Comparación de estrategias de análisis de bases de datos de la producción agropecuaria

Ana Lucía Ferreira López

Maestría en Ciencias Agrarias
Opción Bioestadística

Septiembre 2023

Comparación de estrategias de análisis de bases de datos de la producción agropecuaria

Ana Lucía Ferreira López

Maestría en Ciencias Agrarias
Opción Bioestadística

Septiembre 2023

Tesis aprobada por el tribunal integrado por el Ing. Agr. PhD. Guillermo Siri, el Ing. Agr. PhD. Pablo González y la Ing. Agr. PhD. Virginia Gravina el 26 de setiembre de 2023. Autora: Ing. Agr. Ana Lucía Ferreira López. Director: Ing. Agr. PhD. Jorge Franco Durán: Codirectores: Ing. Agr. PhD. Sebastián Mazzilli e Ing. Agr. PhD. Oswaldo Ernst.

AGRADECIMIENTOS

Agradezco a la Comisión Académica de Posgrado (CAP) por el apoyo económico, a mi tutor y cotutores por orientarme y ofrecerme su tiempo y conocimiento, sin los cuales el desarrollo de esta tesis no habría sido posible. A los compañeros del Departamento de Estadística y Computación, así como a los compañeros de la estación experimental Mario A. Cassinoni y la estación experimental de la Facultad de Agronomía (Salto), que colaboraron con la elaboración de presentaciones orales brindando apoyo y conocimiento. Y especialmente a mis amigos y familiares que han sido parte del proceso de aprendizaje y que han aportado desde diferentes ámbitos.

TABLA DE CONTENIDO

página

RESUMEN

El registro de datos de la producción agropecuaria ha crecido sustancialmente. El volumen de información generado requiere de métodos estadísticos de análisis que aseguren un buen tratamiento de los datos y que permita alguna medida de confiabilidad de los resultados obtenidos. Los objetivos de este trabajo fueron proponer y comparar métodos de análisis estadístico aplicable a bases de datos de este tipo y, utilizando los resultados obtenidos, generar modelos de predicción de rendimiento para soja. Se utilizó el método de agrupamiento de mínima varianza entre grupos de Ward con matrices de distancia de Gower a partir de las variables originales, y euclidiana a partir de las coordenadas principales resultantes de un MFA (análisis factorial múltiple). Se utilizaron modelos lineales mixtos para predicción. El agrupamiento a partir de las distancias euclidianas mostró mayor consistencia en la diferenciación de las prácticas de manejo. Se establecieron tres grupos, dos de ellos con importancia agronómica: el grupo 1, caracterizado por prácticas intensivas y rotaciones diversas, el grupo 2, por una menor intensidad de uso del suelo y una mayor homogeneidad en las rotaciones. El análisis de varianza para el modelo de predicción de rendimiento evidenció un efecto significativo para las variables seleccionadas (fecha de siembra, cultivos antecesores y agregados de nutrientes principalmente) y una diferenciación en grupos de rendimiento. La fecha de siembra fue una de las variables más importantes, se estimó una pérdida potencial de rendimiento de 300 kg por sembrar en fechas posteriores al 8 de diciembre. Con el fin de validar la metodología los resultados, fueron aplicados a una base de datos de trigo. Se pudieron diferenciar grupos y las variables que los discriminaron fueron, en efecto, las ya conocidas o esperadas para modelar el rendimiento.

Palabras clave: clúster, métodos de análisis, modelación, registros de la producción agropecuaria, rendimiento de soja

COMPARISON OF ANALYSIS STRATEGIES OF AGRICULTURAL DATABASES

SUMMARY

The record of agricultural production data has grown substantially. The volume of generated information requires statistical analysis methods that ensure good treatment of the data and that allow some measure of reliability of the obtained results. The objectives of this work were to propose and compare methods of statistical analysis applicable to databases of this type and, using the methodological results, to evaluate yield prediction models for soybean. The minimum variance clustering method between Ward groups with Gower and Euclidean distance matrices was used, based on the principal coordinates resulting from an MFA (Multiple Factor Analysis). The modeling used mixed linear models. The grouping based on Euclidean distances showed greater consistency in the differentiation of management practices. Three groups were established, two of them with agronomic importance: group 1, characterized by intensive practices and diverse rotations, group 2, by a lower intensity of land use and greater homogeneity in rotations. The analysis of variance for the yield prediction model showed a significant effect for the selected variables (planting date, predecessor crops and nutrient additions mainly) and a differentiation in yield groups. The sowing date was one of the most important variables, a potential yield losses of 300 kg was estimated for sowing on dates after December 8th. To validate the methodology, the results were applied to a wheat database. Groups were able to be differentiated and the variables that discriminated them were, in effect, those already known or expected to model performance.

Keywords: analysis methods, cluster, modeling, soybean yield

1. INTRODUCCIÓN

1.1 MARCO GENERAL

El registro de información productiva y su análisis posterior no es algo nuevo en el sector agropecuario. No obstante, para el sector agrícola, debido a la acelerada expansión que ha tenido recientemente (DIEA, 2017), se produjo un aumento significativo de la información disponible. En la actualidad, el tratamiento masivo de datos, dado el acelerado desarrollo de las tecnologías de la información y la comunicación, ha dado lugar a nuevos conceptos y paradigmas en la gestión de la información: minería de datos y *big data* (Amoroso y Costales, 2016).

La necesidad, además, de generar una producción cada vez más eficiente, lo que se conoce como intensificación ecológica de la agricultura (Tittonell, 2014), ha llevado a que las instituciones públicas, federaciones y empresas privadas registren cada vez más información en la búsqueda de la optimización de resultados productivos. En Uruguay, el registro de información de este tipo ha sido utilizado principalmente por el Plan Agropecuario, la Facultad de Agronomía, la Oficina de Estadísticas Agropecuarias, la Federación Uruguaya de Grupos CREA y algunas empresas privadas.

Las bases de datos generadas en una serie de años en las unidades productivas contienen información de alta confiabilidad sobre los sistemas de producción, la toma de decisiones, el uso de recursos y los paquetes tecnológicos. Generalmente, el uso que se le ha dado a esta información en el país es de tipo descriptivo-explicativo, un ejemplo de esto es el Programa de Monitoreo de Empresas Ganaderas del Plan Agropecuario que se hace a partir de la información recolectada en las llamadas Carpetas Verdes.

El análisis estadístico apropiado de las mencionadas bases de datos es fundamental para su correcta interpretación, la generación de hipótesis y la predicción de efectos, con el fin de lograr mejoras en los sistemas de producción y el uso de los recursos. Conocer la confiabilidad predictiva de los métodos a utilizar es de suma importancia.

1.2 MÉTODOS DE ANÁLISIS

En las ciencias biológicas y en el área agrícola se requiere una labor eficiente en la organización y el desarrollo de la investigación sobre modelos estadísticos, con el apoyo de las nuevas tecnologías de la información y la comunicación (Guerra et al., 2003). En la investigación agronómica se han estado utilizando metodologías de análisis aplicadas tanto a datos experimentales como a datos provenientes de muestreos para explicar y predecir resultados productivos y económicos. Con mucha frecuencia, al utilizarse bases de datos de registros productivos, no se valoran los supuestos teóricos de los modelos estadísticos y no se establecen conclusiones válidas a partir de la información analizada (Guerra et al., 2003).

Generalmente, la información registrada en un proceso de observación es tratada, en un primer momento, con el objetivo de describir y resumir sus características más sobresalientes. Esto se conoce como estadística descriptiva y se suele basar en el uso de tablas y gráficos, y en la obtención de medidas resumen (Di Rienzo et al., 2005). La estadística descriptiva es el primer paso en cualquier análisis y más aún en los casos de bases de datos donde las hipótesis y posteriores análisis surgen de la descripción y exploración. Esta información viene dada por características observadas en los individuos que pueden ser medidas con diferentes tipos de variables (discretas y/o continuas). En el caso de la estadística aplicada a las ciencias agrarias, existiendo múltiples factores que determinan los resultados y varias formas de medir la respuesta, es conveniente analizar en conjunto toda la información y se dice, entonces, que el análisis es multivariado.

El modelo lineal (ML) en sus dos principales formas, general y mixto, es, probablemente, la herramienta de inferencia estadística más utilizada en las investigaciones científico-técnicas en el campo de las ciencias biológicas en general y en las agropecuarias en particular. El ML es un método estadístico cuya finalidad es probar hipótesis referidas a los parámetros de posición y variabilidad de dos o más

poblaciones en estudio (Di Rienzo et al., 2005). Este ML general se basa en los supuestos de homogeneidad de varianzas y normalidad.

Cuando se tratan de explicar los resultados finales de los predios agropecuarios, las variables o factores involucrados son muchos. El análisis de componentes principales (PCA, por sus siglas en inglés) ha estado siendo utilizado como un método recurrente para la reducción en el número de variables de tipo continuo que luego serán introducidas en algún análisis de agrupamiento (o análisis de conglomerados) para generar grupos homogéneos (Urruty et al., 2017) y finalmente esos agrupamientos se modelan a través de modelos lineales.

Ernst et al. (2016), buscando explicar brechas de rendimiento en el cultivo de trigo por una posible pérdida de productividad de los suelos vinculada al aumento de la intensidad de uso, trabajó con 1700 registros de FUCREA de producción de trigo (datos de producción). Mediante el análisis de fronteras estocásticas de producción generó una función de ineficiencia que explicaría la diferencia entre el rendimiento alcanzable y el obtenido. En esta función de ineficiencia se incluyó, entre otras variables de manejo, un índice climático generado a partir de PCA.

El trabajo de Urruty et al. (2017) para determinar la robustez del rendimiento de trigo en Francia, también a partir de registros de productores, es otro ejemplo de métodos aplicados. En este caso, el análisis consistió en utilizar PCA para reducir el número de variables para el análisis y con las nuevas coordenadas realizar agrupamiento de productores a través de análisis de agrupamiento, con lo que logró, finalmente, una tipología de productores de la zona bajo estudio que permitió concluir sobre la importancia de la heterogeneidad en las prácticas de manejo. Los sistemas de producción definidos como intensivos en cuanto a rotaciones y uso de agroquímicos demostraron mayor robustez.

Siguiendo con el uso de datos de la producción, Mazzilli et al. (2016) propusieron el uso de fronteras estocásticas para estimar las pérdidas de rendimiento en trigo por efecto del cultivo de invierno antecesor y un trabajo similar fue publicado con respecto al cultivo de soja (Mazzilli y Ernst, 2019) donde se logró determinar las variables que afectan al rendimiento de cultivos y la importancia del uso de prácticas integradas de manejo. En general, el tratamiento de variables de tipo discreto (binarias

y multinomiales) no está contemplado en estos trabajos o se han hecho transformaciones de estas a variables discretas a continuas; este es el caso de los cultivares, por ejemplo, que se transforman a rendimiento potencial (Ernst et al., 2016).

Cuando el objetivo del análisis es el agrupamiento (clasificación) de observaciones, existen alternativas a la transformación de datos antes del análisis de clasificación. Una de ellas es el cálculo de la matriz de distancias de Gower entre individuos u observaciones; esta distancia cumple con las propiedades matemáticas de una distancia euclidiana y puede utilizar variables continuas y discretas o categóricas (binarias, ordinales y nominales). Otra alternativa es la transformación conjunta de todos los tipos de variable mencionados a nuevas coordenadas en una escala continua, con mínima (o ninguna) perdida de información. Esta transformación se logra con el análisis de múltiples factores o análisis factorial multiple (MFA, *Multiple Factor Analysis*; Le Dien y Pagès, 2003).

En este trabajo se comparan ambos métodos de distancia aplicados al agrupamiento jerárquico. No se han encontrado trabajos referidos a la comparación y validación de métodos estadísticos multivariados aplicados en bases de datos de la producción agropecuaria.

1.3 ANÁLISIS FACTORIAL MÚLTIPLE, DISTANCIAS Y AGRUPAMIENTO JERÁRQUICO

1.3.1 Análisis factorial múltiple

El análisis factorial múltiple (MFA) es una técnica de análisis estadístico multivariado que se aplica sobre un conjunto de datos conformados por grupos de variables de diferente naturaleza y fue propuesto por Escofier y Pagès (1994) y Pagès (2002). El método se centra en un análisis multivariado de factores, aplicado a cada uno de los grupos de variables, en el cual cada grupo de variables es balanceado (o estandarizado para lograr el balance), para concluir con un análisis de componentes principales aplicado a las variables balanceadas previamente.

En principio, el MFA se utilizó para grupos de variables continuas de diferente origen y con grandes diferencias en el número de variables por grupo, en cuyo caso dominarían en el PCA aquellos grupos de variables en los que haya número más elevado de ellas; posteriormente se generalizó a la mezcla de diferentes tipos de variable. El método, en general, consiste en aplicar metodologías de reducción de variables para expresar las observaciones en un sistema de coordenadas denominadas coordenadas principales. Se realiza un PCA para cada grupo de variables continuas, un análisis múltiple de correspondencia (MCA, *Multiple Correspondence Analysis*) para los grupos de variables categóricas (incluyendo las binarias) y un análisis de correspondencia (CA, por su sigla en inglés) para variables de frecuencia, expresadas como proporciones.

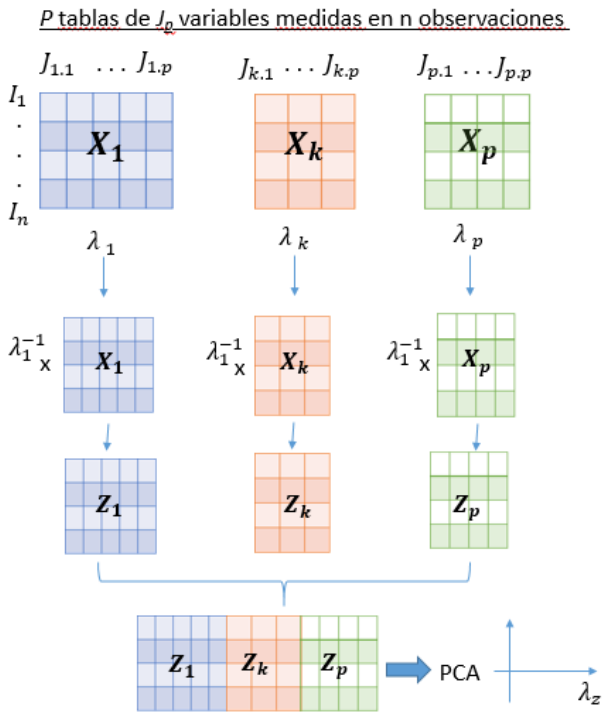

Figura 1. Matriz global de grupos de variables originales convertidas a Z variables estandarizadas (coordenadas). I: individuos. J: variables. λ: auto valores. z: valores estandarizados por λ^{-1}.

El MFA realiza un PCA individual en cada uno de los grupos de variables continuas (estandarizadas o no) K_j y un PCA balanceado (no estandarizado) en las variables discretas. Si llamamos Z_j a la matriz que contiene a las variables discretas, podemos transformar a $Z_j = \{z_{ijk}\}$ a $Y_j = \{y_{ijk}\}$, a través de la fórmula: $y_{ijk} = \frac{(z_{ijk} - w_{jk})}{w_{jk}}$, donde $w_{kj} = \sum_{i \in I} p_i \cdot z_{ijk}$ es la proporción de entradas correspondiente a la columna k_j (k_j = 1,2,…,K_j, número de columnas en la matriz Y_j), p_i es el peso asociado a cada entrada (p_i = 1/I) y $\frac{w_{kj}}{Q_j}$ ($Q_j = \sum_{k \in K_j} w_{jk}$) se utiliza como peso de la columna.

Luego se obtienen los valores propios o autovalores λ_1^j, $j = 1,2,\ldots, J_q + J_c$ para el grupo de variables continuas (J_q) y las discretas (J_c), que corresponden a las direcciones de máxima variabilidad (inercia). En este estudio no existen variables de tipo frecuencia, pero, si las hubiera, se realizaría el CA y el cálculo de sus valores propios.

En el siguiente paso, $1/\lambda_1^j$ es utilizado como ponderador de cada variable del grupo j de variables continuas y $\frac{w_{kj}}{Q_j \lambda_1^j}$ como ponderador de cada variable del grupo j de variables discretas. Una vez realizada la ponderación, todas las variables tienen el mismo peso entre grupos. Geométricamente, esto implica igualar a 1 la inercia axial máxima de cada una de las nubes j. Posteriormente, en el análisis global, esto garantiza que ninguno de los grupos pueda generar por sí solo el primer eje, o primera coordenada principal.

	1... i... I	Pesos columna (λ_1^j)	Pesos fila
Variable k en el grupo de variables cuantitativas J_q	$\frac{x_{ijk} - \bar{x}_{kj}}{s_{kj}}$	$\frac{1}{\lambda_1^j}$	$p_i = \frac{1}{I}$
Variable k en el grupo de variables categóricas J_c	$\frac{z_{ijk} - w_{kj}}{w_{kj}}$	$\frac{w_{kj}}{Q_j \lambda_1^j}$	
Variable k en el grupo de variables de frecuencia J_f	$\frac{f_{ijk} - (f_{i.j}/f_{..j}) * f_{.kj}}{p_i * f_{.kj}}$	$\frac{f_{.kj}}{\lambda_1^j}$	

Figura 2. Transformación de los datos de la tabla inicial, pesos fila y pesos columna para los tres tipos posibles de variables.

Fuente: Marcillo (2017).

Finalmente, las puntuaciones (*scores*) de cada una de las entradas y la contribución a la variabilidad total de cada una de las entradas, variables y grupo de variables son el resultado del MFA (Franco *et al*., 2010).

El MFA facilita la caracterización de cada unidad/individuo por vectores de coordenadas principales. Esto permite convertir los ejes coordenados expresados en su métrica original en nuevos ejes de coordenadas de tipo continuo llamadas coordenadas principales. Estos scores toman en consideración la contribución de cada grupo de variables: a mayor número de variables, menor es la contribución de cada variable en cada eje, puesto que la suma es del 100 %. Se obtiene una distancia global provocada por el MFA, que es una suma ponderada de las distancias entre los individuos *i* y *l* para cada grupo de variables (utilizando la distancia euclidiana):

$$d^2(i,l) = \sum_{j \epsilon J_q} \frac{1}{\lambda_1^j} \sum_{k \epsilon K_j} \left[\frac{x_{ijk} - x_{lkj}}{s_{kj}} \right]^2 + \sum_{j \epsilon J_c} \frac{1}{\lambda_1^j} \sum_{k \epsilon K_j} \frac{1}{Q_j \, w_{kj}} \left[z_{ikj} - z_{lkj} \right]^2 \quad (1)$$

La ecuación 1 representa la contribución de cada grupo de variables a la distancia global, la ponderación dada por el inverso del primer valor propio (que es siempre el máximo) equilibra la influencia de los grupos.

1.3.2 Medidas de distancia

El concepto de distancia entre objetos o individuos permite interpretar geométricamente muchas técnicas clásicas del análisis multivariado a partir de la representación de los objetos como puntos en un espacio métrico de p ≥ 2 dimensiones adecuado.

Se denomina distancia a la longitud de la recta que une dos puntos en un espacio geométrico de cualquier dimensión. Desde un punto de vista formal, para un conjunto de elementos X, se define distancia como cualquier función binaria, $d(a,b)$ de X*X: $\Re^p \rightarrow \Re$, con p número de variables (dimensiones), que verifique las siguientes condiciones:

i) No negatividad:

$$d(a,b) \geq 0 \quad \forall a, b \epsilon X$$

ii) Simetricidad:

$$d(a,b) \leq d(b, a) \quad \forall a, b \epsilon X$$

iii) Desigualdad triangular:

$$d(a,b) \leq d(a,c) + d(c,b) \quad \forall a, b \epsilon X$$

Se denomina distancia euclidiana entre dos puntos $A(x_1, y_1)$ y $B(x_2, y_2)$ a la longitud del segmento de la recta que tiene por extremos a los puntos A y B y se expresa como:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Esta medida de distancia solo se puede aplicar a variables de tipo continuo, por sus operaciones. De todas formas, la distancia euclidiana no es recomendable cuando se trabaja con las variables originales (sin transformación), ya que es sensible a los cambios de escala y presupone que las variables son no correlacionadas.

La distancia de Gower (Gower, 1971), por otro lado, opera con mezcla de variables, es decir, de continuas, nominales y ordinales. Originalmente, la distancia de Gower fue un coeficiente de similaridad definido entre 0 y 1, pero se transforma a una distancia (s) al operar: s(i, j) = 1- d(i, j), donde:

$$d(_i, j) = \frac{\Sigma_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\Sigma_{f=1}^{p} \delta_{ij}^{(f)}} \quad (2)$$

El indicador $\delta_{ij}^{(f)}$ toma valor 1 cuando ambas medidas x_{if} y x_{jf} para la *f- ésima* variable son valores no faltantes; de otra forma, toma el valor 0. El número $d_{ij}^{(f)}$ es la contribución de la *f- ésima* variable a la distancia entre i y j. Si la variable f es binaria o nominal, entonces $d_{ij}^{(f)}$ es definida como:

$$d_{ij}^{(f)} = 1 \; si \; x_{if} \neq x_{jf} \; , d_{ij}^{(f)} = 0 \; si \; x_{if} = x_{jf}$$

Si la variable es nominal, la expresión 2 se convierte la proporción de coincidencias respecto al número total de pares (comparaciones) y coincide con el coeficiente *de simple matching*. Si la variable es continua, entonces $d_{ij}^{(f)}$ es la distancia de Manhattan, estandarizada por el rango para que tome valores en el intervalo [0, 1]:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$$

Donde R_f es el rango de la variable f (Escobedo y Salas, 2008).

1.3.3 Agrupamiento jerárquico

El análisis de conglomerados es una técnica estadística multivariada que busca agrupar elementos o individuos que presenten mayor similitud (menor distancia) entre ellos. Se busca la mayor homogeneidad dentro de los grupos y la mayor heterogeneidad entre grupos.

A partir de la matriz de datos de orden (n x p) con n individuos y p variables, es posible generar dos tipos de matrices que permiten medir las similaridades o las distancias entre pares de individuos: matriz de similaridad y matriz de distancia. Un aumento de la similaridad implica un aumento de la semejanza entre individuos, y toda similaridad de un individuo consigo mismo debería ser igual al máximo valor posible, es decir, 1 (Demey et al., 2011).

El análisis de conglomerados, cuando se realiza con métodos que no asumen una distribución estadística de las variables ni de los grupos de variables, como el método jerárquico utilizado en este trabajo, no requiere de supuestos distribucionales sobre los cuales generar inferencias estadísticas para una población a partir de una muestra. Es un método basado en criterios geométricos y se utiliza fundamentalmente como una técnica exploratoria descriptiva.

Para la comparación de conglomerados se utilizó el índice Rand ajustado (ARI), que es una medida de la similitud entre dos agrupaciones de datos. Es una corrección del índice de Rand, que es una medida básica de similitud entre dos agrupamientos. También se evaluó el índice de concordancia, que establece el porcentaje de acuerdo, es decir, en qué medida hubo coincidencia en la clasificación de las observaciones entre un agrupamiento y otro.

1.4 JUSTIFICACIÓN

Aunque existe un gran volumen de información disponible y se han aplicado técnicas de análisis diversas, no se ha estudiado un método de análisis exploratorio que asegure un buen tratamiento de los datos, ni se ha estudiado la repetitividad de los métodos aplicados para el análisis exploratorio o para las predicciones (modelación).

El principal problema de aplicar métodos de análisis estadístico a este tipo de información es que no se está trabajando bajo condiciones controladas, es decir que se desconoce y no se controla la situación o el entorno a partir del cual se genera la información y esto genera incertidumbre respecto a la extrapolación de los resultados obtenidos. A partir de datos no experimentales no se pueden establecer relaciones de causalidad, solamente se pueden establecer relaciones de tipo descriptivo entre las variables (Lindsey, 1995). En segundo lugar, pero no menos importante, está el hecho de que estos registros no son medidos en una muestra aleatoria de una población objetivo, sino que son registros de productores y técnicos que brindan su información a las instituciones, en este caso FUCREA, y forman parte de un grupo con ciertas características (avanzados, miembros de un grupo o asociación, muy informados, etc.). Estos últimos puntos podrían afectar el alcance de la inferencia estadística a partir de la información disponible. De todas maneras, este tipo de base de datos contiene información confiable y puede, por tanto, ser utilizada para generar hipótesis que conduzcan a nuevas líneas de investigación.

1.5 OBJETIVOS

1.5.1 Objetivo general

El objetivo principal de esta tesis es proponer y estudiar empíricamente el comportamiento de algunos métodos de estadística multivariada para el análisis y la utilización de bases de datos generadas en los procesos de producción.

1.5.2 Objetivos específicos

1. Generar algunas pautas pare el acondicionamiento previo de las bases de datos y la transformación de variables que faciliten el análisis estadístico de la información.
2. Establecer la importancia de las variables en la diferenciación entre grupos de racionalidad productiva.
3. Comparar y seleccionar modelos de predicción de rendimiento que sean extrapolables por grupo de producción (de racionalidad productiva) o a todos los grupos.
4. Estimar otros modelos agronómicos referidos al uso de agroquímicos
5. Estudiar la confiabilidad de las predicciones y la repetitividad de los métodos.

1.6 HIPÓTESIS DE TRABAJO

La hipótesis fundamental de trabajo es que la información de la forma de producción (lugar, componentes del sitio, tecnología y manejo de insumos, etc.) de una unidad de información corresponde a una «racionalidad» en el modo de producir y que diferentes «racionalidades» deben producir diferentes resultados.

A partir de registros de manejo y resultados productivos de unidades de información es posible identificar grupos homogéneos de productores agrícolas y caracterizar su racionalidad productiva, identificar limitantes y predecir resultados ante

diferentes escenarios productivos. La inferencia generada puede ser evaluada en cuanto a su capacidad predictiva.

La capacidad de predecir resultados con bajo margen de error requiere ajustar la metodología de análisis de datos (métodos, técnicas y modelos) para que sea extrapolable a diferentes condiciones (otras bases de datos de origen similar).

1.7 RESULTADOS ESPERADOS

Como se estableció desde el principio, el objetivo es probar diferentes técnicas de agrupamiento ya utilizados en este tipo de bases de datos y proponer una metodología que permita la utilización simultánea de diferentes tipos de variable (discretas y continuas) para la formación de grupos y análisis de modelos. Desde el punto de vista estadístico, importa que los métodos sean correctos y aplicables y, desde lo agronómico, que tengan coherencia y lógica productiva.

• El estudio dará como resultado algunas propuestas comparadas y evaluadas sobre la forma de analizar este tipo de datos. No se espera establecer el método, sino evaluar y comparar metodologías y establecer su nivel de confiabilidad.

• Los productores se pueden agrupar utilizando la información de su forma de producir, esto significa que existe una «racionalidad» en la toma de decisiones que puede ser comprendida y estudiada utilizando sus registros y el resultado productivo.

• Los grupos de «racionalidad» que se puedan generar obtienen resultados productivos diferentes entre grupos y similares dentro del grupo.

• Se puede obtener una metodología (métodos, técnicas y modelos) que tenga la capacidad de predecir resultados con bajo margen de error en diferentes condiciones y que sea extrapolable a otras bases de datos.

2. MATERIALES Y MÉTODOS

2.1 BASE DE DATOS

Se trabajó con dos grupos de bases de datos provenientes de registros de manejo de cultivos y rendimientos de productores integrantes de la Federación Uruguaya de Centros Regionales de Experimentación Agropecuaria (FUCREA). Un grupo de bases tenía incorporado el cálculo de los indicadores de sustentabilidad (Mazzilli, 2018) y otro grupo tenía los datos de manejo de los mismos predios para los que se estimaron los indicadores. La base con datos de manejo contiene información de nueve años de registros de resultados productivos y prácticas de manejo (2006 al 2014), mientras que el cálculo de indicadores se realizó para el período de 2010 a 2014. Con la información de las prácticas de manejo que se consideraron relevantes, se generó una base única con registros de 28 predios que conforman 5862 unidades de información (UI; cada una de ellas es la combinación de: establecimiento, potrero, año, zafra y cultivo sembrado). La primera instancia del trabajo consistió en generar grupos homogéneos (conglomerados) de estas unidades productivas, sin incluir las variables referidas a resultados productivos.

Se revisaron las 28 bases de datos de cada uno de los predios. Estas planillas contenían información de las prácticas de manejo y, si bien mantenían un formato, no todas coincidían exactamente, además, muchas de las variables contenidas eran redundantes. Se extrajo la información relevante de manejo y se fueron resumiendo estas variables (había bases con información para 120 hasta 230 variables). Este proceso de depuración condujo a una única base con 58 variables, de las cuales 6 son identificadoras y 52 de manejo (anexo 1). Esta primera base de datos resumida contiene la información de las variables independientes, no contiene registros de resultados productivos. Estos se resumieron de manera similar; mediante un análisis de correlación se extrajeron las consideradas independientes. Luego de algunas estrategias como generación de nuevas variables y/o resumen de otras, se generó una base única que contenía toda la información (manejo y resultados productivos) y se redujo a los cultivos de trigo y soja. Esta base constaba de 2472 observaciones (el 42

% de la información inicial), de las cuales 1716 correspondían a soja y 757 a trigo. El último ajuste realizado a la base de datos fue la eliminación de las variables que contenían más del 50 % de valores faltantes. Se establecieron métodos de análisis para soja, con un número inicial de 23 variables (7 factores y 16 continuas) y que luego fueron replicados para el cultivo de trigo.

2.2 METODOLOGÍA

2.2.1 Matrices de distancia

Con base en la hipótesis de trabajo planteada de que existen racionalidades en la producción, se realizó un agrupamiento de observaciones a partir de los valores que tomaron las variables independientes en cada una de las UI. Se usaron dos estrategias de agrupamiento, una utilizando la distancia de Gower (1971) y otra utilizando el análisis factorial múltiple (MFA) como método de reducción de p variables originales a k ($k < p$) coordenadas principales para, posteriormente, calcular la distancia euclidiana. Para el agrupamiento se utilizó el método jerárquico mínima varianza dentro de grupos propuesto por Ward (1968).

La distancia de Gower permite que las variables bajo estudio sean de tipo mixto, es decir, nominales, ordinales, binarias simétricas o asimétricas y/o continuas. Por otro lado, a partir del MFA, se calcularon las coordenadas principales (paquete de FactoMineR en R) y, posteriormente, las distancias euclidianas (función «Daisy», paquete Cluster, en R).

2.2.2 Agrupamiento

Una vez obtenidas las matrices de distancias se realizó el dendrograma con la función «hclust» del paquete fastcluster de R y con la opción Ward.D2, que permite el cálculo de sumas de cuadrados y varianzas utilizadas por el método de Ward. La definición del número óptimo de grupos es un problema no resuelto matemáticamente, existen más de 30 propuestas con ventajas y desventajas. Una de estas posibles técnicas

es graficar la suma de cuadrados dentro de grupos y/o los valores del estadístico pseudo-F según el número de grupos. De acuerdo con Calinski y Harabatz en 1974 (Wilkinson, 1994), este valor es la relación de varianzas entre grupos con la varianza dentro de grupos y se calcula como:

$$Pseudo\ F = \frac{(\mathrm{GSS})/\ (\mathrm{K}-1)}{(\mathrm{WSS})/\ (\mathrm{N}-\mathrm{K})}$$

Donde N es el número de observaciones, K es el número de clústeres en cualquier paso del clúster jerárquico, GSS es la suma de cuadrados entre grupos y WSS es la suma de cuadrados dentro de grupos. El número óptimo se establece como el valor en el cual la variación dentro de los grupos y/o la pseudo-F se estabilizan.

Para la interpretación agronómica de los grupos se estudió la importancia de las variables originales en la estructura de los agrupamientos. En el caso de las variables continuas, se utilizaron los valores F (de Fisher) provenientes del análisis de varianza, ANAVA, para cada una de ellas y, para las variables discretas, una prueba de razón de verosimilitud de chi-cuadrado (G test). Cada una de las variables se modeló en función del agrupamiento (variables continuas) o con una tabla de doble entrada: grupo x categorías de la variable (variables categóricas). Para balancear la comparación de los efectos de las variables categóricas en el agrupamiento, el valor de G se dividió entre los grados de libertad de la prueba.

2.2.3 Transformación de variables para la interpretación agronómica de los grupos

Para poder explicar y caracterizar a los grupos desde el punto de vista agronómico, se optó por trabajar con las variables de interés, transformadas a indicadores que resumieron información para tres de ellas: antecesor verano, ciclo del cultivo y fecha de siembra.

Para antecesor verano se calculó el porcentaje de gramíneas, soja de segunda, pradera o campo natural y de los valores faltantes. En cuanto a la variable ciclo del cultivo, se estimaron los porcentajes de ciclos cortos (grupos de madurez menores a

5), medios (entre 5 y 6) y largos (mayores a 6). La variable fecha de siembra se clasificó en temprana (del 10/10 al 20/10), regular (del 20/10 al 22/11), tardía (23/11 al 7/12) y muy tardía (del 8/12 al 14/12) y se estimaron los porcentajes para cada grupo.

Luego de establecidos los grupos, se utilizaron modelos lineales mixtos para comparar los resultados obtenidos por los diferentes grupos (como ya se mencionó, las variables de respuesta no fueron incluidas en el análisis de clasificación) y para modelar las variables respuesta de interés.

2.2.4 Utilización de los agrupamientos en modelos de predicción

Se plantearon modelos de predicción de rendimiento a modo de visualizar si las variables que determinaron el agrupamiento también determinan rendimiento y, de esta forma, generar grupos de rendimiento a partir de las diferentes racionalidades productivas. Los modelos se probaron teniendo en cuenta todas las variables que resultaron determinantes del agrupamiento en el proceso anterior. Se descartaron aquellas que presentaron p-valores mayores a 0,05 en el modelo. Los modelos se realizaron con el objetivo de asociar grupos de producción a racionalidades productivas, y, por otro lado, responder a algunas interrogantes agronómicas planteadas por especialistas en el área.

En primer lugar, se planteó un análisis de varianza para determinar la significancia de las diferentes variables en el agrupamiento y luego se corrieron los diferentes modelos, de efectos fijos, aleatorios y mixtos, con varianzas heterogéneas y homogéneas para grupos.

La modelación del rendimiento se planteó a partir de las variables que explicaron el agrupamiento. Esto para determinar si los grupos formados por las variables prácticas de manejo predecían grupos de rendimiento diferentes. Se incluyeron las variables asociadas a la rotación (antecesores), la fecha de siembra, el número de fertilizantes aplicados, el número de insecticidas aplicados y el agregado de fósforo, azufre y potasio.

En cuanto a los modelos de aplicación, tuvieron como objetivo evaluar la respuesta del rendimiento al uso más o menos intensivo de agroquímicos y, por otro lado, si los balances de nutrientes estaban explicados en mayor medida por el rendimiento o por el agregado de fertilizantes, es decir, si un balance positivo implicaba que se fertilizó lo suficiente o que el cultivo rindió poco. Es importante resaltar que, para el caso de aplicaciones de agroquímicos, en la base de datos solamente es posible analizar la cantidad de aplicaciones y las unidades toxicológicas de la zafra, pero no el objetivo de la aplicación y la eficacia y eficiencia.

Modelo 1: en este modelo se consideraron todos los efectos fijos, a excepción del año, que fue aleatorio, con varianzas homogéneas para grupos. El objetivo de este modelo es estimar la significancia de las variables bajo estudio, además de las estimaciones de los efectos. Se asume en este modelo que, si bien los grupos pueden producir diferentes resultados productivos promedio, no se diferencian en variabilidad.

Modelo 2: en este modelo se consideraron todos los efectos fijos, a excepción del año, que fue aleatorio, con varianzas heterogéneas para grupos. El objetivo de este modelo es estimar la significancia de las variables bajo estudio, además de las estimaciones de los efectos. Al permitir varianzas heterogéneas entre grupos, el modelo permite analizar si los grupos inducen diferentes grados de variabilidad en las variables dependientes.

Modelo 3: se consideraron todos los efectos aleatorios, excepto el grupo que es fijo y varianzas heterogéneas. El objetivo de este modelo es estimar los componentes de varianza de los efectos aleatorios y la proporción de varianza del rendimiento que es explicada por cada una de las variables y por el error. También se predijeron los BLUP (mejor predictor lineal insesgado) para describir la dirección (signo) de los efectos sobre el rendimiento. Se trabajó con los datos imputados para la estimación de los parámetros de todos los modelos.

Modelo 4: modelo de componente de varianzas, con varianzas homogéneas. Variables continuas fijas y las demás como aleatorias.

Modelo 5: modelo de componente de varianzas, con varianzas homogéneas. Variables de clasificación fijas y todas las demás aleatorias.

Se realizaron tres modelos adicionales, relacionados con el balance de nutrientes, a modo de aplicación y para tratar de responder algunas interrogantes puntuales relacionadas con la base de datos. Se modeló el balance de P, S y K en función del agregado de nutriente y del rendimiento. Desde el punto de vista metodológico, estos modelos tienen como objetivo probar la utilidad de los métodos propuestos en la extracción de variables de interés y síntesis de la información.

Para poder realizar el último objetivo del trabajo, la medida de la confiabilidad o alcance de la inferencia, se utilizaron los datos de trigo de la base a modo de validación del método de clasificación. Replicando el análisis en este nuevo conjunto de datos, teniendo en cuenta que se trata de otro cultivo, pero recordando que se trata de probar la metodología y testear que cumpla con sintetizar información, determinar variables de peso en los agrupamientos y la relación de estas variables con variables de interés productivo (respuesta), además de responder ciertas interrogantes particulares de cada cultivo (o individuo estudiado) a partir de la información resumida que se obtiene luego de aplicar la metodología.

Se trabajó con los programas de análisis estadístico R para el agrupamiento y modelación e Infostat y SAS se utilizaron en una primera instancia de estudio y exploración de los modelos.

3. RESULTADOS

3.1 AGRUPAMIENTO

En el proceso que utilizó el MFA como método de reducción de variables a coordenadas, la distribución de las distancias fue muy asimétrica hacia la derecha (figura 3). Las distancias se encontraron entre 0 y 60, un 90 % de estas con valores menores a 11. Las UI que presentaron distancias promedio mayores a 11 con todas las demás se eliminaron por considerarse UI atípicos a la población en estudio. La figura 3 presenta la distribución de las distancias que permanecieron (1,329,265 distancias) en el análisis.

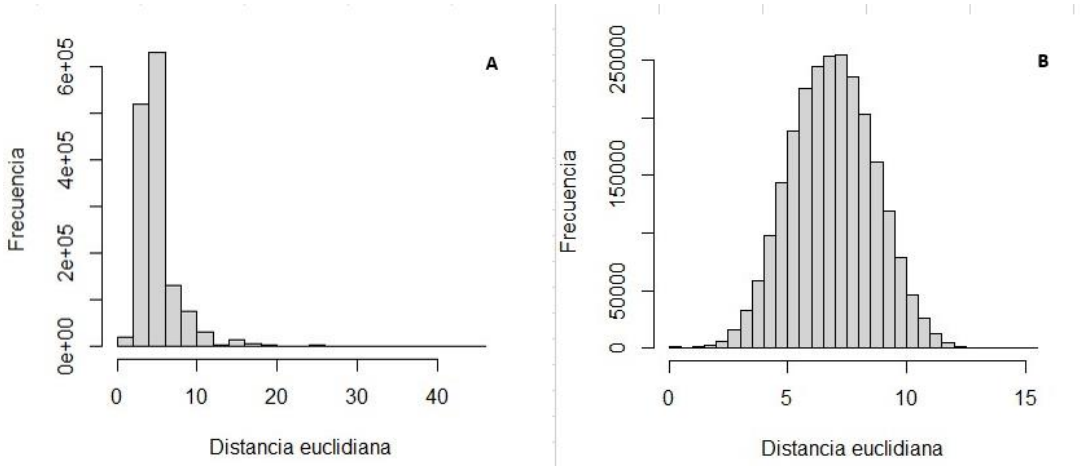

Figura 3. Distribución de distancias euclidianas a partir de MFA con todas las distancias (A) vs. distancias menores a 11 (B)

Eliminando las distancias del decil superior se mejoró la estructura del agrupamiento y se encontraron grupos mejor delimitados (figura 4). Este resultado se obtiene directamente a partir del MFA, con la orden de seleccionar distancias menores a un valor determinado (anexo 3).

Uno de los puntos favorables de utilizar la distancia de Gower es la facilidad que presentó desde el punto de vista operativo y conceptual, puesto que se utilizan los valores de las variables discretas directamente sin ningún tipo de transformación previa. El MFA genera coordenadas a partir de los datos originales; por lo tanto, cada observación tiene una coordenada para cada una de las variables, lo que produce que no haya valores faltantes.

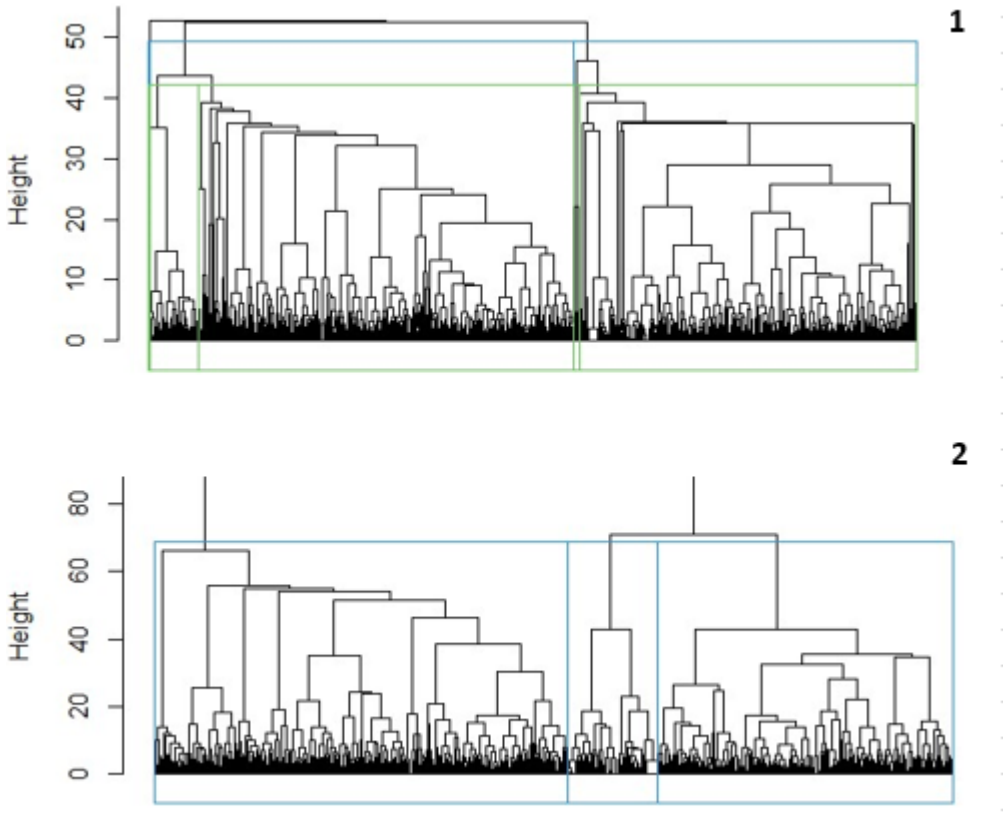

Figura 4. Dendrograma con distancias calculadas a partir de MFA completo (todas las observaciones) (1) vs. dendrograma con distancias calculadas a partir de MFA eliminando observaciones con distancias mayores a 11 (2).

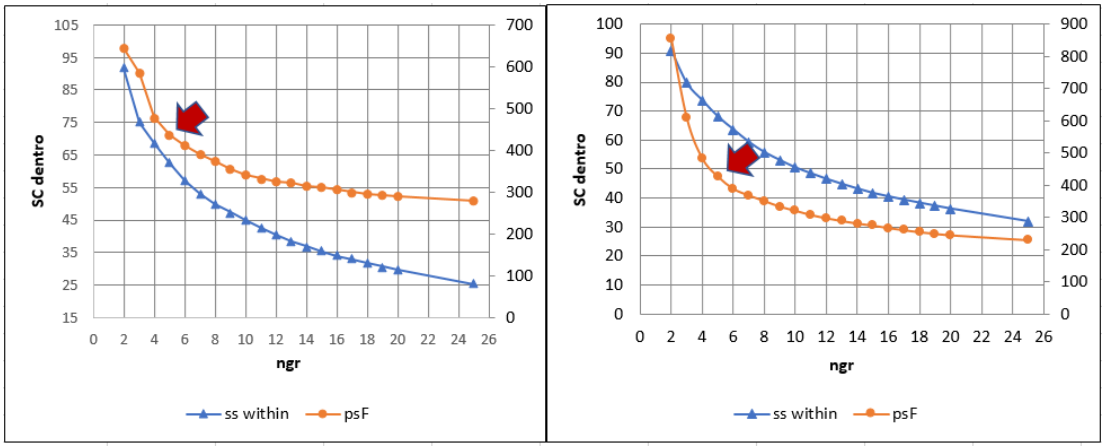

Figura 5. Cambios de valor en las sumas de cuadrados dentro de grupos y del estadístico pseudo-F según número de grupos. SC dentro: suma cuadrado dentro de grupo, ngr: número de grupo, psF: pseudo-F

Para ambos métodos de agrupamiento (Gower-Ward y MFA-Ward) el número óptimo de grupos estuvo entre 3 y 5 (figura 5). Desde el punto de vista estadístico, el agrupamiento a partir de Gower en comparación con el agrupamiento a partir de MFA, dio una concordancia de 0,77 y un ARI de 0,6. La caracterización agronómica se basó en indicadores de prácticas de manejo (variables transformadas) y, como resultado, se estableció que es posible diferenciar 3 grupos a partir del MFA-distancia euclidiana (MFA-DE), con el 90 % de las distancias (cuadro 1).

Cuadro 1. Indicadores utilizados para caracterización de prácticas de manejo por grupo.

Antecesor verano	**Gr1**	**Gr2**	**Gr3**
% gramíneas	38,8	1,8	0
% soja segunda	48,2	26,6	0
% P + CN	5,1	11,9	100
% valores faltantes	3,5	20,2	0
Ciclo cultivo			
% corto	15,2	9,2	0
% medio	74,1	76,6	76,9
% largo	7,5	0,2	0
FS			
% temprana	49,8	8,2	23,1
% regular	25,9	22,6	46,2
% tardía	16,6	32	23,1
% muy tardía	7,8	37,2	7,7

FS: fecha de siembra. P+N pradera y campo natural.

Esta lógica de agrupamiento tiene que ver con los tipos de prácticas involucradas en cada uno de los grupos. El agrupamiento de cinco generó dos grupos bien diferenciados en cuanto a antecesores de verano e invierno, pero luego se observó que uno de los grupos se diferenciaba por un alto porcentaje (más del 60 %) de valores faltantes, razón por la cual se optó por la descripción de tres grupos de prácticas de manejo, considerando las variables involucradas.

En el agrupamiento MFA-DE se observó un grupo que aparece sin cambios tanto en el agrupamiento de 5 como en el agrupamiento de 3. Este grupo, conformado por

13 observaciones, contiene todas las UI que utilizaron en la rotación el campo natural como antecesor de verano. En el agrupamiento según Gower, este grupo no se mantiene y por eso se optó por la descripción de los grupos a partir del MFA-DE.

De acuerdo con los indicadores que aparecen en el cuadro 1, el grupo 1 se clasificó como intensivo y diverso debido al alto porcentaje de uso de gramíneas y de cultivos de segunda. En referencia al ciclo del cultivo, si bien la mayoría utiliza ciclos medios, en este grupo se vio una presencia relativamente alta de ciclos cortos y de ciclos largos, las fechas de siembra predominantes fueron las tempranas, del 10 de octubre al 10 de noviembre.

Cuadro 2. Promedio de rendimiento de soja por grupo a partir de HMFA-DE.

Grupo	Rendimiento (kg/ha)	DE	EE	N.º obs.
1	2589	673	24.4	761
2	2147	676	22.2	928
3	2218	674	187	13

DE: desviación estándar del grupo, EE: error estándar, N.º obs.: número de observaciones por grupo.

El grupo 2 presentó un menor porcentaje de gramíneas y de soja de segunda. En cuanto a los ciclos del cultivo, evidenció un mayor uso de ciclos intermedios, y se caracteriza por fechas de siembra muy tardías respecto a los otros dos grupos. Presentó una alta proporción de siembras muy tardías, desde el 8 al 14 de diciembre, específicamente.

Finalmente, el grupo 3, si bien contiene un número menor de observaciones, como ya se dijo, se mantiene junto bajo cualquier número de grupos seleccionado. Son las unidades que presentan mayor proporción de campo natural como cultivo antecesor; por tanto, la variabilidad en los rendimientos podría explicarse por el manejo del barbecho antes de la siembra y porque no es una práctica que se mantenga en el tiempo en la medida en que no hay un aumento del área de cultivos sobre campo natural.

Cuadro 3. Valores de F y chi-cuadrado obtenidos a partir de anova y prueba de verosimilitud (Gtest) para variables continuas y discretas, respectivamente, según agrupamiento.

Variables discretas	gl	G/gl	Variables continuas	media	desvío	F
FS	6	83,9	Entrada P	32,9	22,7	231,77
AI	16	71,1	Entrada K	11,3	25	151,85
BQ	6	39,8	AA	3,3	2,6	106,06
AV	22	35,6	Entrada S	2	5,2	71,395
Ciclo cultivo	6	26,0	AF	3,8	0,4	63,95
Sistema de siembra	6	7,3	UT mam	3,8	0,4	39,3
			UTAH (posiembra)	1,6	5,2	32,47
			DH	35,2	6,9	27,31
			UTAT	6,3	1,1	27,16
			DS	78,4	22,9	22,8
			UTMH (posiembra)	3,3	0,3	16,05

gl: grados de libertad. G: valor asociado a la prueba chi-cuadrado FS: fecha de siembra. AI: antecesor invierno. BQ: realiza barbecho quimico (si/no). AV: antecesor verano. AA: años de agricultura continua. AF: aplicaciones de fertilizantes (cantidad total de aplicaciones en el ciclo del cultivo). UT mam: unidades toxicológicas mamíferos totales. UTAH: unidades toxicológicas/abeja herbicidas. DH: distancia entre hileras. UTAT: unidades toxicológicas/abeja total. DS: densidad de siembra en Kg semilla/ha. UTMH: unidades toxicológicas/mamífero herbicida.

De las variables continuas, según los valores de F, provenientes del ANOVA, las de mayor importancia fueron la entrada de fósforo (P), en primer lugar, seguido por la entrada de potasio (K) y los años en agricultura. En cuarto lugar, pero con un valor importante de F, aparece la entrada de azufre (S). En el caso de las variables discretas, la fecha de siembra (FS) y los antecesores de verano (AV) e invierno (AI) fueron los que mostraron mayor poder discriminante en el agrupamiento. Si bien la presencia del barbecho químico fue considerable, no se tomó en cuenta, ya que no estaba completa

la información de la fecha de inicio de este ni de las condiciones en que se mantuvo el barbecho previo a la siembra de la soja.

3.2 MODELOS DE PREDICCIÓN

Se estimaron modelos de predicción de rendimiento para determinar si las variables que definieron el agrupamiento también tenían importancia para determinar el rendimiento y, de esta forma, ver si existen grupos de racionalidad productiva asociada al resultado productivo.

Cuadro 4. Prueba de razón de verosimilitud para la prueba de homogeneidad de varianzas dentro de grupo. Homo: varianzas homogéneas. Hete: varianzas heterogéneas.

	Modelo	gl	AIC	BIC	logLik	Test	L. Ratio	p-valor
homo	1	27	26410,80	26557,24	-13178			
hete	2	29	26411,17	26568,45	-13176	1 vs. 2	3,633	0,1626

Según la prueba de razón de verosimilitud (cuadro 4), no se evidenció heterogeneidad de varianzas; se adoptaron, por lo tanto, los modelos con homogeneidad de varianzas.

Si las varianzas se definen homogéneas, entonces el grupo puede incluirse como efecto aleatorio en los modelos agronómicos explicativos para permitir una inferencia de tipo amplio, esto es, que las conclusiones de los otros efectos en los modelos son aplicables al conjunto de las UI. Si, por el contrario, las varianzas fuesen heterogéneas, se debería trabajar con un modelo mixto para cada uno de los grupos, con su respectiva varianza y el grupo como efecto fijo, lo que implicaría que las conclusiones solo sean válidas para ese grupo en particular (no sería válida la extrapolación a todos los grupos).

3.2.1 Modelo 1 de predicción de rendimiento

Fue posible modelar el rendimiento en función del agrupamiento, el año, el antecesor de verano e invierno, la fecha de siembra y el agregado de azufre. Este modelo asume varianzas homogéneas y todos los efectos fijos a excepción del año, que es aleatorio.

$$R = \beta_0 + G_i + a_j + AV_k + AI_l + FS_m + \beta_1 P_n + \beta_2 S_o + \varepsilon_{ijklmno}$$

Dónde: R es rendimiento de soja en kg/ha. G_i, grupo según MFA. a_j, efecto del j-ésimo año. AV_k, efecto del k-ésimo antecesor de verano. AI_l, efecto del l-ésimo antecesor de invierno. FS_m,efecto de la m-ésima fecha de siembra. β_1 ,coeficiente de regresión asociado al agregado de P. β_2 ,coeficiente de regresión asociado al agregado de S. $\varepsilon_{ijklmno}$, efectos residuales al modelo.

Cuadro 5. Resultados del análisis de varianza para el modelo 1.

	gl	F	p-valor
(Intercepto)	1	948	<0,0001
Grupo3E	2	95	<0,0001
FS	1	136	<0,0001
AV	11	4	<0,0001
AI	8	4	0,0002
P	1	20	<0,0001
S	1	16	0,0001

Grupo3E: grupo a partir de MFA, FS: fecha de siembre, AV: antecesor de verano, AI: antecesor de invierno, P: agregado de fósforo, S: agregado de azufre.

El agregado de los nutrientes fósforo y azufre apareció como una variable de importancia a la hora de determinar el rendimiento de soja (cuadro 5). En cuanto a la estimación de los promedios de rendimiento en kg de soja por ha, la única diferencia

significativa (al 5 %) se evidenció entre los grupos 1 y 2, los cuales habían sido clasificados como intensivo y diverso (G1) y de intensidad intermedia y poco diverso (G2) (cuadro 6).

Cuadro 6. Prueba de diferencia de medias para rendimiento de soja (kg/ha) según grupo

Grupo	Rend. medio	LI	LS	Prueba[+]
2	2186	1927	2446	a
1	2436	2175	2696	b
3	3236	1920	4553	ab

[+] Valores seguidos por diferente letra difieren estadísticamente ($p \leq 0,05$).

Cuadro 7. Contrastes entre fechas de siembra para rendimientos de soja.

Contraste	Estimación	t.ratio	p-valor
Temprana vs. media	-85,2	-1915	0,2219
Temprana vs. tardía	47,8	1	0,8249
Temprana vs. muy tardía	423,1	7367	< 0,0001
Media vs. tardía	133,0	2724	0,0329
Media vs. muy tardía	508,3	10008	< 0,0001
Tardía vs. muy tardía	375,3	8740	< 0,0001

De acuerdo a los contrastes para fechas de siembra (cuadro 7), sembrar después del 8 de diciembre genera pérdidas potenciales de rendimiento significativamente diferentes (al 5 %) con respecto al resto de las fechas, al igual que las fechas medias si se las compara con las tardías.

3.2.2 Modelo 2 de predicción de rendimiento

En este modelo se utilizaron las mismas variables que en el modelo 1 y también se asumieron varianzas homogéneas, tomando todos los efectos como aleatorios. El

objetivo de este procedimiento es estimar la contribución de las variables sobre la variabilidad total. Es un método de componentes de varianzas. No se busca la estimación de medias, sino la estimación del efecto de cada uno de los niveles de las variables medidas sobre la variabilidad de la variable dependiente y la predicción del efecto que se conoce como BLUP (mejor predictor lineal insesgado).

Cuadro 8. Estimación de componentes de varianza según variables independientes.

Param. Cov.	Estimación	% cov	% sin res
ngr3E	20527	4,1	13,9
Año	42843	8,5	29,0
AV	11230	2,2	7,6
AI	25669	5,1	17,3
FS	47551	9,4	32,1
P	18,8	0,0	0,0
S	120	0,0	0,1
Residual	358565	70,8	
Total	506524		
Total s/res	147959		

Param. Cov.: variables del modelo. % cov.: porcentaje de la variabilidad total aportado por cada variable. % sin res.: porcentaje de variabilidad aportado por cada variable sin tomar en cuenta la variabilidad del residual.

La varianza residual es un 70,8 % de la variabilidad total. Del 30 % que se relaciona con las variables, las que presentaron mayor contribución fueron la fecha de siembra, el año y el antecesor de invierno.

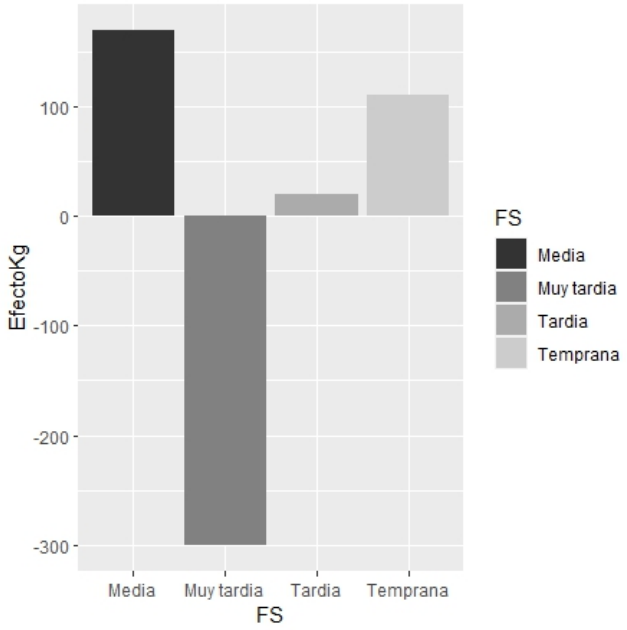

Figura 6. BLUP según fecha de siembra (FS). Estimaciones del modelo 2. Los BLUP indican ganancias o pérdidas potenciales de rendimiento respecto al promedio.

Una vez obtenidos los BLUP, se estimó una pérdida potencial de rendimiento de 300 kg de soja por sembrar en fechas tardías, más particularmente, luego del 8 de diciembre. Mientras que las siembras en fecha estuvieron cerca de los 170 kg por encima del promedio.

3.2.3 Modelos de aplicación: balance de nutrientes

Estos modelos de aplicación se plantearon ante las interrogantes de técnicos conocedores de las bases de datos con el objetivo de responder preguntas de interés agronómico, referidas al uso de agroquímicos; para lo cual fue necesario proponer alguna metodología además de la modelación. Se propuso modelar el balance en función del rendimiento y del agregado de cada uno de los nutrientes: nitrógeno (N), fósforo (P) y potasio (K), y determinar cuál presentó una mayor magnitud en la determinación del balance.

$$\hat{\beta}_{xstd} = \frac{s\,x}{s\,y}\,\hat{\beta}_x$$ (Ecuación 2)

Los coeficientes asociados al agregado (β_1) y al rendimiento (β_2) se encuentran en diferentes unidades, kg de nutriente y de grano de soja respectivamente, por lo cual se los estandarizó como se muestra en la ecuación 2.

Donde: $\hat{\beta}_{xstd}$ es el coeficiente de regresión estandarizado. $s\,x$ es el desvío de la variable independiente. $s\,y$ es el desvío de la variable respuesta y $\hat{\beta}_x$ es el coeficiente de regresión asociado a la variable independiente (beta 1 agregado y beta 2 rendimiento).

Cuadro 9. Coeficientes de regresión y de regresión estandarizados para rendimiento y agregado de nutrientes en los modelos de balances de nutrientes

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{xstd}$	$\hat{\beta}_{xstd}$	$\hat{\beta}_{xstd}\,/\hat{\beta}_{xstd}$
P	0,97	-0,012	1,02	-0,41	2,49
S	0,94	-0,002	0,98	-0,35	2,79
K	0,99	-0,023	0,89	-0,55	1,62

Se pudo establecer la relación entre ambos coeficientes y se observó que el agregado tiene un peso mayor en el balance respecto al rendimiento. Para P, por ejemplo, es dos veces y media mayor el peso de lo que se agrega respecto a lo que se lleva el cultivo en rendimiento, para K, más de 1,5 veces y, para S, la importancia en cantidad de lo que se agrega es casi tres veces mayor en relación con lo que el cultivo extrae en rendimiento (cuadro 9).

3.3 VALIDACIÓN

El proceso de validación se realizó utilizando los mismos procesos aplicados en el análisis de soja a la base de datos para trigo, que es un cultivo conocido. Las variables de interés que entran como independientes son las mismas (anexo 2);

entonces, dado que las variables que definen el rendimiento en trigo no son las mismas que en soja, se planteó la hipótesis de que los métodos aplicados en el análisis de soja pudiesen identificar grupos de rendimiento de trigo, así como las variables que lo definen.

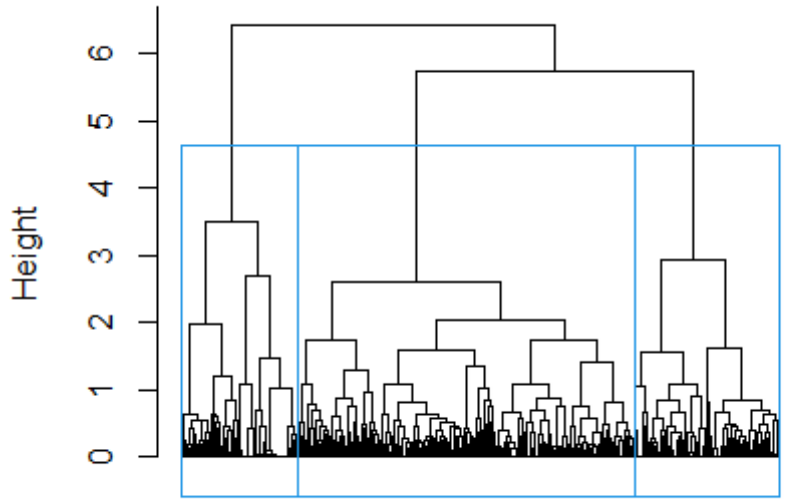

Figura 7. Dendrograma generado para las observaciones de trigo

En este caso, el agrupamiento a partir de la distancia de Gower mostró una mejor definición de los grupos y también se optó por tres grupos. La fecha de siembra para trigo se clasificó como: muy temprano (anteriores al 10 de mayo), temprana (del 10 al 30 de mayo), media (del 30 de mayo al 20 de junio) o tardía (del 20 de junio en adelante).

Cuadro 10. Valores de F y chi-cuadrado para las variables modeladas como dependientes del grupo para trigo.

Variables discretas	G	gl	G/gl	Variables continuas	F
Ciclo del cultivo	969	8	121	Entrada de N	353
				Entrada de P	67
Fecha de siembra	72	4	18	Número de años en agricultura	32
				Entrada de S	20
Antecesor verano	160	28	6	Entrada de K	13
Antecesor invierno	179	32	6	Distancia entre hileras	2

Ngr3E: número de grupo según MFA (distancia euclidiana). Gl: grados de libertad. G: valor asociado a la prueba chi-cuadrado.

En la determinación del agrupamiento para trigo, aplicando los métodos de ANAVA y prueba de razón de verosimilitud chi-cuadrado, se estableció que las variables discretas de mayor importancia fueron el ciclo del cultivo, la fecha de siembra y los antecesores. De las variables continuas, las más importantes fueron entrada de N y de P y años de agricultura.

3.3.1 Modelo de predicción de rendimiento para trigo

En el caso de trigo, se trabajó con un modelo lineal mixto de predicción de rendimiento, con las variables determinantes del agrupamiento establecidas en EL CUADRO 10.

$$R = \beta_0 + g_i + a_j + AV_k + AI_l + FS_m + CC_n + \beta_1 N_o + \beta_2 nF_p + \varepsilon_{ijklmnop}$$

Donde: R, rendimiento de trigo en kg/ha. G_i ,grupo. a_j, efecto del j-ésimo año. AV_k, efecto del k-ésimo antecesor de verano. AI_l, efecto del l-ésimo antecesor de invierno. FS_m, efecto de la m-ésima fecha de siembra. CC_n, efecto del n-ésimo ciclo del cultivo. β_1, coeficiente de regresión asociado al agregado de N. β_2, coeficiente de regresión asociado al número de fertilizantes aplicados y $\varepsilon_{ijklmnop}$, error experimental.

Cuadro 11. Análisis de la varianza para el modelo 3 aplicado en los datos de trigo.

	gl	F	p-valor
(Intercepto)	1	4244638	<0,0001
G	2	77124	0,0005
N	1	71072	0,0079
CC	3	27108	0,0444
FS	2	51908	0,0058
AI	15	47535	<0,0001
AV	12	25666	0,0026
NF	1	90620	0,0027

G: grupo; N: agregado de nitrógeno; CC: ciclo del cultivo; FS: fecha de siembra; AV: antecesor verano; AI: antecesor invierno; NF: número de fertilizantes aplicados.

Cuadro 12. Prueba de diferencia de medias para rendimiento de trigo por grupo

Grupo	Rendim?	LI	LS	Prueba[+]
3	1433	708	2159	a
1	2555	1171	3940	ab
2	2880	2220	3540	b

[+] Valores seguidos por diferente letra difieren estadísticamente ($p \leq 0,05$). LI: límite inferior del intervalo al 95 %. LS: límite superior del intervalo al 95 %.

4. DISCUSIÓN Y CONCLUSIÓN

Tanto el uso de la distancia de Gower como la reducción a coordenadas principales a partir de MFA y distancia euclidiana mostraron agrupamientos que se pudieron describir y clasificar desde el punto de vista estadístico y agronómico. Dependiendo del conocimiento biológico de las bases de datos, es válido optar por el uso de una u otra medida de distancia. Si bien la distancia de Gower trabaja con las variables originales (sin transformaciones en las mediciones) y es invariante ante el cambio de escalas (que puede hacerse con una estandarización en el caso euclidiano), los indicadores agronómicos (cuadro 1) muestran que el método de Ward a partir de coordenadas del MFA resultó más satisfactorio para caracterizar los agrupamientos referidos a la producción de soja.

Cuando se realizó la validación del método utilizando los datos de trigo, la distancia de Gower generó grupos bien definidos y que se pudieron caracterizar. Si bien con el método a partir de MFA se generaron grupos, la descripción agronómica no fue tan sólida como en el primer caso (las prácticas por grupos no establecían ningún patrón claro).

Es válido optar por uno u otro método de cálculo de distancias desde el punto de vista estadístico, el análisis agronómico permite una mejor interpretación de los resultados y la generación de posibles hipótesis de trabajo, y, a través de indicadores o del conocimiento biológico (agronómico, eventualmente social, dependiendo de la naturaleza de las bases de datos), optar por una u otra distancia.

En la determinación del agrupamiento, se pudo establecer que la prueba F para las variables continuas y la prueba chi-cuadrado de razón de verosimilitud (cuadros 3 y 10) resultaron satisfactorias a la hora de determinar el peso de las variables en el agrupamiento. Algunas mediciones que no fueron claras o que no aportaron información significativa por la forma en que se tomaron se debieron eliminar del análisis, las variables redundantes o poco claras fueron eliminadas en el proceso de acondicionamiento de las bases de datos. Las variables de mayor importancia (mayores valores de F o G/gl) fueron las que luego se tomaron en cuenta para la modelación y predicción del rendimiento. Debe considerarse la importancia de tener la información

completa (y no redundante) y que su recolección tenga objetivos claros para que pueda ser utilizada en el estudio.

Una vez obtenido el ranking de variables para los agrupamientos, se utilizaron en los modelos de predicción de rendimiento. A través del ANAVA se demostró que estas variables seleccionadas a partir de los valores de F o G (chi-cuadrado) tuvieron efecto sobre el rendimiento, esto es, que la racionalidad productiva de los grupos se vio reflejada en el rendimiento.

Los grupos 1 y 2 del agrupamiento de soja mostraron diferencias significativas en rendimiento. El mejor resultado lo obtuvo el grupo definido como diverso e intensivo en cuanto a las prácticas de manejo.

En cuanto a los modelos de predicción, si bien se estimaron coeficientes y efectos, no se debe perder de vista el gran peso que tiene sobre la variabilidad el agrupamiento. En la descomposición de la varianza, el grupo explica cerca del 70 % de esta; tal vez siguiendo líneas de investigación experimentales se podría establecer una cuantificación más precisa de los efectos de las variables en el rendimiento o bien estudiar métodos que cuantifiquen mejor el efecto del agrupamiento en la varianza total.

Si bien en el modelo de componentes de varianza el P y S aparecen con los valores mínimos de aporte a la variabilidad, fueron significativas en el modelo. Teniendo en cuenta que la base de datos va hasta el 2015, es importante, ya que hasta el momento no había muchos estudios sobre el S en soja.

Si se observan las prácticas incluidas en los modelos, se pudo determinar que la fecha de siembra tuvo un impacto en la determinación del rendimiento que se pudo cuantificar en 300 kg menos de rendimiento potencial por sembrar en fechas muy tardías (luego del 8 de diciembre). Vitantonio et al (2020) obtuvieron una estimación de 39 kg menos de rendimiento potencial por día a partir de siembras del 30 de octubre en Argentina.

A partir de la información del cuadro 9 de agregados de insumo y rendimiento de soja podría plantearse algunas hipótesis de investigación, o interrogantes, que tengan que ver con el agregado sustentable de nutrientes y cuánto se va en el agua de escurrimiento y sedimentos de suelo.

El conjunto de técnicas estudiadas cumplió con el objetivo de proponer una metodología de estudio que se adapte a las bases de datos de la producción agropecuaria y similares, donde se pretenda generar nuevas hipótesis de investigación (desde la estadística exploratoria) o para probar hipótesis planteadas a través de la modelación.

La dependencia de los resultados, incluyendo la validación de la metodología, respecto a los datos utilizados (dependientes de los datos) debe ser resaltada, puesto que se trata de un estudio empírico en lo que se relaciona con los resultados agronómicos. La utilización de dos rutas de acción (o métodos de análisis) desde luego es dependiente de los datos, pero parte de ella es la definición secuencial de las decisiones: (1) depuración de la información (que, en este caso, implicó una pérdida del 42 % de las observaciones, (2) utilización de dos medidas de distancia, las dos bien estudiadas en la literatura estadística, (3) utilización de un método jerárquico de agrupamiento apropiado y que podría ser reemplazado por algún otro si los resultados no parecen ser suficientemente convincentes para el investigador (y que no implicaría más que un cambio de algoritmo) y (4) pruebas estadísticas de razón de verosimilitud para la selección de los modelos de interpretación, particularmente en lo relacionado con la definición de los grupos como fijos o aleatorios que conducirá a extrapolaciones globales o por grupo.

Pensamos, entonces, que se pueden explorar estos caminos al enfrentar información de datos no obtenidos experimentalmente, pero que tienen una característica muy importante y es que son datos muy objetivos, recogidos sobre el camino productivo y para nada despreciables en la búsqueda de mejores formas o racionalidades de producción.

5. BIBLIOGRAFÍA

Amoroso Y, Costales D. 2016. Big Data: una herramienta para la administración pública. Ciencias de la Información, 47(3): 3-8.

Demey JR, Pla L, Vicente-Villardón J L, Di Rienzo JA, Casanoves F. 2011. Medidas de distancia y similitudes. En: Demey JR (Ed.). Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos. Turrialba: CATIE. (n.º 384). 47-59.

Di Rienzo JA, Casanoves F, González LA, Tablabda EM, Díaz MP, Robledo CW, Balzarini MG. 2005. Estadística para las ciencias agropecuarias. Córdoba: s. e. 107-186.

DIEA (Dirección de Estadísticas Agropecuarias). 2017. Producción [En línea]. En: Anuario estadístico agropecuario. Montevideo: MGAP (Ministerio de Ganadería, Agricultura y Pesca). Consultado el 20 diciembre de 2021. Disponible en: https://descargas.mgap.gub.uy/DIEA/Anuarios/Anuario2017/DIEA-Anuario2017.pdf

Ernst OR, Kemanian AR, Mazzilli SR, Cadenazzi M, Dogliotti S. 2016. Depressed attainable wheat yields under continuous annual no-till agriculture suggest declining soil productivity. Field Crops Research, 186: 107-116.

Escobedo MT, Salas JA. 2008. P. CH. Mahalanobis y las aplicaciones de su distancia estadística. Cultura Científica y Tecnológica, 5(27): 13-20.

Escofier B, Pagès, J. 1994. Multiple factor analysis (AFMULT package). Computational statistics & data analysis, 18(1): 121-140.

Franco J, Crossa J, Desphande S. 2010. Hierarchical multiplefactor analysis for classifying genotypes based on phenotypic and genetic data. Crop Science, 50(1): 105-117.

Guerra CW, Cabrera A, Fernández L. 2003. Criterios para la selección de modelos estadísticos en la investigación científica. Revista Cubana de Ciencia Agrícola, 37(1): 3-10.

Gower JC. 1971. A general coefficient of similarity and some of its properties.

Biometrics 27:857-874

Le Dien S, Pagès J. 2003. Hierarchical Multiple Factor Analysis: Application to the comparison of sensory profiles. Food Quality and Preference, (5-6): 397-403.

Lindsey JK. 1995. Modelling Frequency and Count Data. Oxford University Press, Oxford.

Marcillo J. 2017. Análisis espaciotemporal multivariante de las valoraciones de las edificaciones en la ciudad de Cuenca (Ecuador). Tesis de maestría. Barcelona, España. Universitat Politècnica de Catalunya. 131 p.

Mazzilli SR, Ernst OR, De Mello VP, Pérez CA. 2016. Yield losses on wheat crops associated to the previous winter crop: Impact of agronomic practices based on on-farm analysis. European Journal of Agronomy, 75: 99-104.

Mazzilli SR, Ernst OR. 2019. Rapeseed-to-Wheat Yield Ratio in Different Production Environments and Effects on Subsequent Summer Crops Yields. Agrosystems Geosciences & Environment, 2: 1-4.

Mazzilli S. 2018. Sustentabilidad ambiental y económica en predios agrícola-ganaderos: Un sistema de indicadores objetivos aplicable en el campo. Montevideo: INIA. (Serie técnica n.º 65). 1-70.

Pagès J. 2002. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. Revue de Statistique Appliquée, 50(4): 5-37.

Tittonell P. 2014. Ecological intensification of agriculture—sustainable by nature. Current Opinion in Environmental Sustainability, 8: 53-61. https://doi.org/10.1016/j.cosust.2014.08.006

Urruty N, Guyomard H, Tailliez-Lefebvre D, Huyghe C. 2017. Variability of winter wheat yield in France under average and unfavourable weather conditions. Field Crops Research, 213: 29-37.

Wilkinson L. 1994. SYSAT for DOS, Advanced Applications. American Statistical Association, 6: 75-77.

6. ANEXOS

6.1 ANEXO 1: consistencia de los valores de f para todos los análisis de varianza de variables respuesta según grupo, para: ngr: número de grupo, numero (3, 4 o 5), g: gower, e: euclidiana, s: con datos imputados

F value	**Variable**					
Grupo	Rend.	Biomasa de residuos	Balance de N	Balance de P	Balance de K	Balance de S
ngr3G	55,8	58,0	39,8	37,1	0,6	1,4
ngr4G	41,1	43,8	28,9	25,0	2,0	1,0
ngr5G	37,8	38,4	28,0	18,8	2,7	1,7
ngr3E	89,8	73,0	83,0	80,1	14,5	5,5
ngr4E	73,3	61,4	63,9	54,0	9,7	3,7
ngr5E	55,6	47,0	47,9	43,1	7,6	2,8
ngr3S	36,7	28,2	32,0	33,0	6,2	15,5
ngr4S	28,3	22,4	24,1	24,7	6,3	11,8
ngr5S	27,2	21,2	24,1	20,3	5,1	11,6

6.2 ANEXO 2: título de las variables bajo estudio y recomendación de referencia, simplificación utilizada (necesaria para facilitar el análisis)

Variable independiente	Referencia	Variable respuesta
Antecesor verano	d7	Rendimiento (0 %)
Antecesor invierno	d8	Biomasa áerea (Mg ha-1)
Sistema de siembra	d9	Biomasa raíces (Mg ha-1)
Número de años con agricultura	d10	Biomasa total corregido (s/grano)
Número de cultivos en SD	d11	Biomasa grano corregida
Número de cultivos totales	d12	1- PRODUCTIVIDAD
Aplicación de herbicida (sí, no)	d13	2- MASA DE RESIDUOS
Barbecho: químico o laboreo	d14	3- USO DEL AGUA
Fecha de siembra (en días desde el 1 de enero de cada año)	d15	4- ENTRADAS C AL SUELO
Material (variedad) pasará a grupo de madurez	d16	Entrada N al suelo

Kg/ha de semilla	d17	Salidas de N al suelo
Distancia entre hilera	d18	5- BALANCE APARENTE DE N
Unidades toxicológicas mamíferos herbicidas presiembra	d46	Entrada P2O5 al suelo
Unidades toxicológicas mamíferos herbicidas posiembra	d47	Salidas de P2O5 al suelo
Unidades toxicológicas abejas herbicidas presiembra	d48	6- BALANCE APARENTE DE P2O5
Unidades toxicológicas abejas herbicidas posiembra	d49	Entrada K al suelo
Numero de fertilizantes aplicados	d50	Salidas de K al suelo
Numero de insecticidas aplicados	d53	7- BALANCE APARENTE DE K
Numero de fungicidas aplicados	d56	Entrada S al suelo
	v28	Salidas de S al suelo
	v29	8- BALANCE APARENTE DE S
	v30	UT MAM - Herbicidas
	v31	UT MAM - Fungicidas
	v32	UT MAM - Insecticidas
	v33	9- UT MAMÍFEROS
	v34	UT ABJ - Herbicidas
	v35	UT ABJ - Fungicidas
	v36	UT ABJ - Insecticidas
	v37	10- UT ABEJAS

6.3 ANEXO 3: *script* de agrupamiento

```
require(funModeling); require(missMDA); require(FactoMineR)
require(fastcluster); require(graphics); require(dplyr); require(cluster)
ts<-read.csv2(file="tg.csv", header=T,stringsAsFactors = F, dec = ".")
df_status(ts) ##esta va para soja con SOJAINDDEP##
attach(ts)
ts$d7 <- as.factor(d7)
ts$d8 <- as.factor(d8)
ts$d9 <- as.factor(d9)
```

```
ts$d13 <- as.factor(d13)
ts$d14 <- as.factor(d14)
ts$d16a <- as.factor(d16a)
ts$d15a <-  as.factor(d15a)
df_status(ts)

# 7 factores y 16 continuas, a ts2
ts2 <- (ts[,c(7,8,9,13,14,16,18,10,15,19,20,28,30,32,33,34,42,36:39,48,49)])
df_status(ts2)

#### proceso gower "incompleto"
dgowd <- daisy(ts2, metric="gower")
dgowm <- as.matrix(dgowd); dim(dgowm)
hcpc <- hclust(dgowd, method="ward.D2")
save(hcpc, file="hcpcgow.RData")
plot(hcpc, labels=F, hang=-1, xlab="", main="Ward.D2 clustering gow" )
rect.hclust(hcpc, k=3, border=4)       ## ojo grupos en la figura
rect.hclust(hcpc, k=5, border=10)       ## ojo grupos en la figura
rect.hclust(hcpc, k=6, border=3)
gro <- c(1:nrow(dgowm))
for(i in 2:10){
  pp <- cutree(hcpc, k=i)
  gro <- data.frame(gro,pp)
}
str(gro)
colnames(gro)=c("nobs",paste("ngr",2:10, sep=""))
save(gro, file="gowgr2a10.RData")
write.csv(gro,"gowgr2a10.csv",row.names=F)
```

```
sum(is.na(dgowm))
diag(dgowm) <- NA
sum(is.na(dgowm))
sum(dgowm==0, na.rm=T)   ## 1192/2 = 596 observaciones con distancia = 0.0
save(dgowm,file="dgowm.RData")
save(dgowd,file="dgowd.RData")
cerosdgow <- which(dgowm==0, arr.ind=T)
write.csv(cerosdgow,"cerosdgow.csv")

### proceso MFA incompleto

mfa1 <- MFA(ts2, group=c(7,16), type = c("n","s"), ncp=100)
#print(mfa1, file="mfa1.txt")
save(mfa1, file="mfa1.RData")
load("mfa1.RData")
eigen <- mfa1$global.pca$eig; dim(eigen); eigen
coord <- mfa1$global.pca$ind$coord; dim(coord)
write.csv(coord,"coord.csv")
vcontrib <- mfa1$global.pca$var$contrib; dim(vcontrib)
write.csv(vcontrib,"vcontrib.csv")
vcorr <- mfa1$global.pca$var$cor; dim(vcorr)
write.csv(vcorr,"vcorr.csv")
#sink()
c0 <- read.csv("coord.csv"); coord <- c0[,-1]; dim(coord)
head(colnames(coord),10)
max(coord) ; min(coord)
deud <- daisy(coord[,1:40], metric= "euclidean")## 40 coord 90% varianza tot
```

```
sum(is.na(deud))
sum(deud==0)
deum <- as.matrix(deud) ; dim(deum) ## pedir las distancias
#1702*(1701/2)
length(deud)
x2 <- eigen(1-deum, symmetric=T, only.values = T)
summary(x2$values)
negativos <- x2$values[x2$values<0]; length(negativos) ##ojo cuando le de NA porque me dice que tiene missing
hist(deud, main= "Distribucion de distancias euclidianas")
sum(is.na(deum))
diag(deum) <- NA; sum(is.na(deum))
sum(deum==0, na.rm=T) ##hay 1086/2= 543 distancias que son 0 que no son diagonales(es un vector)

cerosdeuminc <- which(deum==0,arr.ind = T)   ##########################
write.csv(cerosdeuminc,"cerosdeuminc.csv")
hcpc <- hclust(deud, method="ward.D2")
save(hcpc, file="hcpcdeu.RData")
plot(hcpc, labels=F, hang=-1, xlab="", main="Ward.D2 clustering deu" )
rect.hclust(hcpc, k=3, border=4)      ## ojo grupos en la figura
rect.hclust(hcpc, k=5, border=3)      ## ojo grupos en la figura
gro <- c(1:nrow(deum))
for(i in 2:20){
  pp <- cutree(hcpc, k=i)
  gro <- data.frame(gro,pp)
}
str(gro)
```

```
colnames(gro)=c("nobs",paste("ngr",2:20, sep=""))
save(gro, file="deugr2a20.RData")
write.csv(gro,"deugr2a20.csv",row.names=F)

mdeu <- colMeans(deum, na.rm=T)
summary(mdeu)
hist(mdeu)
length(mdeu[mdeu >=7])
bigmdeu <- which(mdeu >= 7, arr.ind=T)
write.csv(bigmdeu,"bigmdeu.csv")

### imputacion FAMD para las variables dependientes
require(missMDA)
imputs2 <- imputeFAMD(ts2, ncp = 10)
length(imputs2)
dim(imputs2$tab.disj); colnames(imputs2$tab.disj)
dim(imputs2$completeObs); colnames(imputs2$completeObs)
### tabla con valores imputados y FAMD sobre la tabla
ts3 <- imputs2$completeObs; dim(ts3)
df_status(ts3)
save(ts3,file="ts3.RData")

famd3 <- FAMD(ts3, ncp=50)
save(famd3, file="famd3.RData") ###un 13% de valores missing en la ts2.
## contribucion de las variables
contfamd3 <- famd3$var$contrib; dim(contfamd3)
write.csv(contfamd3,"contfamd3.csv")
coord <- famd3$ind$coord; dim(coord)
```

```
max(coord) ; min(coord)
write.csv(coord,"coordfamd3.csv")

raices <- famd3$eig; raices   ## 33 explican 92.3 %
deud <- daisy(coord[,1:40], metric= "euclidean")
sum(is.na(deud))
sum(deud==0)
deum <- as.matrix(deud) ; dim(deum) ## pedir las distancias
#1702*(1701/2)
length(deud)
x2 <- eigen(1-deum, symmetric=T, only.values = T)
summary(x2$values)
x3 <- x2$values[x2$values<0]; length(x3); sum(x3); summary(x3)
### un solo eigen < 0.0
sum(x2$values<0) ; length(x2$values); sum(x2$values<0)/length(x2$values)
hist(deud, main= "Distribucion de distancias euclidianas")
qdeud <- quantile(deud, probs=seq(0,1,0.05))
hist(deud[deud<=16], main= "Distribucion de distancias euclidianas <= 16")
### multidimensional scaling
require(smacof)
deudsmac <- smacofSym(deud, ndim=3)
names(deudsmac)
coordsmac <- deudsmac$conf;head(coordsmac)
write.csv(coordsmac,file="coordsmac.csv")
sum(is.na(deum))
diag(deum) <- NA; sum(is.na(deum))
sum(deum==0, na.rm=T) ## hay 1086/2 = 543 distancias = 0.0
cerosdeumimp <- which(deum==0, arr.ind = T)
```

```
write.csv(cerosdeumimp, "cerosdeumimp.csv")
### media de distancia por observacion
mdist <- colMeans(deum,na.rm=T); length(mdist); summary(mdist)
sum(mdist > 11); bigmdist <- which(mdist > 11, arr.ind = T)
hist(deum[-bigmdist,-bigmdist])
write.csv(bigmdist,"bigmdistIMPU.csv")
deum11 <- deum[-bigmdist,-bigmdist]
deud11 <- as.dist(deum11)
obsSel <- colnames(deum11)

ts4sel <- ts3[-bigmdist,]; dim(ts4sel)
ts4del <- ts3[bigmdist,]; dim(ts4del)
save(ts4sel,file="ts4sel.RData")
save(ts4del,file="ts4del.RData")

### cluster analysis con datos imputados y seleccionados
require(fastcluster)
hcpc <- hclust(deud11, method="ward.D2")
save(hcpc, file="hcpc11.RData")
# load("hcpc11.RData")
plot(hcpc, labels=F, hang=-1, xlab="", main="Ward.D2 clustering eud, 1574 obs" )
rect.hclust(hcpc, k=3, border=4)      ## ojo grupos en la figura
rect.hclust(hcpc, k=4, border=3)      ## ojo grupos en la figura
rect.hclust(hcpc, k=5, border=5)      ## ojo grupos en la figura
summary(hcpc)
str(hcpc)
gro <- colnames(deum11)
for(i in 2:10){
```

```
  pp <- cutree(hcpc, k=i)
  gro <- data.frame(gro,pp)
}
dim(gro)
colnames(gro)=c("nobs",paste("ngr",2:10, sep=""))
save(gro, file="gr2a10_1574obs.RData")
write.csv(gro,"gr2a10_1574obs.csv",row.names=F)

write.csv(ts4del,"ts4del.csv")
write.csv(ts4sel,"ts4sel.csv")

### junta archivos
dim(ts3)
write.csv(ts3,"ts3.csv")

### cluster analysis con datos imputados y sin seleccion

load(file="famd3.RData"); names(famd3)
coord <- famd3$ind$coord; dim(coord)
deuimp <- daisy(coord, metric = "euclidean")
hcpc <- hclust(deuimp, method="ward.D2")
save(hcpc, file="hcpcImp.RData")
# load("hcpc11.RData")
plot(hcpc, labels=F, hang=-1, xlab="", main="Ward.D2 clustering eud, imputados"
)
rect.hclust(hcpc, k=3, border=4)      ## ojo grupos en la figura
rect.hclust(hcpc, k=4, border=3)      ## ojo grupos en la figura
rect.hclust(hcpc, k=5, border=2)      ## ojo grupos en la figura
```

```
summary(hcpc)
str(hcpc)
gro <- hcpc$labels
for(i in 2:20){
  pp <- cutree(hcpc, k=i)
  gro <- data.frame(gro,pp)
}
dim(gro)
colnames(gro)=c("nobs",paste("ngr",2:20, sep=""))
save(gro, file="gr2a10_imp_allobs.RData")
write.csv(gro,"gr2a10_imp_allobs.csv",row.names=F)
  ### el grupo 4 tiene solo 13 obs que pueden pasarse al 5 o sacarlas
which(gro$ngr5 == 4, arr.ind=T)
### obs: 38  203  206  530  569  589  638  911  954  960 1003 1009 1358

tds3 <- read.csv("ts3DEP-INDEP-GRUP-DIM.csv")
save(tds3,file="tds3.RData")
```

6.4 ANEXO 4: MINERÍA DE DATOS CON BASES DE DATOS DE LA PRODUCCIÓN AGROPECUARIA. ALGUNAS ESTRATEGIAS DE ANÁLISIS

Lucía Ferreira[1], Jorge Franco[2], Sebastián Mazzilli[3], Oswaldo Ernst[4]

[1] Ing. Agr., Profesor Asistente, Depto. de Biometría, Estadística y Computación, EEMAC. luciaf@fagro.edu.uy

[2] Ing. Agr., Profesor Agregado, Depto. de Biometría, Estadística y Computación, EEMAC. jfranco@fagro.edu.uy

[3] Ing. Agr., Profesor Asistente, Depto. de Producción Vegetal, EEMAC. smazzilli@fagro.edu.uy

[4] Ing. Agr., Profesor Titular, Depto. Producción Vegetal EEMAC, EEMAC. oernst@fagro.edu.uy

RESUMEN

El registro de datos obtenidos directamente por técnicos y productores avanzados en el proceso de la producción agropecuaria ha crecido sustancialmente (DIEA, 2017). El volumen de información generado requiere de métodos estadísticos de análisis que aseguren un buen tratamiento de los datos y que permitan alguna medida de confiabilidad de los resultados obtenidos. Los objetivos de este trabajo fueron proponer y comparar métodos de análisis estadístico aplicable a bases de datos de este tipo (no experimentales) con el fin de identificar diferentes racionalidades en la producción. Se utilizó el método de agrupamiento de mínima varianza entre grupos de Ward con matrices de distancia de Gower y euclidiana a partir de las coordenadas principales resultantes de un análisis factorial múltiple (MFA, por sus siglas en inglés). El agrupamiento a partir de las distancias euclidianas mostró mayor consistencia en la diferenciación de las prácticas de manejo para el cultivo de soja. Se establecieron tres grupos de racionalidad productiva (formas de combinar insumos y prácticas agrícolas para lograr un resultado), dos de ellos de importancia agronómica: el grupo 1, caracterizado por prácticas intensivas y rotaciones diversas, el grupo 2, por una menor intensidad de uso del suelo y una mayor homogeneidad en las rotaciones. Con el fin de validar la metodología, los resultados fueron aplicados a una base de datos de trigo. Se pudieron diferenciar grupos, y las variables que los discriminaron fueron, en efecto, las ya conocidas o esperadas para modelar el rendimiento en este cultivo.

Palabras clave: análisis de conglomerados, distancias, métodos de agrupamiento, registros de la producción agropecuaria, soja.

DATA MINING WITH DATABASES OF AGRICULTURAL PRODUCTION. SOME ANALYSIS STRATEGIES

The recording of data obtained directly by advanced technicians and producers in the agricultural production process has grown substantially (DIEA, 2017). The volume of information generated requires statistical analysis methods that ensure good treatment of the data and that allow some measure of reliability of the results obtained. The objectives of this work were to propose and compare methods of statistical analysis applicable to databases of this type in order to identify different rationales in production. The minimum variance clustering method between Ward groups with Gower and Euclidean distance matrices from the principal coordinates resulting from a multiple factorial analysis (MFA) was used. The grouping based on Euclidean distances showed greater consistency in the differentiation of management practices. Three groups of productive rationality were established (ways of combining inputs and agricultural practices to achieve a result), two of them of agronomic importance: group 1, characterized by intensive practices and diverse rotations, group 2, by a lower intensity of use of the soil and a greater homogeneity in the rotations. In order to validate the methodology, the results were applied to a wheat database. Groups could be differentiated, and the variables that discriminated them were, in effect, those already known or expected to model the yield in this crop.

Keywords: agricultural production records, cluster analysis, distances, soybean

INTRODUCCIÓN

El análisis del registro de información obtenida por técnicos y productores en los procesos productivos no es algo nuevo en el sector agropecuario. No obstante, para el sector agrícola, debido a la acelerada expansión que ha tenido en los últimos años (DIEA, 2017), se produjo un aumento significativo de la información disponible. En la actualidad, el tratamiento masivo de datos, dado el acelerado desarrollo de las tecnologías de la información y la comunicación, ha dado lugar a nuevos conceptos y paradigmas en la gestión de la información: minería de datos y *big data* (Amoroso y Costales, 2016). La necesidad, además, de generar una producción cada vez más eficiente y respetuosa del ambiente, lo que se conoce como intensificación ecológica de la agricultura (Tittonell, 2014), ha llevado a que las instituciones públicas, federaciones y empresas privadas registren cada vez más información en la búsqueda de la optimización de resultados productivos. En Uruguay, el registro de información de este tipo ha sido utilizado principalmente por el Plan Agropecuario, la Oficina de Estadísticas Agropecuarias, la Oficina de Estadísticas Agropecuarias (DIEA), Federación Uruguaya de Grupos CREA (FUCREA) y algunas empresas privadas.

Las bases de datos generadas en una serie de años en las unidades productivas contienen información relevada por los técnicos en el mismo proceso productivo y son, por tanto, datos objetivos y confiables, relacionados a la toma de decisiones, el uso de recursos y los paquetes tecnológicos utilizados. Generalmente, el uso que se le ha dado a esta información en el país es de tipo descriptivo-explicativo; un ejemplo de esto es el Programa de Monitoreo de Empresas Ganaderas del Plan Agropecuario que se hace a partir de la información recolectada en las llamadas Carpetas Verdes.

La confiabilidad predictiva de este tipo de información no experimental requiere de un análisis más cuidadoso, en términos estadísticos, que permita la generación de hipótesis y predicción de efectos para ser utilizada, posteriormente, en programas de investigación agronómica.

Aunque existe un gran volumen de información y que se han aplicado diversas técnicas de análisis, no se ha estudiado un método de análisis exploratorio que asegure un buen tratamiento de los datos, ni se ha estudiado la repetitividad de los

métodos aplicados para el análisis exploratorio o para las predicciones (modelación). Uno de los problemas de aplicar métodos de análisis estadístico a este tipo de información es que no ha sido obtenida bajo condiciones controladas (experimentales), lo que limita la validez de la extrapolación de los resultados: «A partir de datos no experimentales no se pueden establecer relaciones de causalidad, solamente se pueden establecer relaciones de tipo descriptivo entre las variables» (Lindsey, 1995).

El objetivo de este trabajo es estudiar, proponer y validar métodos de análisis multivariado que faciliten la exploración y utilización de estas bases de datos en la generación de nuevas líneas de investigación. La hipótesis de trabajo es que existen grupos de diferentes racionalidades en la combinación de los factores de producción que inducen diferentes resultados.

MATERIALES Y MÉTODOS

Datos: se trabajó con bases de datos (BD) provenientes de registros de manejo de cultivos y rendimientos de productores integrantes de la Federación Uruguaya de Centros Regionales de Experimentación Agropecuaria (FUCREA). A partir de dos bases de datos que contienen registros de resultados productivos y prácticas de manejo de ocho años (2006-2014) para 28 predios, se resumió la información de acuerdo a las prácticas de manejo más relevantes desde el punto de vista técnico. Esta base contiene 5862 unidades de información (UI), cada una de ellas es la combinación de establecimiento, cultivo, potrero, año y zafra y tiene asociadas variables de manejo que deberían explicar los resultados productivos si la hipótesis de trabajo resultara ser cierta.

Se realizó un proceso de depuración/selección en dos etapas: control de la calidad de las observaciones y eliminación de variables que presentaron más de 50 % de valores faltantes. Este proceso condujo a una BD con 2472 observaciones y 35 variables. Las observaciones corresponden a los cultivos de soja y trigo, 1716 y 757, respectivamente. De las 35 variables, 6 son identificadoras, 19 de manejo (variables independientes, de naturaleza discreta y continua) y 10 de resultados productivos (dependientes). La información proveniente del cultivo de soja se utilizó en el proceso

de estudio y generación de metodologías de análisis y, la del cultivo de trigo, para su validación.

Métodos: los métodos de clasificación deben adaptarse al tipo de variables en el proceso, que en este caso es una mezcla de variables continuas y discretas. Para la búsqueda de grupos de racionalidades productivas (el agrupamiento de las UI) se propusieron dos estrategias de clasificación, a partir de las cuales se generaron grupos de observaciones (*cluster analysis*). La propuesta de este trabajo fue utilizar un método jerárquico de agrupamiento a partir del cálculo de distancias entre las observaciones. Para este agrupamiento hay, por lo menos, tres propuestas: El método de mezcla de máxima verosimilitud se ha utilizado para análisis multivariado de variables normales provenientes de mezcla de poblaciones y que a menudo involucran variables de tipo categóricas como es este caso. Lawrence y Krzanowski (1996) propusieron trabajar con el modelo gaussiano condicional homogéneo (LM, *Location Model*). En 1998, Franco *et al.* propusieron el modelo de localización modificado (MLM, por sus siglas en inglés). En el 2010, Franco, Crossa y Desphande trabajaron con el análisis múltiple de factores (MFA, por sus siglas en inglés) para clasificar genotipos basados en caracteres fenotípicos y genéticos.

En este trabajo se compararon dos estrategias: la primera consistió en calcular directamente una distancia entre las observaciones (UI) utilizando todas las variables (discretas y continuas) y, posteriormente, el análisis de conglomerados a partir de estas distancias; esta primera estrategia ofrece cierta facilidad en la manipulación de los datos ya que trabaja con las variables originales. De las tres propuestas para el agrupamiento, se optó por utilizar el MFA (Le Dien y Pagès, 2003). Esta consiste en generar nuevas coordenadas para cada observación en un sistema de variables continuas (sistema de coordenadas principales) de dimensiones reducidas, calcular una distancia utilizando estas nuevas coordenadas y, finalmente, realizar el agrupamiento basado en la última distancia calculada. La distancia utilizada en la primera estrategia permite utilizar variables continuas ordinales, nominales y binomiales para su cálculo y fue propuesta por Gower (1971); en la segunda estrategia se utilizó la distancia euclidiana basada en las coordenadas principales, producto del análisis de factores múltiples (MFA).

La distancia de Gower (1971) opera con mezcla de variables, es decir, continuas, nominales y ordinales. Es un coeficiente de distancia definida entre 0 y 1 y cumple las propiedades de una métrica euclidiana siempre que en los datos originales no haya (o sean muy pocos) valores faltantes.

$$d(_i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \quad , 0 < d_{ij} < 1 \qquad (1)$$

El indicador $\delta_{i\,j}^{(f)}$ toma valor 1 cuando ambas medidas x_{if} y x_{jf} para la *f- ésima* variable son valores no faltantes; de otra forma, toma el valor 0. El número $d_{i\,j}^{(f)}$ es la contribución de la *f- ésima* variable a la distancia entre i y j. Si la variable f es binaria o nominal, entonces $d_{i\,j}^{(f)}$ es definida como:

$$d_{ij}^{(f)} = 0 \; si \; x_{if} \neq \; x_{jf} = 0 \; si \; x_{if} = \; x_{jf}$$

Si la variable es nominal, la expresión (1) se convierte en el número de coincidencias entre el número total de posibles pares y coincide con el coeficiente de *simple matching*, proporción de desacuerdos. Si la variable es cuantitativa de intervalo, entonces $d_{ij}^{(f)}$ está dado por:

$$d_{ij}^{(f)} = \frac{\left| x_{if} - \; x_{jf} \right|}{R_f}$$

Donde R_f es el rango de la variable f (Gower, 1971).

La segunda alternativa es un método propuesto para mezclar variables de diferentes tipos donde se generan nuevas coordenadas en una escala continua con mínima (o ninguna) pérdidas de información. Esta transformación se logra con el análisis de múltiples factores o análisis factorial múltiple (MFA, por sus siglas en inglés). El MFA es una técnica de análisis estadístico multivariado que se aplica sobre un conjunto de datos conformados por grupos de variables de diferente naturaleza; fue propuesto por Escofier y Pagès (2008).

En principio, el MFA se utilizó para grupos de variables continuas de diferente origen y con grandes diferencias en el número de variables por grupo, en cuyo caso

dominarían en el PCA aquellos grupos de variables en los que haya número más elevado de ellas; posteriormente se generalizó a la mezcla de diferentes tipos de variable. El método, en general, consiste en aplicar metodologías de reducción de variables para expresar las observaciones en un sistema de coordenadas denominadas coordenadas principales. Se realiza un PCA para cada grupo de variables continuas, un análisis múltiple de correspondencia (MCA, por sus siglas en inglés) para los grupos de variables categóricas (incluyendo las binarias) y un análisis de correspondencia (CA, por sus siglas en inglés) para variables de frecuencia, expresadas como proporciones. Se obtienen los autovalores de cada grupo de variables y se estandariza la matriz dividiendo entre este para obtener *scores* o coordenadas a partir de las cuales se realiza PCA.

Una vez obtenidas estas coordenadas, se realizó el análisis de conglomerados utilizando la matriz de distancias obtenidas a partir del número de coordenadas (que son continuas y están balanceadas) que explicaban el 90 % de la variabilidad (31 coordenadas principales). Se utilizó la distancia euclidiana entre dos puntos $A(x_1, y_1)$ y $B(x_2, y_2)$, que se define como la longitud del segmento de la recta que tiene por extremos a los puntos A y B y se expresa como:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Para el agrupamiento se utilizó el método de mínima varianza dentro de grupos de Ward (1968) con cada una de las matrices de distancia: Gower con los datos originales, y euclidiana con las coordenadas principales a partir del MFA.

Para determinar el número de conglomerados, se optó por una de las técnicas posibles, que consiste en graficar la suma de cuadrados dentro de grupos y/o los valores del estadístico pseudo-F (denominado así porque se calcula como la relación de varianzas entre/dentro de grupos) según el número de grupos. El número óptimo se establece como el valor en el cual la variación dentro de los grupos y/o la pseudo-F se estabilizan.

Para la interpretación agronómica de los grupos se estudió la importancia de las variables originales en la estructura de los agrupamientos. En el caso de las variables

continuas se utilizaron los valores F (de Fisher) provenientes del análisis de varianza, ANAVA, para cada una de ellas y, para las variables discretas, una prueba de razón de verosimilitud de chi-cuadrado (G test). Cada una de las variables se modeló en función del agrupamiento (variables continuas) o con una tabla de doble entrada de tipo grupo por categoría de la variable (variables categóricas). Para balancear la comparación de los efectos de las variables categóricas en el agrupamiento, el valor de G se dividió entre los grados de libertad de la prueba.

Finalmente, a través de modelos lineales mixtos, se modeló la relación de rendimiento con las variables obtenidas de las pruebas de anova y de chi- cuadrado para probar que si existe un efecto significativo de las mismas, entonces los grupos se diferencian por prácticas de manejo que resultan en rendimientos diferentes. Se utilizó un modelo de efectos aleatorios (2), dado la ventaja de que puede ser extrapolable a otras situaciones (a diferencia de los modelos de efectos fijos).

$$R = \beta_0 + g_i + \alpha_j + av_k + ai_l + fs_m + \beta_1 p_n + \beta_2 S_o + \varepsilon_{ijklmno} \quad (2)$$

Donde: R =rendimiento de soja en kg/ha. g_i = grupo según MFA. a_j = efecto del j-ésimo año. av_k = efecto del k-ésimo antecesor de verano. ai_l =efecto del l-ésimo antecesor de invierno. fs_m = efecto de la m-ésima fecha de siembra. β_1 = coeficiente de regresión asociado al agregado de P. β_2 = coeficiente de regresión asociado al agregado de S. $\varepsilon_{ijklmno}$ = efectos residuales al modelo.

RESULTADOS Y DISCUSIÓN

Aplicación de metodología en soja: tanto el uso de la distancia de Gower como la reducción a coordenadas principales a partir de MFA y distancia euclidiana mostraron agrupamientos que se pudieron describir y clasificar desde el punto de vista estadístico y agronómico.

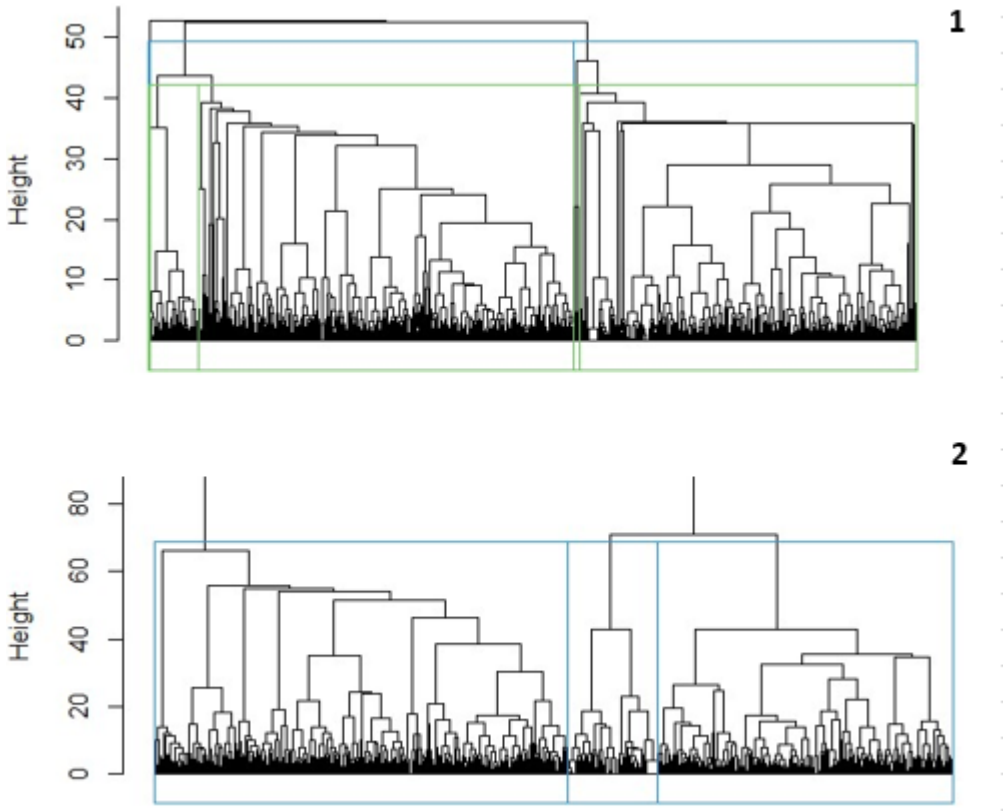

Figura 1: dendrograma con distancias calculadas a partir de MFA completo (1) (todas las observaciones) vs. con eliminación de distancias mayores a 11 (2).

Dependiendo del conocimiento técnico de las bases de datos es conveniente utilizar los dos caminos y comparar para optar por una de las opciones de medida de distancia. Si bien la distancia de Gower trabaja con las variables originales y es invariante ante el cambio de escalas, el uso de la distancia euclidiana a partir del MFA produjo resultados más consistentes para caracterizar los agrupamientos referidos a la producción de soja.

En el proceso que utilizó el MFA como método de reducción de variables a coordenadas, la distribución de las distancias fue muy asimétrica hacia la derecha (figura 2). Las distancias se encontraron entre 0 y 60, un 90 % de estas con valores menores a 11. Las UI que presentaron distancias promedio mayores a 11 respecto a todas las demás se eliminaron por considerarse atípicas a la población en estudio (valores mal ingresados por errores de operador). La figura 3 presenta la distribución de las distancias antes y después de remover este 10 %. Se evidenció una mejora en la

distribución y en la estructura del agrupamiento. Este procedimiento se realiza directamente luego del MFA, al indicar que se trabaje con las distancias menores a determinado valor; no es necesario realizar todo el procedimiento desde el inicio. Se suman las distancias mayores a 11 y se las guarda en un archivo (función sum, distancias *sum(mdist > 11); bigmdist <- which(mdist > 11, arr.ind = T*) y luego se eliminan de la matriz de distancias que se generó en el paso anterior (*deum11 <- deum[-bigmdist,-bigmdist]*) para volver a generar el histograma y el análisis de conglomerados a partir de la nueva matriz.

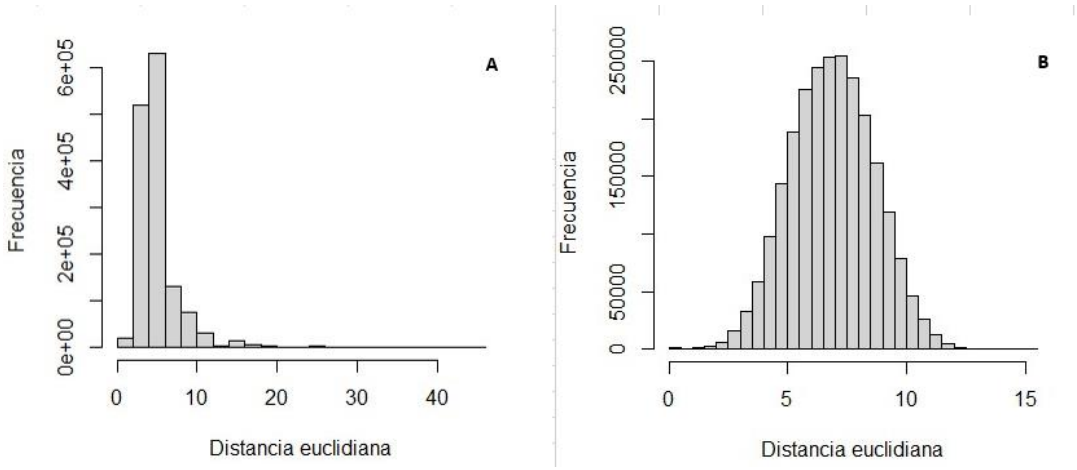

Figura 2: distribución de distancias euclidianas a partir de MFA con todas las distancias (A) vs. distancias menores a 11 (B).

De acuerdo con el resultado obtenido gráficamente de los valores de pseudo-F según el número de grupos, se determinó que el número óptimo de grupos para trabajar fue tres. Se calcularon las proporciones (tabla 1) de los niveles de las variables que resultaron importantes en la determinación de los grupos para ver si efectivamente la clasificación estadística fue concordante con la descripción agronómica. Los indicadores fueron propuestos por técnicos con conocimiento del cultivo, para caracterizar a los grupos de forma genérica a partir de las variables más relevantes de acuerdo a las prácticas de manejo.

Tabla 1. Distribución de las categorías para antecesor de verano, ciclo del cultivo y fecha de siembra (FS) dentro de cada grupo.

Antecesor verano	**Gr1**	**Gr2**	**Gr3**
% gramíneas	38,8	1,8	0
% soja segunda	48,2	26,6	0
% P + CN	5,1	11,9	100
% valores faltantes	3,5	20,2	0
Ciclo cultivo			
% corto	15,2	9,2	0
% medio	74,1	76,6	76,9
% largo	7,5	0,2	0
FS			
% temprana	49,8	8,2	23,1
% regular	25,9	22,6	46,2
% tardía	16,6	32	23,1
% muy tardía	7,8	37,2	7,7

P +CN: pasturas y campo natural. Gr1, Gr2 y Gr3: grupos 1, 2 y 3 respectivamente

Por otra parte, se realizó análisis de varianza y una prueba chi-cuadrado de razón de verosimilitud para las variables continuas y discretas que entraron al análisis de conglomerados, respectivamente. En la tabla 2 se ven las variables en orden descendente de importancia en definir el agrupamiento. Las variables de mayor importancia (mayores valores de F o G/gl) fueron las que luego se tomaron en cuenta para la modelación y predicción del rendimiento.

Tabla 2. Valores de F, promedio y desvió para cada una las variables continuas modeladas como dependientes del grupo. Valores chi-cuadrado de máxima verosimilitud/grados de libertad para las variables discretas.

Variables discretas	gl	G/gl	Variables continuas	media	desvío	F
FS	6	83,9	Entrada P	32,9	22,7	231,77
AI	16	71,1	Entrada K	11,3	25	151,85
BQ	6	39,8	AA	3,3	2,6	106,06
AV	22	35,6	Entrada S	2	5,2	71,395
Ciclo cultivo	6	26,0	AF	3,8	0,4	63,95
Sistema de siembra	6	7,3	UT mam	3,8	0,4	39,3
			UTAH (posiembra)	1,6	5,2	32,47
			DH	35,2	6,9	27,31
			UTAT	6,3	1,1	27,16
			DS	78,4	22,9	22,8
			UTMH (posiembra)	3,3	0,3	16,05

gl: grados de libertad. G: valor asociado a la prueba chi-cuadrado FS: fecha de siembra. AI: antecesor invierno. BQ: realiza barbecho quimico (si/no). AV: antecesor verano. AA: años de agricultura continua. AF: aplicaciones de fertilizantes (cantidad total de aplicaciones en el ciclo del cultivo). UT mam: unidades toxicológicas mamíferos totales. UTAH: unidades toxicológicas/abeja herbicidas. DH: distancia entre hileras. UTAT: unidades toxicológicas/abeja total. DS: densidad de siembra en Kg semilla/ha. UTMH: unidades toxicológicas/mamífero herbicida.

A través del análisis de varianza se demostró que estas variables seleccionadas a partir de los valores de F o G (chi-cuadrado) tuvieron efecto sobre el rendimiento; se puede concluir que las prácticas de manejo que caracterizaron a cada uno de los grupos tienen un efecto significativo sobre el rendimiento. Cuando se corrió el modelo (2) las variables grupo, fecha de siembra, antecesor verano e invierno, agregado de P y S fueron significativas (p-valor <0,0002 para *ai* y p-valor <0.0001 para las demás variables). La BD utilizada registra variables con reconocido efecto sobre el rendimiento de soja y trigo, cuantificado a partir de experimentos realizados bajo condiciones controladas y diseño estadístico definido. Todas las variables antes mencionadas son parte de las propuestas de manejo impulsadas a partir de resultados de estos experimentos.

El conjunto de técnicas estudiadas cumplió con el objetivo de proponer una metodología de estudio que se adapte a las bases de datos de la producción agropecuaria, donde se pretenda generar nuevas hipótesis de investigación (desde la estadística exploratoria) o probar hipótesis planteadas a través de la modelación.

Validación: cuando se aplicó el método utilizando los datos de trigo, la distancia de Gower generó grupos bien definidos y que se pudieron caracterizar de acuerdo con las prácticas agronómicas. El agrupamiento a partir de las coordenadas del MFA para los datos de trigo no mostró una definición tan clara de los grupos, sin embargo a partir de las distancias de Gower se pudo establecer un agrupamiento satisfactorio desde el punto de vista técnico.

Se definieron tres grupos. La fecha de siembra para trigo se clasificó como: muy temprano (anteriores al 10 de mayo), temprana (del 10 al 30 de mayo), media (del 30 de mayo al 20 de junio) y tardía (del 20 de junio en adelante).

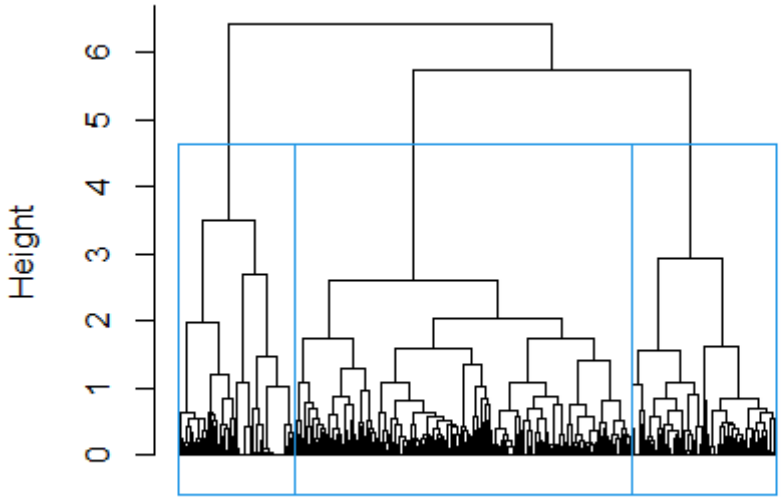

Figura 3: dendrograma a partir de matriz de distancias de Gower para datos de trigo.

Tabla 3. Valores de F y chi-cuadrado para las variables modeladas como dependientes del grupo para trigo.

Variables discretas	G	gl	G/gl	Variables continuas	F
Ciclo del cultivo	969	8	121	Entrada de N	353
				Entrada de P	67
				Número de años en	
Fecha de siembra	72	4	18	agricultura	32
				Entrada de S	20
Antecesor verano	160	28	6	Entrada de K	13
				Distancia entre	
Antecesor invierno	179	32	6	hileras	2

G: valor asociado a la prueba chi-cuadrado, Ngr3E: número de grupo según MFA (distancia euclidiana). gl: grados de libertad. F: valor F de Fisher para ANOVA.

En la determinación del agrupamiento para trigo, aplicando los métodos de ANAVA y prueba de razón de verosimilitud chi-cuadrado, se estableció que las variables discretas de mayor importancia fueron el ciclo del cultivo, la fecha de siembra y los antecesores. De las variables continuas, las más importantes fueron entrada de N y P y años de agricultura. Estos resultados muestran que, al obtenerse variables conocidas, ya sea por evaluación de cultivares o por ensayos de fertilidad en trigo, la metodología logra, en una serie de pasos sencillos de seguir, resumir información contenida en las BD que es de utilidad a la hora de generar nuevas líneas de investigación o de obtener información en términos de minería de datos, donde no es tan clara desde un inicio la información contenida en dichas bases.

Da, además, una idea de cómo corregir, con criterios técnicos y/o estadísticos, posibles problemas que se pueden presentar a la hora de analizar bases de datos de la producción. Lo más destacable en este sentido es que la metodología presenta confiabilidad estadística debido a la repetitividad en diferentes cultivos con diferentes requerimientos y manejos. Esta metodología logró obtener resultados similares a los obtenidos bajo investigación para trigo (García, 2004).

REFERENCIAS

Amoroso, Y. y Costales, D. 2016. Big Data: una herramienta para la administración pública. Ciencias de la Información 47(3): 3-8.

DIEA (Dirección de Estadísticas Agropecuarias). 2017. Anuario estadístico agropecuario. https://www.gub.uy/ministerio-ganaderia-agricultura-pesca/comunicacion/publicaciones/anuario-estadistico-diea-2017

Escofier, B. y Pagès, J. 2008. Analyses factorielles simples et multiples: Objectifs, méthodes et interpretation. Dunod, Paris, p157-172. ISBN 978-2-10-053809-6

Franco, J., Crossa, J. y Desphande, S. 2010. Hierarchical multiple-factor analysis for classifying genotypes based on phenotypic and genetic data. Crop Science 50(1): 105-117. https://doi.org/10.2135/cropsci2009.01.0053

Franco, J., Crossa, J., Villaseñor, J., Taba, S. y Eberhart, S. 1998. Classifying genetic resources using categorical and continuous variables. Crop Science 38 (6): 1688-1696. https://doi.org/10.2135/cropsci1998.0011183X003800060045x

García, A. 2004. Manejo de la fertilización con nitrógeno en trigo y su interacción con otras prácticas agronómicas. INIA 144:1-57.

Gower, J. 1971. A general coefficient of similarity and some of its properties. Biometrics 27(4): 857-874. https://doi.org/10.2307/2528823

Lawrence, C. y Krzanowski, W. 1996. Mixture separation for mixed-mode data. Statistics and computing 6: 84-92. https://doi.org/10.1007/BF00161577

Le Dien, S. y Pagès, J. 2003. Hierarchical Multiple Factor Analysis: Application to the comparison of sensory profiles. Food Quality and Preference, 14 (5-6): 397-403. https://doi.org/10.1016/S0950-3293(03)00027-2

Lindsey, J. 1995. Modelling Frecuency and Count Data. Oxford University Press, Oxford, p300. ISBN: 9780198523314

Tittonell, P. 2014. Ecological intensification of agriculture-sustainable by nature. Current Opinion in Environmental Sustainability 8: 53-61. https://doi.org/10.1016/j.cosust.2014.08.006