



UNICEN
Universidad Nacional del Centro
de la Provincia de Buenos Aires

FCE UNICEN
ECONOMICAS

**UNIVERSIDAD NACIONAL DEL CENTRO
DE LA PROVINCIA DE BUENOS AIRES**

**FACULTAD DE CIENCIAS ECONÓMICAS
LICENCIATURA EN ECONOMÍA EMPRESARIAL**

Trabajo Final de Graduación

**Análisis integral de la eficiencia y productividad en la
producción de soja: un estudio empírico basado en encuestas
a productores agropecuarios**

Barcelonna, Sofía

**Director/a:
Hoyos Maldonado, Daniel**

Contenido

Introducción	3
Marco teórico	6
La función de producción agrícola	6
Análisis de eficiencia técnica.....	8
Factores climáticos y sostenibilidad en la producción de soja	10
Metodología	11
Diseño de estudio.....	11
Contextualización del problema	11
Selección de muestra y recolección de datos.....	12
Aspectos teóricos y metodológicos	12
Metodología de Clustering	13
Consideraciones éticas y limitaciones	15
Futuras direcciones de investigación	15
Resultados.....	17
Análisis exploratorio de datos	17
Evaluación de multicolinealidad entre variables	20
Evaluación de variables dummy con baja variabilidad	23
Perfil estructural y sociodemográfico de los establecimientos	25
Perfil operativo y de gestión productiva	27

Análisis de conglomerados de productores	31
Análisis de eficiencia.....	34
Evaluación del Modelo de Tobit.....	39
Conclusiones.....	44
Conclusiones de los resultados	44
Conclusión general del trabajo	46
Futuras direcciones de investigación	47
Bibliografía	49
Anexo	56
Anexo I. Preguntas realizadas en la encuesta a productores agropecuarios	56
Anexo II. Metodología y diagnósticos del clustering	60
Anexo III. Tablas y archivos de soporte del Clustering.....	62
Anexo IV. Resultados del Análisis Envolvente de Datos (DEA)	63
Anexo V. Resultados del Modelo de Tobit	68
Anexo VI. Script de Python ejecutado en Google Colab	69

Introducción

El sector agrícola engloba todas las actividades productivas dedicadas al cultivo de la tierra con el propósito de obtener materias primas de origen vegetal utilizadas mayoritariamente como insumos del sector manufacturero.

En la actualidad, la producción de soja se destaca como la más relevante entre los productos agrícolas en Argentina. El país ocupa una posición destacada a nivel global como uno de los principales productores y exportadores de esta oleaginosa y sus derivados (aceite y harina).

Según el Instituto Nacional de Estadísticas y Censos (INDEC, 2024), las exportaciones del sector oleaginoso de Argentina alcanzaron los 22.282 millones de dólares, representando el 28,3% del total de las exportaciones de bienes del país. Dentro del sector oleaginoso, el complejo soja fue el más destacado, con exportaciones por 19.624 millones de dólares, lo que representó el 24,6% del total de las exportaciones argentinas en 2024 (INDEC, 2024).

A nivel global, la soja es uno de los cultivos más importantes debido a su alto contenido proteico y su uso en la producción de aceite vegetal. La demanda global de derivados de la soja ha crecido significativamente en las últimas décadas, impulsada por su uso en la alimentación animal; además se consolida una creciente demanda como insumo de productos destinados al consumo humano como el tofu, la leche de soja y otros alimentos (INASE, 2023). Este crecimiento ha posicionado a la soja como un cultivo clave en las estrategias productivas de muchos países. En el caso argentino, la soja representa uno de los principales pilares del sector agroexportador: solo en la campaña 2023/24, el complejo sojero generó más de 16.000 millones de dólares en exportaciones netas, consolidándose como un actor fundamental en la economía nacional (Bolsa de Comercio de Rosario, 2024).

En Argentina, además de su impacto económico, la producción de soja desempeña un papel crucial en el desarrollo social de las regiones productoras, generando empleo y contribuyendo al desarrollo regional. Según un informe de la Bolsa de Comercio de Rosario (2024), la cadena sojera genera aproximadamente 404.183

puestos de trabajo, representando el 11,7% del total de empleos en las cadenas agroalimentarias y el 2,5% del empleo total en el país.

Sin embargo, enfrenta desafíos significativos, como el cambio climático, la degradación del suelo y la necesidad de prácticas agrícolas más sostenibles. El calentamiento global representa una amenaza creciente para la producción agrícola nacional, provocando sequías extensas y pérdida de suelo fértil, lo que afecta gravemente la producción agrícola (Agroavances, 2024).

En este contexto, la optimización de los insumos y la adopción de tecnologías avanzadas resultan fundamentales para mejorar la eficiencia y sostenibilidad del sector. La implementación de prácticas sostenibles (como la rotación de cultivos, el uso de abonos orgánicos y técnicas de control biológico de plagas) busca equilibrar la productividad agrícola con la responsabilidad de proteger el medio ambiente (ReporteAsia, 2024).

La gestión eficiente de los recursos productivos es clave para la rentabilidad del cultivo. Comprender cómo interactúan insumos como la tierra, el trabajo, el capital y la tecnología permite a los productores tomar decisiones estratégicas que maximicen la productividad y reduzcan costos. En este sentido, la función de producción en el sector agrícola describe la relación entre la cantidad de insumos utilizados y la producción obtenida, permitiendo evaluar el impacto de cada factor en la eficiencia productiva. No obstante, su cálculo es complejo debido a la interacción de múltiples variables, como la calidad del suelo, la disponibilidad de agua, el manejo agrícola y las condiciones climáticas (Banco Mundial, 2015).

Las innovaciones tecnológicas, como la agricultura de precisión (AP), están transformando el sector mediante el uso de herramientas avanzadas que optimizan la gestión de los recursos y mejoran la productividad. La AP permite ajustar las prácticas agrícolas a las condiciones específicas de cada parcela, incrementando la eficiencia y reduciendo el impacto ambiental (CREA, 2024).

Diversos estudios econométricos y estadísticos han evaluado la eficiencia y productividad en la producción de soja. Por ejemplo, García Bernado (2018) analiza la

rentabilidad del cultivo de soja en Argentina a través de un enfoque empírico, considerando variables como los costos de producción, los rendimientos por hectárea y los precios de mercado.

En este contexto, el objetivo principal de este trabajo es realizar un análisis de la eficiencia y productividad en la producción de soja en el área de Tandil durante la campaña 2025-2026.

Para evaluar las distintas dimensiones de la eficiencia de los factores de producción, se realizará un análisis estadístico de las respuestas de los productores, que permitirá comparar unidades productivas. Este enfoque busca adaptar los modelos empíricos a la realidad local y, para ello, se realizará una encuesta a productores agropecuarios a fin de capturar la diversidad de prácticas productivas y de las condiciones del terreno afectado a esta actividad.

Marco teórico

La función de producción agrícola

Según Debertin (2012), una función de producción se define como una regla que asigna a cada valor en un conjunto de variables (dominio de la función) un único valor en otro conjunto de variables (rango de la función).

Dentro de los modelos económicos para estimar funciones de producción, la función Cobb-Douglas ha sido ampliamente utilizada en estudios agrícolas (Cobb & Douglas, 1928). Cruz Lauracio (2019) menciona que esta función es especialmente aplicable para representar la relación entre los insumos utilizados en la producción y la cantidad de productos obtenidos. En este modelo, la variable dependiente es la cantidad de producto (kg/ha), mientras que las variables explicativas incluyen el trabajo (jornal/ha), el capital (horas máquina/ha) y la tierra (ha). El mismo ha demostrado ser eficaz para estudiar la productividad total de los factores y analizar la eficiencia en la asignación de recursos (Banco Mundial, 2020), considerando que los factores productivos influyen positivamente en la producción, manteniendo las demás variables constantes (*ceteris paribus*).

La función de producción Cobb-Douglas se expresa de la siguiente manera (Cobb & Douglas, 1928):

$$Q = A * (L^\alpha) * (K^\beta)$$

Donde:

- Q: cantidad de producción.
- A: factor de productividad total de los insumos (eficiencia técnica).
- L: cantidad de trabajo utilizada (jornales/ha u horas-máquina/ha).
- K: cantidad de capital utilizada (jornales/ha u horas-máquina/ha).
- α y β : coeficientes de elasticidad que representan la contribución relativa de cada insumo a la producción total, con $\alpha+\beta=1$.

En esta especificación, la suma de las elasticidades determina los rendimientos a escala: si $\alpha+\beta=1$ hay rendimientos constantes, si $\alpha+\beta<1$ son decrecientes y si $\alpha+\beta>1$ son crecientes.

En particular, los coeficientes de elasticidad asociados con los factores productivos reflejan la importancia relativa de cada uno de ellos en la determinación de la producción agrícola, sugiriendo que aumentos en la utilización de estos factores podrían resultar en incrementos significativos en la producción.

Estudios recientes, como el de Pérez et al. (2020), han encontrado que la elasticidad del capital tiende a ser mayor en regiones con alta mecanización, lo que indica que las inversiones en tecnología y maquinaria pueden aumentar la productividad de manera desproporcionada. Esto refuerza la necesidad de incluir factores adicionales en los modelos de producción agrícola para capturar mejor las variaciones específicas del sector.

Por su parte, Córdoba (2014) observa que tanto el rendimiento de los cultivos como las propiedades del suelo exhiben patrones espaciales que pueden ser estudiados y utilizados para mejorar la gestión agrícola.

En Argentina, el uso de tecnologías de Agricultura de Precisión (AP) ha permitido medir y gestionar de manera diferenciada la variabilidad espacial intralote de propiedades del sitio, como el suelo, el terreno y los nutrientes, así como los rendimientos. Esto permite a los productores ajustar su toma de decisiones en función de datos en tiempo real, reduciendo desperdicios y maximizando la productividad (Gebbers & Adamchuk, 2010).

En el ámbito local, diversos estudios han abordado la eficiencia en sistemas agrícolas y agroindustriales argentinos. Investigaciones del INTA Pergamino sobre soja y maíz muestran que la variabilidad en la calidad de insumos y en las condiciones agroclimáticas genera diferencias significativas en rendimientos y eficiencia operativa (INTA, 2023). Por su parte, estudios de la EEAOC en Tucumán han demostrado que la eficiencia técnica en cultivos extensivos está estrechamente vinculada a la adopción de buenas prácticas agrícolas, más allá del tamaño de la explotación (EEAOC, 2020).

Asimismo, investigaciones impulsadas por CONICET han destacado que la productividad total de factores en la agricultura argentina ha crecido de manera sostenida pero heterogénea, con brechas de desempeño notables entre regiones y productores (CONICET, 2018). Estos antecedentes refuerzan la pertinencia de analizar la eficiencia en Tandil como un caso específico que puede contribuir a la discusión nacional.

Desde una perspectiva aplicada, la AP no solo mejora la eficiencia técnica, sino que también contribuye a la sostenibilidad. Bongiovanni y Lowenberg-Deboer (2004) destacan que la aplicación precisa de fertilizantes y pesticidas reduce el impacto ambiental, minimizando la contaminación del suelo y del agua. Además, estudios recientes han encontrado que la adopción de AP mejora significativamente la eficiencia productiva en cultivos como la soja (Sharma & Baliyan, 2022). No obstante, la adopción de la AP enfrenta desafíos importantes. Díaz y Gómez (2017) identifican tres barreras clave para su implementación:

- Altos costos iniciales: La inversión en tecnología y capacitación puede ser prohibitiva para pequeños y medianos productores.
- Falta de conocimiento técnico: La curva de aprendizaje y la capacitación limitada dificultan la adopción de herramientas avanzadas.
- Resistencia al cambio: Algunos productores prefieren continuar con métodos tradicionales debido a la incertidumbre en los beneficios a corto plazo.

En tanto, Perren (2008) se enfocó en la determinación de costos en cultivos específicos. Para ello, creó una matriz de costo por hectárea y determinó el punto de equilibrio de producción del cultivo en estudio. Con base en estos resultados, se llevaron a cabo análisis de viabilidad económica-financiera.

Análisis de eficiencia técnica

La eficiencia técnica mide la capacidad de una unidad de producción para obtener el máximo rendimiento posible con un conjunto de insumos. En la agricultura, es fundamental para maximizar la productividad y reducir costos. Un método

ampliamente utilizado para evaluar esta eficiencia es el Análisis Envolvente de Datos (DEA), un enfoque no paramétrico que permite comparar explotaciones agrícolas con distintas combinaciones de insumos y productos (Charnes, Cooper & Rhodes, 1978).

El DEA identifica las unidades más eficientes y mide la distancia de otras respecto a esta "frontera" de eficiencia, permitiendo detectar ineficiencias y áreas de mejora en el uso de insumos. Esto facilita la implementación de prácticas agrícolas más eficientes y sostenibles, optimizando recursos clave para la rentabilidad del cultivo de soja. Este método ha sido aplicado en numerosos estudios agrícolas para evaluar la eficiencia relativa entre productores y detectar diferencias en técnicas de producción, acceso a tecnología y condiciones climáticas (Cooper, Seiford & Tone, 2007; Fried, Lovell & Schmidt, 2008).

Como señalan Fried, Lovell y Schmidt (2008) el DEA se ha convertido en una de las herramientas más versátiles y ampliamente utilizadas para medir eficiencia técnica, dado que no requiere una especificación funcional previa y permite evaluar la eficiencia relativa de múltiples unidades que utilizan distintos niveles de insumos para producir varios productos. (p. 11).

Este método no paramétrico permite comparar explotaciones agrícolas con distintas combinaciones de insumos y productos, identificando aquellas que operan sobre la frontera eficiente y midiendo la distancia relativa de las demás unidades respecto a dicha frontera (Charnes, Cooper & Rhodes, 1978). Esto facilita la detección de ineficiencias en el uso de los recursos y orienta posibles mejoras en la gestión productiva.

En Argentina, diversas investigaciones han empleado el DEA para evaluar la eficiencia en explotaciones agropecuarias, tanto en cultivos extensivos como intensivos. Por ejemplo, Berbel y Rodríguez-Ocaña (2007) destacan su aplicabilidad en el análisis de eficiencia en regiones con heterogeneidad estructural y productiva, como ocurre en el sistema agrícola pampeano.

Factores climáticos y sostenibilidad en la producción de soja

Las condiciones climáticas desempeñan un papel fundamental en la productividad de la soja, influyendo directamente en los rendimientos. Lobell y Burke (2010) destacan que las variaciones en temperatura y precipitaciones pueden afectar significativamente la producción, especialmente en zonas donde la variabilidad climática es alta.

En este sentido, Hatfield et al. (2011) argumentan que las estrategias de manejo adaptativo, como la selección de variedades resistentes y la optimización del riego, son esenciales para mitigar los efectos del cambio climático.

Metodología

Diseño de estudio

El sitio de investigación es el partido de Tandil, ubicado en la provincia de Buenos Aires, Argentina. La población de estudio está conformada por establecimientos rurales localizados en este municipio que se dedican a actividades agrícolas, específicamente a la producción de soja.

Tandil cuenta con una superficie total de aproximadamente 477.020 hectáreas destinadas a explotaciones agropecuarias, distribuidas en más de 1.000 establecimientos productivos (Cadena 103, s.f.). Durante la campaña agrícola 2017/2018, se sembraron alrededor de 134.110 hectáreas con soja, representando el 40,4% del total de la superficie cultivada en el partido (El Eco de Tandil, 2018).

Estos datos reflejan la centralidad de la soja como principal cultivo extensivo de la región, superando ampliamente a otras producciones como el maíz y el girasol. Comprender cómo optimizar la producción de soja en esta zona resulta fundamental para mejorar la eficiencia y sostenibilidad del sector agropecuario local.

El diseño de investigación propuesto es exploratorio-descriptivo, adoptando un enfoque cuantitativo. Este diseño permitirá identificar las variables relevantes en la producción de soja y analizar su impacto en la eficiencia productiva.

Contextualización del problema

El cultivo de soja en la región de Tandil es de gran importancia económica. Según datos del Ministerio de Agricultura, Ganadería y Pesca de la Nación (MAGyP, 2023), en la campaña 2022/2023 el partido de Tandil registró una superficie sembrada de aproximadamente 50.000 hectáreas de soja, con una producción total estimada en 150.000 toneladas y un rendimiento promedio de 3.000 kg/ha. Estos indicadores posicionan a Tandil como un actor relevante en la producción sojera dentro de la provincia de Buenos Aires. Comprender cómo optimizar la producción de soja en esta

región es crucial para los productores locales, ya que permite mejorar la eficiencia productiva y la sostenibilidad de sus sistemas agrícolas.

Selección de muestra y recolección de datos

Para recopilar los datos necesarios para desarrollar el estudio se construirá una muestra de productores agropecuarios, sobre la base de la técnica denominada muestreo “bola de nieve”. Este método permite construir una muestra no probabilística, adecuada para este tipo de estudio, que habilitará obtener datos susceptibles de ser cotejados con otras fuentes de naturaleza secundaria como, por ejemplo, el censo agropecuario desarrollado por el INDEC.

Los productores agropecuarios serán contactados y recibirán un vínculo en el que accederán a un formulario digital donde se encuestarán datos referidos al tamaño de la explotación, el tipo de tecnología utilizada, las prácticas agrícolas, entre otros tópicos referidos al objeto de este estudio. En el Anexo I se acompañan las preguntas que estarán incluidas en el formulario de la encuesta.

El cuestionario será validado mediante revisión por expertos y pruebas piloto.

Aspectos teóricos y metodológicos

Una vez recolectados y revisados los datos, se llevará a cabo un proceso de limpieza para garantizar la calidad y consistencia de la información recopilada. Se eliminarán posibles errores, valores atípicos o datos faltantes.

Los datos definitivos serán normalizados y estandarizados para asegurar la comparabilidad de los valores obtenidos. Se aplicarán técnicas estadísticas para estimar la función de producción, la eficiencia y productividad de la producción de soja en el partido de Tandil.

A los efectos del estudio de la eficiencia, se aplicará el Análisis de la Envolvente de Datos (DEA). En el caso específico de la producción de soja en Tandil, el DEA

permitirá, desde un enfoque de producción, comparar la eficiencia de las diferentes explotaciones agrícolas integrantes de la muestra. Además, este estudio se complementará con un análisis de regresión de variables censuradas (modelo Tobit); de esta manera, se procurará identificar determinantes exógenos que afectan la eficiencia productiva (Battese & Coelli, 1995).

A fin de identificar los factores que explican las diferencias de eficiencia entre unidades productivas, se empleó un modelo de regresión de variables censuradas (modelo Tobit). Esta metodología permite analizar la influencia de variables exógenas (como las condiciones climáticas, las características del suelo o el tipo de tecnología utilizada) sobre los puntajes de eficiencia obtenidos, considerando que estos se encuentran limitados entre 0 y 1 (Tobin, 1958; Greene, 2003).

Metodología de Clustering

Para la identificación de tipologías de productores se aplicará el análisis de conglomerados, una técnica multivariante que permite clasificar unidades en grupos homogéneos a partir de sus similitudes, maximizando la cohesión interna y las diferencias entre los grupos (Hair et al., 2019). En el contexto agronómico, esta metodología permitirá captar la heterogeneidad estructural y de gestión existente entre los productores de soja, contribuyendo a la construcción de perfiles que faciliten la interpretación de las distintas estrategias productivas (Everitt et al., 2011).

En una primera instancia se empleará el análisis de conglomerados jerárquico, utilizando el método de Ward con distancia euclídea al cuadrado. Este procedimiento se selecciona por su capacidad para minimizar la varianza intragrupal y proporcionar una representación dendrítica de fácil interpretación, lo cual permitirá explorar la existencia de estructuras naturales en los datos sin la necesidad de establecer a priori el número de grupos (Ward, 1963; Everitt et al., 2011).

La determinación del número de conglomerados se realizará a partir de la aplicación de métricas consolidadas en la literatura. Se recurirá al método del codo, que posibilita observar la evolución de la varianza intragrupal en función del número de

grupos, y al índice silhouette, que pondera la cohesión y la separación entre los conglomerados (Rousseeuw, 1987; Ketchen & Shook, 1996). De esta manera, se asegurará una decisión metodológica fundamentada y se evitará la arbitrariedad en la elección del número óptimo de grupos.

Una vez establecido el número de conglomerados, se aplicará el algoritmo no jerárquico K-means, considerado uno de los métodos más eficientes y ampliamente utilizados para la clasificación definitiva de observaciones (MacQueen, 1967; Jain, 2010). Este procedimiento se fundamenta en la asignación iterativa de cada observación al centroide más cercano, optimizando la homogeneidad intragrupal. Su implementación, en combinación con el análisis jerárquico, permitirá robustecer la validez de la segmentación, siguiendo las recomendaciones metodológicas de complementar ambos enfoques (Kaufman & Rousseeuw, 2005).

Como estrategia de validación y apoyo interpretativo, se utilizará el Análisis de Componentes Principales (PCA), con el propósito de proyectar los conglomerados en un espacio reducido que conserve la mayor proporción de varianza. Este procedimiento permitirá representar gráficamente la separación entre los grupos y verificar la coherencia de la estructura obtenida (Jolliffe & Cadima, 2016). El PCA, en este sentido, no modificará la segmentación, pero aportará claridad comunicacional y fortalecerá la solidez de los hallazgos.

El empleo combinado de estas metodologías garantizará un análisis riguroso y replicable, basado en criterios técnicos consolidados. El análisis de conglomerados, de este modo, se integrará como un complemento idóneo al estudio de eficiencia y productividad, proporcionando tipologías empíricas que reflejen la diversidad productiva y tecnológica de los productores de soja en Tandil (Hair et al., 2019; Kaufman & Rousseeuw, 2005).

Introducción al entorno de trabajo

El análisis de los datos empíricos de este estudio se llevará a cabo utilizando el lenguaje de programación Python, una herramienta de código abierto ampliamente

reconocida en la investigación científica por su versatilidad y precisión en el tratamiento de grandes volúmenes de datos. Python ha ganado popularidad debido a su amplio ecosistema de bibliotecas especializadas en análisis de datos, estadísticas y visualización, lo que lo convierte en una opción robusta y flexible para el análisis cuantitativo en diversas disciplinas (Van Rossum, 2009; McKinney, 2010).

Para ejecutar este análisis de manera eficiente y accesible, se utilizará Google Colaboratory (Colab), una plataforma en la nube que facilita la ejecución de código Python sin necesidad de instalaciones locales. Colab permite el uso gratuito de recursos computacionales escalables, lo que resulta en una opción ideal para trabajar con datasets complejos y realizar análisis econométricos de forma remota (Google Research, 2020). Esta plataforma no solo ofrece un entorno accesible y flexible, sino que también facilita la integración con bibliotecas clave como pandas, numpy, matplotlib y seaborn, herramientas fundamentales para la manipulación, análisis y visualización de datos en este trabajo.

El uso de Python junto con Google Colab se justifica por su capacidad para automatizar procesos de análisis, generar visualizaciones interactivas de alta calidad y garantizar la reproducibilidad del trabajo, elementos clave en la investigación empírica.

Consideraciones éticas y limitaciones

Se tuvieron en cuenta consideraciones éticas en la recolección de datos, asegurando que los participantes estén informados y den su consentimiento explícito. También se reconocieron las posibles limitaciones del estudio, como la representatividad de la muestra y la precisión de las respuestas en las encuestas.

Futuras direcciones de investigación

El estudio puede abrir la puerta a futuras investigaciones, como la evaluación del impacto de nuevas tecnologías o la comparación de prácticas entre diferentes regiones. También se considerará la realización de un análisis de sensibilidad para evaluar cómo

los cambios en los parámetros del modelo afectan los resultados, aportando una comprensión más robusta de la relación entre insumos y producción.

Resultados

Análisis exploratorio de datos

Con el objetivo de comprender la estructura, calidad y distribución de los datos recolectados mediante encuestas estructuradas, se realizó un análisis exploratorio exhaustivo como etapa previa a la aplicación de metodologías econométricas y no paramétricas. Este análisis constituye un insumo indispensable para detectar valores atípicos, evaluar la completitud de las respuestas y establecer criterios objetivos de depuración de variables.

La base está conformada por 60 observaciones y más de 70 variables que capturan múltiples dimensiones de la producción agraria, entre ellas aspectos técnicos, económicos, organizacionales y contextuales. Se consideraron de especial interés cinco variables centrales que reflejan con claridad el funcionamiento del sistema productivo: rendimiento de soja por hectárea, costo de producción, superficie trabajada, años de experiencia del productor y porcentaje de inversión en tecnología. Estas variables permiten abordar con rigor la caracterización de los establecimientos desde una perspectiva integral, incorporando tanto factores físicos como de gestión y conocimiento.

Cabe destacar que las variables de experiencia del productor y superficie de la explotación no fueron relevadas de forma directa como valores numéricos continuos, sino que se estimaron a partir de intervalos categóricos provistos en el cuestionario. Para su inclusión en los análisis, se asignaron valores promedio representativos dentro de cada rango, conforme a criterios estandarizados de imputación utilizados en estudios similares. Esta aproximación, si bien introduce un nivel de agregación, permite incorporar dichas variables conservando su validez analítica.

El examen gráfico de estas variables permitió identificar formas de distribución heterogéneas. El rendimiento de soja, expresado en kilogramos por hectárea, se distribuye de manera moderadamente asimétrica, con un promedio cercano a los 3.277 kg/ha y un rango que oscila entre 1.900 y 5.500 kg/ha. Esta variabilidad evidencia diferencias técnicas sustantivas entre unidades productivas. El costo de producción

muestra una dispersión considerable, con registros entre \$200 y \$3.700 por hectárea, lo que sugiere estrategias económicas diversas por parte de los productores.

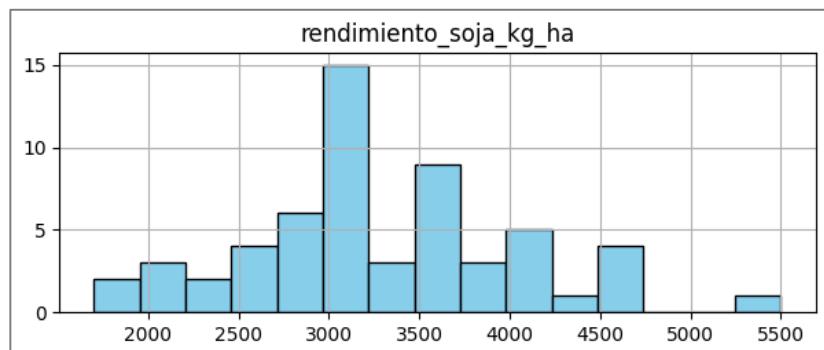


Figura 1. Rendimiento de soja en kilogramos por hectárea. *Fuente: elaboración propia en Python en base a datos de encuesta.*

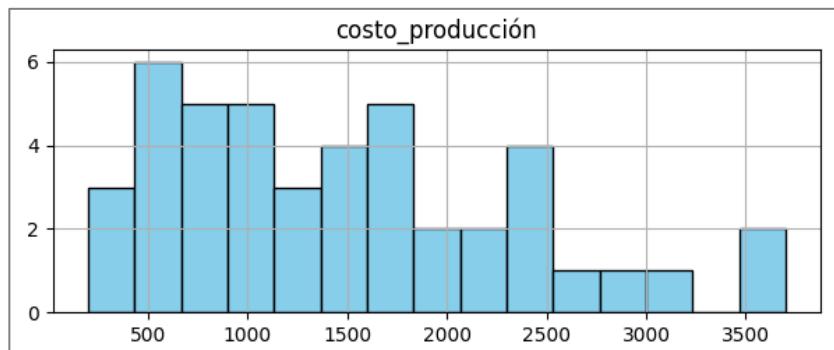


Figura 2. Costo de producción por hectárea. *Fuente: elaboración propia en Python en base a datos de encuesta.*

En lo que respecta a la escala de operación, la superficie trabajada evidencia una distribución bimodal, diferenciando a pequeños productores (con menos de 300 ha) de aquellos que gestionan explotaciones de gran escala (superiores a 1.000 ha). La trayectoria productiva, capturada a través de la variable años de experiencia, se concentra mayoritariamente en valores superiores a los 15 años, aunque se identifican también casos con menos de cinco años de antigüedad. Por su parte, el porcentaje de inversión en tecnología presenta una fuerte concentración en valores bajos, con algunos casos aislados de alta intensidad tecnológica.

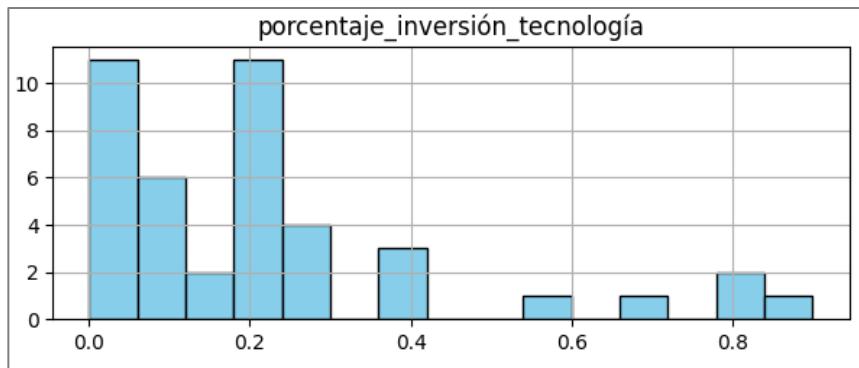


Figura 3. Porcentaje de inversión en tecnología por año. *Fuente: elaboración propia en Python en base a datos de encuesta.*

La elección de estas variables se sustenta tanto en su relevancia teórica como en su capacidad empírica para capturar elementos clave del proceso productivo. Su inclusión en los análisis constituye una aproximación robusta a la realidad heterogénea del sector y permite identificar diferencias sustantivas en cuanto a capacidad técnica, escala, conocimiento y adopción tecnológica.

Asimismo, se realizó un análisis del nivel de completitud de las variables para eliminar aquellas que tengan más del 35% de datos nulos. Del total de indicadores, un subconjunto presentó niveles significativos de datos faltantes. Las variables con mayor proporción de valores ausentes fueron el porcentaje de inversión en tecnología (30%) y el costo de producción (27%), ambas consideradas de alta relevancia analítica. En menor medida, también se registraron faltantes en variables como si arrienda o no la tierra, rendimiento de soja, y algunas dummies relacionadas con financiamiento y asesoramiento. Este diagnóstico permitió anticipar eventuales restricciones para su uso en análisis cuantitativos y guiar decisiones metodológicas en las etapas subsiguientes.

El gráfico siguiente resume visualmente el porcentaje de valores faltantes por variable, destacando aquellas con mayor impacto en la calidad de la información disponible.

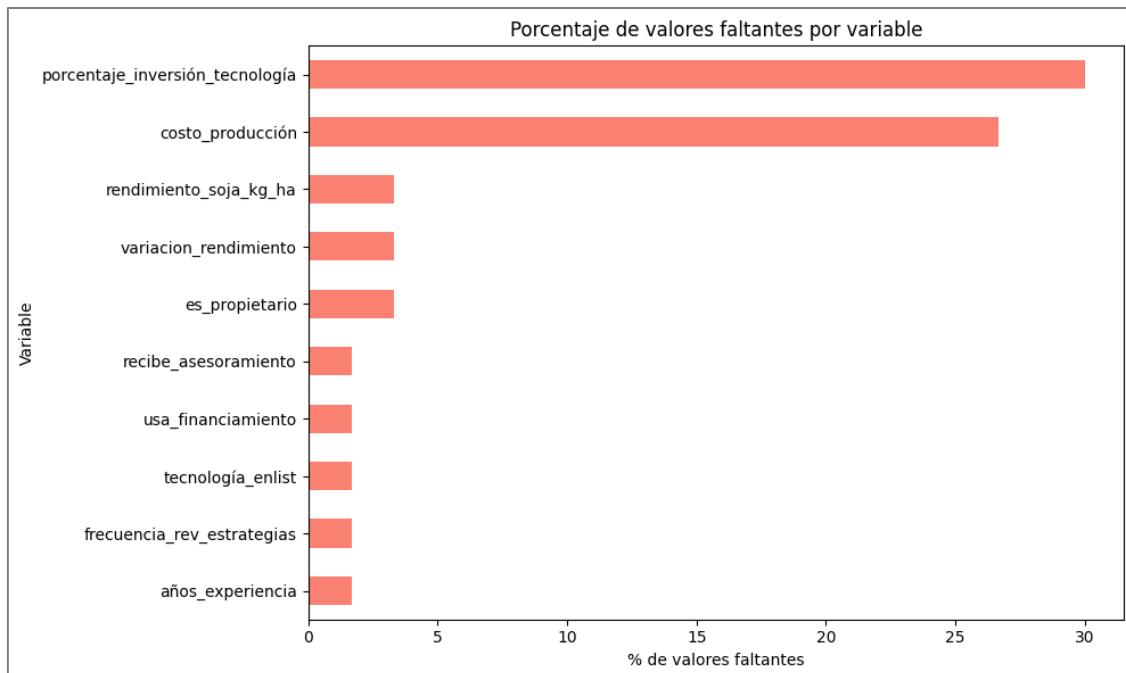


Figura 4. Porcentaje de valores faltantes por variable en la base de datos utilizada.

Fuente: elaboración propia en Python en base a datos de encuesta.

Evaluación de multicolinealidad entre variables

Se llevó a cabo un exhaustivo procedimiento de depuración de variables explicativas, orientado a mitigar los riesgos asociados a la multicolinealidad y garantizar una especificación parsimoniosa, teóricamente fundamentada y estadísticamente robusta. El análisis incluyó tanto variables numéricas como variables categóricas binarias (dummies), contemplando distintos criterios de exclusión en función del tipo de dato.

En el caso de las variables continuas, se utilizó como criterio principal el coeficiente de correlación de Pearson, una medida clásica de asociación lineal. Se estableció como umbral crítico un valor absoluto igual o superior a 0,6 ($|r| \geq 0,6$), a partir del cual la colinealidad comienza a adquirir relevancia desde el punto de vista econométrico. Para facilitar la identificación de asociaciones elevadas, se construyó una matriz de correlación entre todas las variables numéricas seleccionadas, y se visualizó la información mediante un mapa de calor. Como resultado de este procedimiento, no

se detectaron pares de variables que superaran el umbral establecido, por lo que no fue necesario eliminar variables numéricas por colinealidad. Esta conclusión refuerza la independencia relativa entre los indicadores cuantitativos incorporados en la modelización, y valida su inclusión conjunta en las especificaciones posteriores.

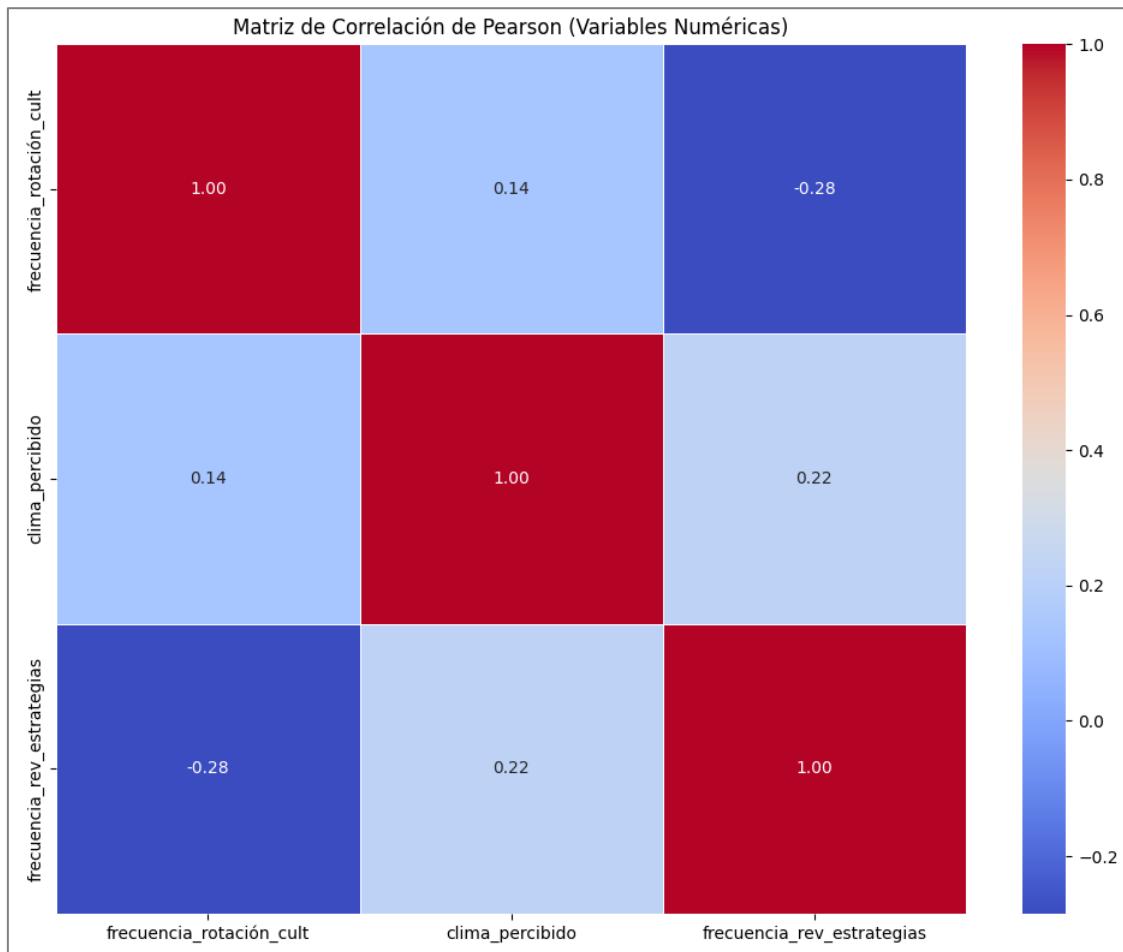


Figura 5. Matriz de correlación de Pearson para variables numéricas. *Fuente: elaboración propia en Python.*

En forma paralela, se evaluó la redundancia informativa entre variables binarias mediante el coeficiente Phi, que representa una medida específica de correlación para variables dicotómicas. Se construyó una matriz Phi completa y su correspondiente visualización gráfica, con el fin de identificar empíricamente los patrones de asociación entre las dummies. Para el análisis, se adoptó nuevamente el umbral de $|\text{Phi}| \geq 0,6$, señalando aquellos pares que presentaban una alta superposición conceptual o

redundancia en la codificación. El procedimiento arrojó cinco pares de variables con correlaciones significativas, los cuales se detallan en la tabla correspondiente. A partir de dicha evidencia, se procedió a la depuración del conjunto categórico, con criterios similares a los aplicados en el caso anterior: prioridad teórica, calidad de captura y variabilidad. En consecuencia, fueron eliminadas cuatro variables del conjunto: desafio_cambio_clim, fact_asesoramiento, fact_insumos e insumos_fungicida. Todas ellas compartían un alto grado de solapamiento informativo con otras variables similares (como desafio_fluct_precio, fact_tecnología o insumos_insecticida) lo cual comprometía la independencia de los predictores y dificultaba la interpretación de los coeficientes estimados.

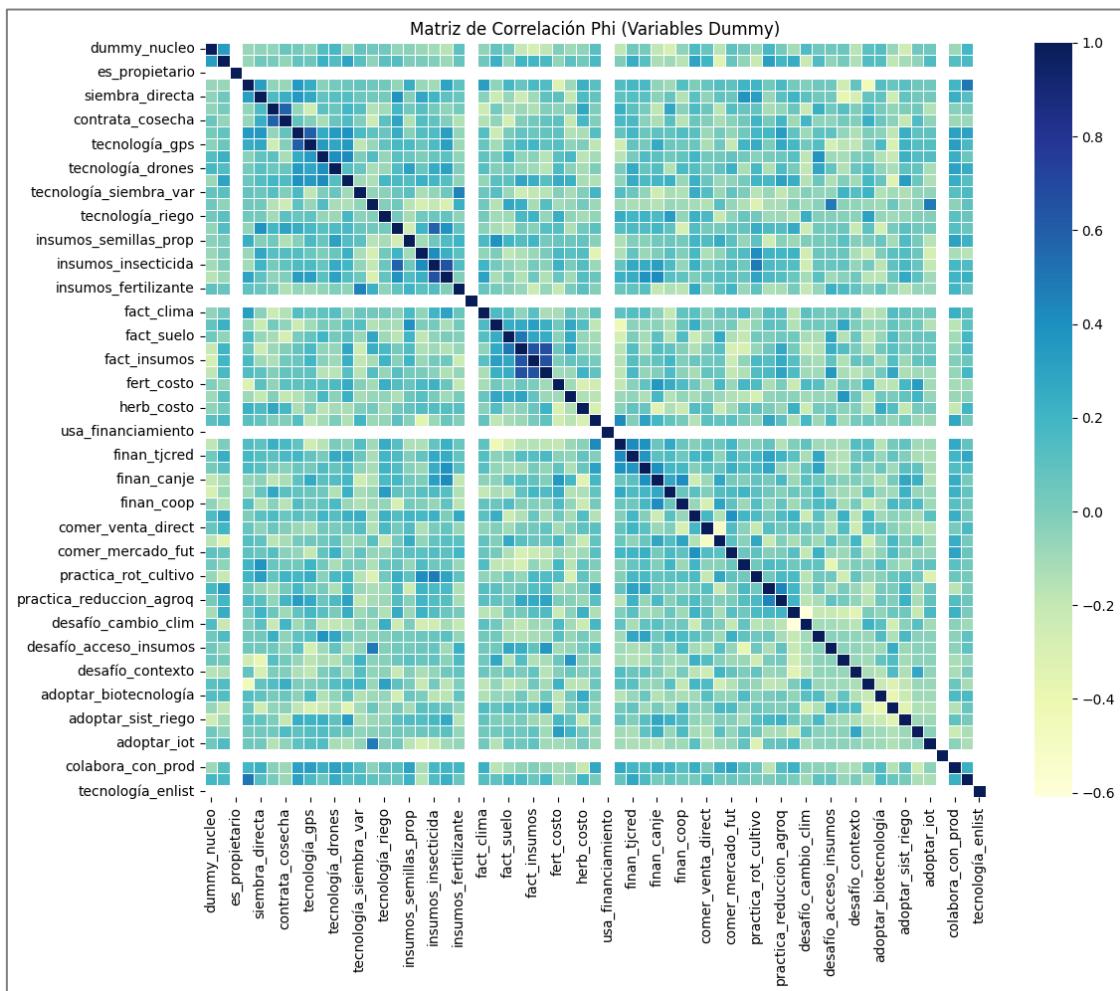


Figura 6. Matriz de Correlación Phi para variables dummy. *Fuente: elaboración propia en Python.*

Finalizado este proceso, se reconstruyeron las matrices de correlación para constatar la eliminación efectiva de las asociaciones problemáticas. Se verificó que el conjunto final de variables binarias no presentaba correlaciones elevadas entre sí, consolidando una base de datos más limpia, no redundante y mejor adaptada a los requerimientos del modelo econométrico Tobit.

En suma, este procedimiento de depuración constituye un paso metodológico indispensable para evitar sesgos en la estimación de parámetros, reducir errores estándar inflados y garantizar una interpretación económica coherente. La eliminación de variables altamente correlacionadas no solo fortalece la credibilidad del modelo, sino que también permite identificar con mayor precisión cuáles son las características efectivamente relevantes en la explicación de los niveles de eficiencia técnica observados en los productores de soja analizados.

Evaluación de variables dummy con baja variabilidad

Se implementó un proceso riguroso de depuración de variables categóricas binarias (dummies), con el objetivo de preservar la validez técnica del modelo, garantizar una adecuada identificación de los coeficientes y asegurar la capacidad explicativa de las especificaciones resultantes.

El criterio utilizado en esta etapa fue la detección de variables con baja variabilidad, definida como una proporción de ceros o unos superior al 85 por ciento del total de observaciones. Esta condición compromete la utilidad analítica de la variable, al limitar su capacidad para discriminar entre unidades y generar variación explicativa, además de aumentar el riesgo de colinealidad estructural o resultados estadísticamente débiles.

Más allá de la justificación técnica, se adoptó una lectura contextual del fenómeno de baja variabilidad, entendiendo que dicha homogeneidad puede reflejar características estructurales del sistema productivo. En efecto, muchas de las variables eliminadas expresan prácticas ampliamente difundidas en el sector, como la siembra directa o el uso de fertilizantes, o decisiones tecnológicas todavía incipientes, como la

adopción de tecnologías emergentes, lo que justifica su escasa dispersión en la muestra. La falta de variabilidad, en este sentido, no constituye un defecto del instrumento de recolección, sino un emergente de la realidad productiva observada.

A partir de este criterio, se eliminaron veintidós variables dummy que presentaban niveles de concentración superiores al umbral establecido. Estas variables se vinculan a tecnologías ampliamente adoptadas (como siembra directa, uso de herbicidas o rotación con soja), prácticas generalizadas entre los productores (como recibir asesoramiento técnico o colaborar con otros productores), tecnologías emergentes aún no adoptadas masivamente (como IoT, plataformas Enlist o sistemas de riego), percepciones poco difundidas (como ciertos factores de afectación vinculados al suelo, clima o contexto), así como instrumentos financieros o comerciales poco utilizados (como Sociedades de Garantía Recíproca o venta directa).

Entre las variables eliminadas por presentar una alta proporción de ceros se encuentran: sem_costo, finan_sgr, insumos_fertilizante, comer_mercado_fut, tecnología_siembra_var, desafio_contexto, desafío_plagas, adoptar_genética, desafío_malezas, tecnología_monitores, desafío_acceso_insumos, adoptar_iot y tecnología_riego.

En tanto, las variables eliminadas por presentar una alta proporción de unos fueron: insumos_insecticida, recibe_asesoramiento, insumos_herbicida, practica_rot_cultivo, dummy_nucleo, rotación_con_soja, insumos_semillas_gen, fact_clima y siembra_directa.

Este procedimiento, sustentado en principios estadísticos y en una lectura contextual de la información empírica, permite asegurar que las variables dummies retenidas en el modelo conserven un grado mínimo de variabilidad, condición indispensable para su inclusión econométrica. Asimismo, contribuye a mejorar la parsimonia del modelo y a reforzar la robustez interpretativa de los coeficientes estimados.

Perfil estructural y sociodemográfico de los establecimientos

La muestra total del estudio está compuesta por 60 productores agropecuarios. Sin embargo, en algunas variables específicas se registraron respuestas incompletas, por lo que los porcentajes que se presentan a continuación fueron calculados únicamente sobre los casos válidos para cada categoría.

El 81,67% de los productores se localiza en la provincia de Buenos Aires, seguido por Córdoba con un 8,33%, Entre Ríos con un 5%, y Santa Fe con un 3,33%. La muestra también incluye un pequeño porcentaje de participantes de La Pampa (1,67%), lo que evidencia la focalización regional del estudio, pero con cierta heterogeneidad territorial.

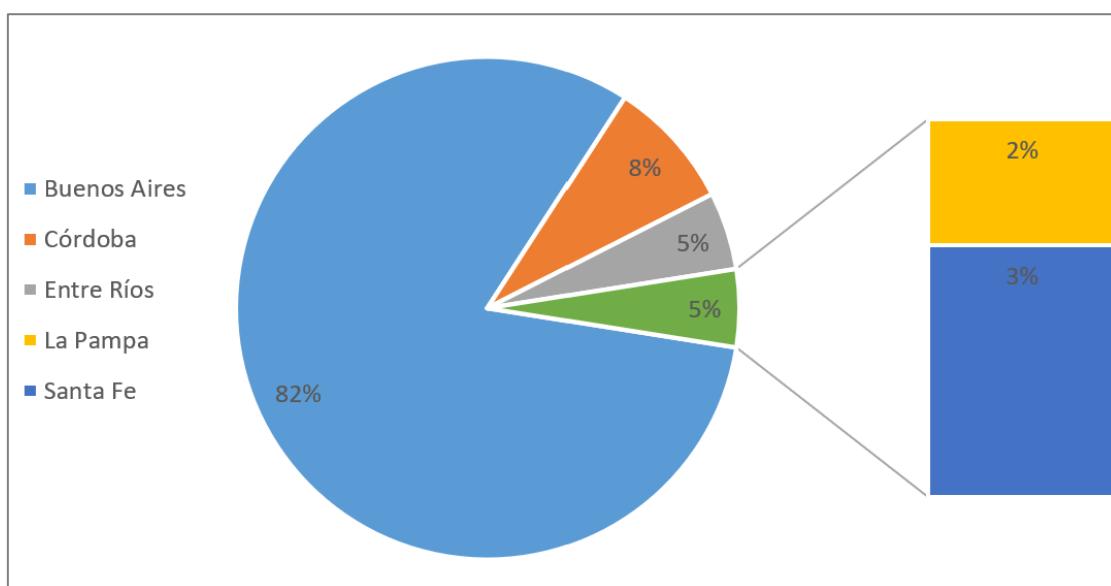


Figura 7. Distribución por provincias de la muestra. *Fuente: elaboración propia en base a datos de encuesta.*

A nivel de localidades, se destacan Tandil (18,33%), Azul (18,33%) y Lobería (5%) son las más mencionadas. Con una proporción del 3,33% cada una, aparecen Benito Juárez, Coronel Pringles, Necochea, Olavarría y Saladillo; todas pertenecientes a la provincia de Buenos Aires. Sigue una extensa lista de localidades con presencia marginal del 1,67% que representa la diversidad territorial del área pampeana.

Desde el punto de vista legal, el 61,66% de los productores declaró operar bajo una estructura societaria, mientras que el 38,33% lo hace de manera unipersonal. Esta tendencia evidencia una predominancia de esquemas organizativos más formales y posiblemente de mayor escala operativa.

Respecto de los años de experiencia en la actividad agropecuaria, se obtuvieron 59 respuestas válidas. Entre quienes sí respondieron, el 37% cuenta con más de 20 años de trayectoria, lo que refleja un alto nivel de conocimiento acumulado en el sector. El 22% posee menos de 5 años de experiencia, lo que indica la presencia de nuevos actores en la actividad. Por su parte, un 21% declaró entre 5 y 10 años de experiencia, y otro 20% entre 11 y 20 años. Esta distribución sugiere una combinación equilibrada entre productores consolidados y nuevas generaciones que se incorporan al negocio agropecuario.

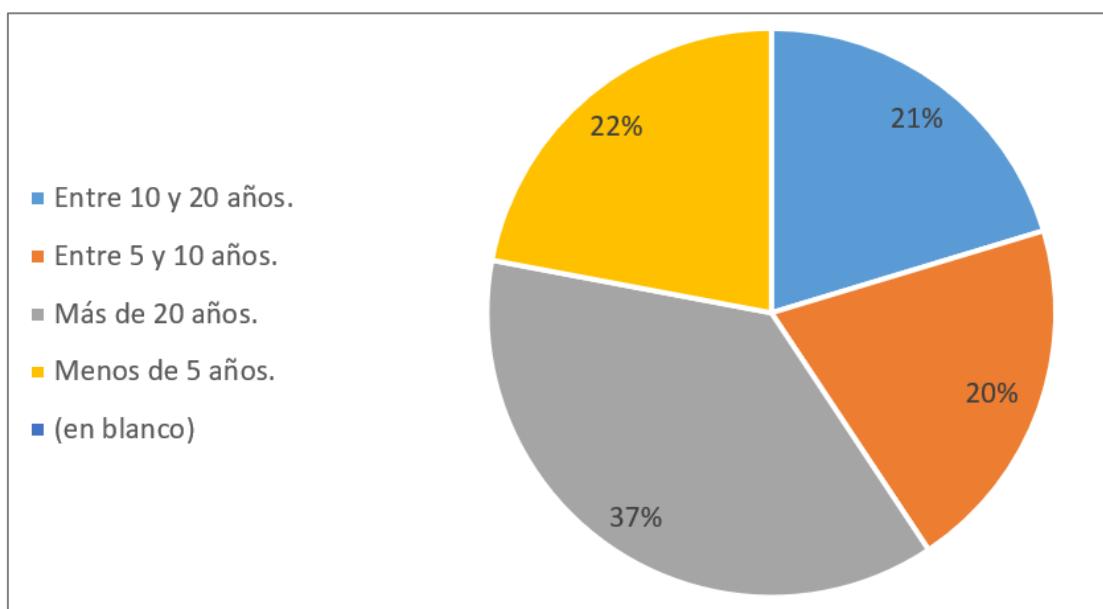


Figura 8. Distribución de los años de experiencia de la muestra. *Fuente: elaboración propia en base a datos de encuesta.*

Respecto al tamaño de la explotación, el 47% de los casos relevados corresponde a establecimientos de entre 100 a 500 hectáreas, mientras que un 32% opera más de 1.000 hectáreas. En menor proporción, representando un 13,33% se encuentran los que cultivan entre 500 a 1.000 hectáreas. Por último, solo el 6,66% corresponde a menos de

100 hectáreas. Esta distribución refleja una participación predominante de medianos y grandes productores, con capacidad operativa ampliada.

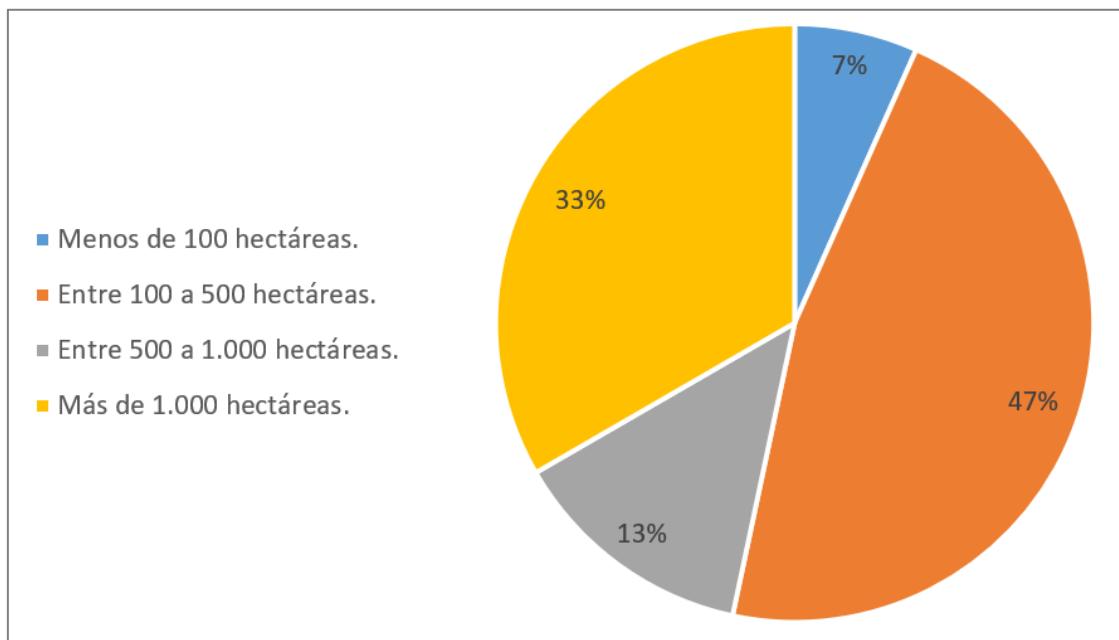


Figura 9. Distribución del tamaño principal de la explotación. *Fuente: elaboración propia en base a datos de encuesta.*

Finalmente, sobre la tenencia de la tierra, se obtuvieron 58 respuestas completas. El 53,44% de los productores trabaja en campos propios, mientras que el 46,55% opera bajo la modalidad de arrendamiento. Estos datos reflejan la flexibilidad del modelo productivo argentino y la relevancia del arrendamiento como herramienta para expandir la superficie operada.

Perfil operativo y de gestión productiva

El análisis de las variables operativas permite comprender las estrategias de producción, adopción tecnológica y capacidades de planificación de los establecimientos agropecuarios incluidos en la muestra.

1. Tecnologías adoptadas y planificación estratégica

El 78,33% de los encuestados tiene incorporado algún tipo de tecnología de agricultura de precisión en su proceso productivo, mientras que el 21,66% restante no utiliza ninguna de estas tecnologías. Dentro de los que han adoptado alguna tecnología, el 87,23% utiliza GPS, el 25,53% emplea sensores, el 38,29% usa drones, el 44,68% emplea software agrícola, el 10,64% aplica siembra variable, el 4,26% utiliza monitores de rendimiento y el 2,13% adopta riego. Estos datos muestran una alta adopción de tecnologías como el GPS y el software agrícola, mientras que herramientas más especializadas, como siembra variable y monitores de rendimiento, tienen una adopción mucho más reducida.

Un 80% de los productores indicó que efectivamente emplea tecnologías ENLIST en sus sistemas productivos, lo que refleja una adopción significativa de esta tecnología en el control de malezas. La soja ENLIST es una herramienta clave para el manejo de malezas resistentes, ya que combina resistencia a herbicidas como glifosato y 2,4-D, lo que facilita el control de especies difíciles de erradicar. Su adopción ha crecido notablemente en Argentina, especialmente en regiones con alta presión de malezas, lo que ha llevado a una mayor inversión por parte de los productores en la implementación de esta tecnología para mejorar el rendimiento y reducir costos asociados al control de malezas (Corteva, 2022; Agroempresario, 2022; Don Mario, 2022).

Esta dispersión en los porcentajes confirma que la adopción de tecnologías es desigual: algunas herramientas (GPS y Enlist) están ampliamente difundidas, mientras que otras como riego, monitores y siembra variable tienen un uso marginal. Esta constatación refuerza la necesidad de considerar la heterogeneidad tecnológica en el análisis de eficiencia y en las políticas de transferencia tecnológica.

En relación con la planificación operativa, el 58,9% de los productores indicó revisar sus estrategias antes de cada temporada, un 32,1% lo hace trimestralmente y solo un 8,9% espera a situaciones de emergencia. Esta información sugiere que una mayoría de los productores mantiene un enfoque anticipatorio y sistemático en la gestión del ciclo productivo.

2. Prácticas agronómicas y sustentabilidad

En términos de prácticas agrícolas, el método de siembra directa es el predominante, utilizado por el 78,9% de los productores. Un 17,5% combina siembra directa con métodos convencionales, y solo un 3,5% se mantiene exclusivamente en el esquema convencional.

En relación con las rotaciones de cultivos, el 93% declaró haber implementado rotaciones con soja en la última campaña. La frecuencia de rotación es alta: el 86% rota todos los años, mientras que el resto lo hace con menor periodicidad.

Las prácticas sostenibles también tienen una presencia destacada. La mayoría de los productores mencionó conservación del suelo y rotación de cultivos como estrategias prioritarias, combinadas en muchos casos con reducción del uso de químicos y eficiencia en el uso del agua. En conjunto, se observa una orientación general hacia prácticas de sostenibilidad, con diferentes grados de profundidad y complejidad técnica.

Consultados por los principales desafíos enfrentados en la campaña, 57 productores respondieron que el mayor obstáculo fue la fluctuación de precios (54,4%), seguido por el cambio climático (21,1%). Otras preocupaciones incluyeron el contexto político-económico (7%), plagas, malezas y dificultades en el acceso a insumos y financiamiento, aunque en menor proporción.

3. Costos, rendimientos y factores de eficiencia

El rendimiento promedio de soja y su evolución reciente fueron informados por la mayoría de los encuestados. El 43,6% indicó que el rendimiento ha aumentado respecto de campañas anteriores, el 38,2% señaló que se mantuvo estable y el 18,2% reportó una disminución.

El costo promedio estimado por hectárea se ubica en torno a los USD 1.482, aunque con variabilidad según zona y nivel tecnológico. Entre los insumos que más impactaron en ese costo, se destacan los herbicidas (20,8%), fertilizantes (25%) y el alquiler de tierras (16,7%), en línea con la estructura típica de costos del cultivo de soja.

Los factores que afectan el rendimiento y aquellos que determinan la eficiencia fueron diversos, incluyendo aspectos técnicos (manejo, semillas, fertilización), climáticos (precipitaciones, temperaturas) y organizativos (planificación, logística, asesoramiento). Esta multiplicidad de respuestas confirma que la eficiencia es percibida como una variable multicausal y sistémica.

Por otro lado, la mayoría de los productores se encuentra involucrada en procesos colaborativos y de formación técnica. El 91,1% declaró contar con asesoramiento técnico, el 61,4% manifestó participar en esquemas de colaboración con otros actores del sector, y el 77,2% asiste a jornadas técnicas. Estos datos reflejan un ecosistema productivo con alto nivel de profesionalización y vinculación con redes técnicas, comerciales y de conocimiento.

4. Financiamiento y comercialización

El 63,2% de los productores accede a financiamiento externo, mientras que el 36,8% opera con fondos propios. Entre quienes utilizan financiación, las herramientas más frecuentes fueron el canje de granos, las tarjetas agropecuarias y los préstamos bancarios. También se mencionaron cooperativas, sociedades de garantía recíproca y esquemas comerciales directos.

Respecto a la forma de comercialización, el 45,6% lo hace a través de acopiadores o cooperativas, mientras que un 19,3% realiza ventas directas sin intermediarios. Otras formas incluyen operaciones para exportación, ventas en mercados futuros y participación en esquemas combinados. Este panorama refleja una diversidad de estrategias, donde prevalece una estructura tradicional de comercialización, pero con presencia creciente de esquemas más directos o integrados.

En conjunto, el perfil operativo de la muestra revela un conjunto de explotaciones que, si bien presentan grados de heterogeneidad en su tamaño y ubicación, comparten una tendencia clara hacia la planificación estratégica, la adopción de tecnologías y la incorporación de prácticas sostenibles, en un entorno desafiante caracterizado por alta volatilidad e incertidumbre.

Análisis de conglomerados de productores

Con el fin de ordenar la heterogeneidad observada en la muestra se implementó un análisis de conglomerados sobre tres rasgos estructurales y de gestión: superficie operada (ha), años de experiencia declarados y porcentaje de inversión tecnológica. Previamente se estandarizaron las variables mediante z-scores para garantizar comparabilidad en una métrica común y evitar que la escala de “superficie” domine la distancia euclídea empleada por el algoritmo. Sobre esa base se estimó un modelo K-means y, para su lectura visual, se proyectaron las observaciones al plano de dos componentes principales (PCA 2D), lo que permite apreciar la separación entre grupos en un espacio de baja dimensión sin imponer restricciones paramétricas adicionales. Como contraste, se calculó además un clustering jerárquico con criterio de Ward, cuyo dendrograma ofrece una vista alternativa de la estructura de proximidades en la muestra.

La visualización principal se presenta en la Figura 10. Los diagnósticos y gráficos de soporte se presentan en el Anexo II (Figs. A2.1–A2.2), y las tablas en el Anexo III (Tablas A3.1–A3.2).

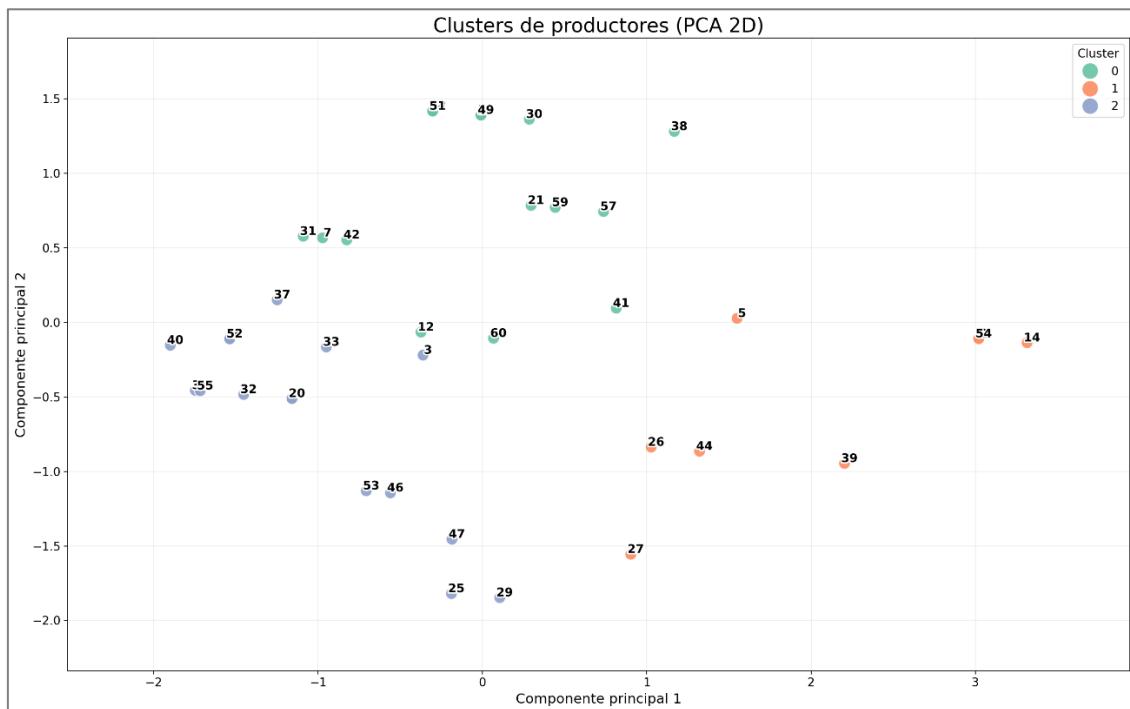


Figura 10. Clusters de productores (PCA 2D). Proyección a dos componentes principales de las variables estandarizadas (z-score): superficie (ha), años de experiencia

y porcentaje de inversión tecnológica. Colores según asignación K-means ($k=3$).

Fuente: elaboración propia en Python en base a datos de encuesta.

La partición con $k=3$ separa nítidamente tres perfiles. Se observa un grupo de gran escala con fuerte adopción tecnológica (superficie media ~ 1.250 ha e inversión $\sim 0,57$), un grupo mediano con baja inversión y trayectoria limitada (superficie ~ 506 ha; inversión $\sim 0,13$; experiencia $\sim 4,5$ años) y un grupo mediano con alta experiencia pero tecnología moderada (superficie ~ 488 ha; inversión $\sim 0,17$; experiencia ~ 22 años). La robustez de esta lectura se respalda con el dendrograma de Ward y el diagnóstico de K (codo y silhouette), incluidos en el Anexo II. Las estadísticas completas por conglomerado y la asignación ID al cluster se reportan en el Anexo III (Tablas 3.1–3.2).

La elección del número de conglomerados se abordó con dos heurísticas estándar. Por un lado, el método del codo mostró la caída de la inercia al aumentar κ , con ganancias decrecientes a partir de $\kappa = 4$. Por otro lado, el índice silhouette alcanzó su máximo en $\kappa \approx 4$ con un valor promedio cercano a 0,46, indicador de una estructura de grupos de calidad intermedia: la cohesión intra-grupo y la separación inter-grupo son razonables, aunque no extremas, lo cual es esperable en muestras reales con superposición parcial de perfiles. Para mantener parsimonia y facilitar la exposición se reportan gráficos y estadísticas con tres conglomerados, y se deja constancia de que con cuatro grupos surge un subconjunto adicional de perfil intermedio sin alterar las conclusiones cualitativas principales. Los gráficos de soporte para la elección de K se reportan en el Anexo (Anexo II y Anexo III).

Los resultados del K-means con $\kappa = 3$ describen tres tipologías bien diferenciadas. Un primer grupo reúne explotaciones de gran escala, con una superficie promedio del orden de 1.250 ha, niveles elevados de inversión tecnológica (en términos relativos dentro de la muestra) y una dotación de experiencia cercana a dos décadas. Estas unidades tienden a ubicarse hacia los extremos del primer componente principal en la proyección PCA, lo que sugiere que la variable “tamaño” aporta un eje dominante de variación combinado con la adopción tecnológica. Un segundo grupo agrupa establecimientos medianos con superficie media próxima a 500 ha, baja inversión

tecnológica y trayectoria limitada (poco más de cuatro años en promedio); en el plano PCA se los observa cercanos entre sí y relativamente apartados del grupo de gran escala, lo que coincide con la lectura del dendrograma, donde se forman racimos compactos a distancias cortas. Un tercer grupo, también de escala media (alrededor de 480–500 ha), se caracteriza por alta experiencia (más de 22 años en promedio) combinada con inversión tecnológica moderada.

La comparación de medias y medianas dentro de cada cluster confirma estas caracterizaciones y sugiere que “experiencia” y “tecnología” no siempre coevolucionan: existen perfiles con trayectoria consolidada pero adopción tecnológica contenida, y perfiles de gran escala con fuerte inversión en herramientas, rasgo consistente con estrategias de intensificación vía capital.

El dendrograma de Ward refuerza esta lectura al mostrar dos macroagrupamientos iniciales: por un lado, el conjunto vinculado a unidades de mayor escala y, por el otro, los subconjuntos de escala media diferenciados por experiencia e inversión tecnológica. La coincidencia cualitativa entre la partición jerárquica y K-means sugiere que la estructura de distancias en el espacio estandarizado es estable a la elección del algoritmo, lo cual aporta robustez al ejercicio descriptivo. La proyección PCA facilita, además, la identificación de gradientes: el primer componente está asociado principalmente a escala e inversión tecnológica, mientras que el segundo recoge variación ligada a la experiencia, de modo que los grupos se separan sobre ambos ejes y no solo por tamaño.

En términos de interpretación, la segmentación permite distinguir tres estrategias de organización productiva presentes en la muestra: una estrategia intensiva en capital y escala, otra apoyada en capital humano y aprendizaje acumulado con menor inversión tecnológica relativa, y una tercera de adopción tecnológica acotada y menor trayectoria. Esta tipología será útil para el análisis posterior, en la medida en que habilita comparar los puntajes de eficiencia y otros indicadores de desempeño entre grupos, sin necesidad de referir a identificadores individuales. Cabe señalar que la elección de variables para el clustering se circunscribió a rasgos disponibles y comparables para toda la muestra; por lo tanto, las conclusiones se entienden como descriptivas y no implican relaciones causales. Finalmente, la estabilidad de las agrupaciones al pasar de $\kappa = 3$ a $\kappa = 4$

(donde emerge un subgrupo intermedio con adopción tecnológica parcial) sugiere que las diferencias observadas no responden a outliers puntuales sino a perfiles efectivamente presentes en los datos.

Análisis de eficiencia

I. Aplicación del Análisis Envolvente de Datos (DEA)

Con el objetivo de medir cuán eficientemente se usan los recursos en las explotaciones encuestadas, se aplicó el Análisis Envolvente de Datos (DEA) en su versión de rendimientos variables de escala (VRS) y orientación a insumos. En este marco, la eficiencia técnica se entiende como la capacidad de alcanzar el mismo rendimiento de soja utilizando una menor cantidad de recursos. La orientación a insumos es consistente con la lógica de gestión agronómica: el productor decide qué y cuánto aplicar, más que el precio o el nivel final de producción.

La estimación abarca 46 explotaciones bonaerenses con información completa. Se utilizó como producto el rendimiento de soja (kg/ha) y, como insumos, la superficie trabajada, el uso de fertilizantes, herbicidas, insecticidas y fungicida (capturados como decisiones de manejo), la tercerización de siembra y de cosecha, y los años de experiencia del productor. Los costos monetarios se excluyeron porque introducían ruido en la frontera; además, el DEA no requiere homogeneizar unidades monetarias para comparar combinaciones de factores. El problema se resolvió por programación lineal y, para cada unidad, se obtuvo su puntaje de eficiencia $\theta \in [0,1]$ y el conjunto de referentes que define su proyección sobre la frontera.

Los resultados muestran heterogeneidad y, a la vez, un grupo relevante en la frontera. El promedio de eficiencia fue 0,663 y la mediana 0,643, con un mínimo de 0,241 y un máximo de 1,000. Aproximadamente el 43,5% de las explotaciones resultó eficiente ($\theta=1$). Operativamente, una unidad con $\theta=0,66$ podría en promedio reducir 34% su uso total de insumos sin resignar rendimiento si replica la combinación de decisiones de sus referentes. Esta lectura es radial: en variables binarias de manejo (por

ejemplo, “usa/no usa fungicida”) el mensaje no es “usar una fracción menor”, sino acercar el patrón de decisiones al de las unidades de la frontera.

La Figura 11 sugiere una distribución asimétrica con dos rasgos claros: un bloque de observaciones en la zona media ($\approx 0,25-0,50$) y un pico en el extremo derecho asociado a la eficiencia plena. La Figura 12 confirma esa asimetría: la mediana se ubica por encima de 0,60, con una caja que se extiende hacia valores altos y algunos casos rezagados en la cola inferior. Esta morfología es habitual en aplicaciones DEA porque la frontera actúa como techo natural y concentra observaciones en 1, mientras que las ineficiencias relativas se dispersan según la calidad del mix de manejo.

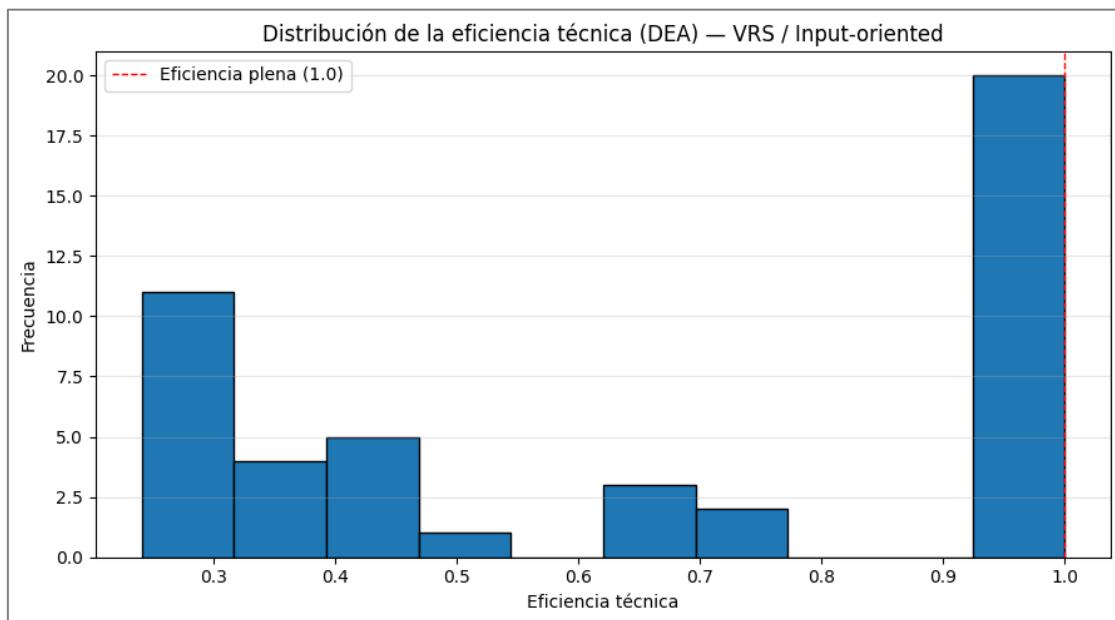


Figura 11. Distribución de la eficiencia técnica (DEA) — VRS, orientación a insumos. Histograma de θ ; línea punteada en 1,0 (eficiencia plena). *Fuente: elaboración propia en Python.*

Un aspecto central para la transferencia de aprendizaje es la concentración de referentes. La construcción de la frontera muestra que un subconjunto acotado de explotaciones opera repetidamente como benchmark. Entre ellas destacan los IDs 30, 31, 32 y 43 (los de mayor frecuencia) y, por detrás, otros casos que también aparecen de forma reiterada. Esta concentración es informativa por dos razones: indica que existen

combinaciones de escala, manejo y tercerización que funcionan de facto como estándares observados en la muestra, y facilita la comparación operativa, porque brinda puntos de apoyo concretos para revisar procesos y reasignar prácticas. El listado completo de referentes por productor y la frecuencia con que cada ID actúa como tal se reporta en el Anexo IV.

La lectura sustantiva de los puntajes refuerza la idea de brechas gestionables. El núcleo eficiente convive con un conjunto amplio de explotaciones en rangos intermedios; la distancia a la frontera no parece explicarse por un único factor aislado, sino por combinaciones de decisiones (qué insumos se aplican, cómo se secuencian, qué tareas se tercerizan y con qué oportunidad). Cuando las variables son continuas (por ejemplo, superficie) la proyección radial provee una meta cuantitativa implícita en la solución; cuando las variables son dicotómicas, la proyección es cualitativa y señala la dirección del ajuste (alinear el patrón de manejo con el de los referentes).

En conjunto, el ejercicio sugiere que existe un grupo sólido de establecimientos técnicamente eficientes y, al mismo tiempo, márgenes sustantivos de mejora para el resto. La evidencia es coherente con un diagnóstico donde la gestión (más que una sola característica estructural) organiza las diferencias de desempeño. Esta información resulta útil para orientar la priorización de prácticas y la focalización de asistencia técnica: primero, identificar el referente relevante de cada productor; luego, contrastar las decisiones de manejo y, finalmente, cerrar brechas replicando combinaciones que ya demostraron ser eficientes en condiciones comparables.

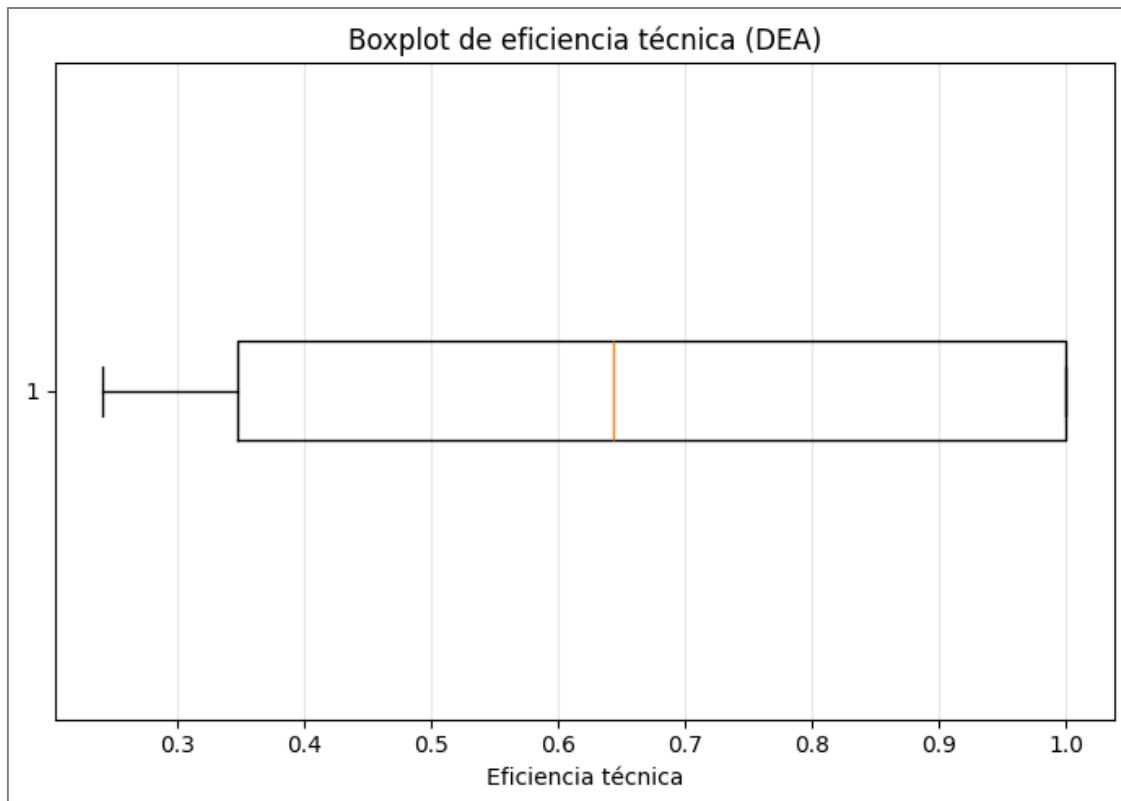


Figura 12. Boxplot de la eficiencia técnica (DEA). Resumen de la dispersión de θ en la muestra. *Fuente: elaboración propia en Python.*

El detalle de puntajes, referentes y proyecciones radiales figura en el Anexo IV.

2. Eficiencia técnica por tipologías de productores

Con el fin de verificar si la eficiencia varía según el perfil productivo, se vinculó la asignación de clusters ($k = 3$, construidos con superficie, años de experiencia y % de inversión tecnológica) con los puntajes de eficiencia DEA. El cruce alcanza 31 casos con ambas fuentes. A nivel descriptivo, las distribuciones por grupo son próximas entre sí: en el cluster 0 (medianos con alta experiencia y adopción tecnológica moderada) la media es 0,698 (mediana 1,000; desvío 0,374), con 58,3% de unidades eficientes; en el cluster 1 (gran escala, alta inversión tecnológica) la media asciende a 0,723 (mediana 0,760; desvío 0,301), con 42,9% eficientes; en el cluster 2 (medianos con baja inversión y menor experiencia) la media es 0,647 (mediana 0,589; desvío 0,330), con 41,7%

eficientes. Las medianas ya sugieren un patrón: el grupo de mayor experiencia concentra más observaciones sobre la frontera (mediana = 1), mientras que el de menor trayectoria e inversión presenta valores centrales más bajos.

Para dimensionar la incertidumbre muestral, se calcularon intervalos de confianza aproximados del 95% para la media de cada grupo ($\text{media} \pm 1,96 \cdot \text{SE}$). Dadas las varianzas y los tamaños muestrales observados, los rangos son amplios y entre-solapados:

- Cluster 0: ~ [0,49; 0,91]
- Cluster 1: ~ [0,50; 0,95]
- Cluster 2: ~ [0,46; 0,83]

Este solapamiento anticipa lo que confirman los test: la prueba no paramétrica de Kruskal–Wallis no rechaza la igualdad de distribuciones ($H = 0,206$; $p = 0,9021$), y los contrastes pareados (Bonferroni) tampoco arrojan diferencias significativas. El tamaño de efecto asociado es muy bajo, lo que respalda la conclusión de ausencia de diferencias estadísticamente detectables entre tipologías con esta muestra.

Desde una perspectiva sustantiva, el resultado es claro: la eficiencia técnica aparece transversal a los perfiles de tamaño, experiencia e inversión considerados para el clustering. Dentro de cada grupo conviven explotaciones en la frontera y otras con márgenes de mejora. El contraste “gran escala/alta tecnología” vs. “medianos/baja tecnología” no garantiza diferencias en eficiencia; más bien, la combinación fina de decisiones de manejo y organización (qué insumos se aplican, cómo se terceriza, qué prácticas se integran y en qué momentos) sigue siendo el eje explicativo. La mediana = 1 en el cluster de alta experiencia sugiere, no obstante, que el aprendizaje acumulado podría asociarse a una mayor probabilidad de operar en la frontera; es un indicio, no una prueba concluyente, y debe leerse con cautela por el tamaño muestral y el “techo” propio del DEA (acumulación de 1s).

El boxplot por cluster (Figura 13) ilustra bien la situación: dispersiones internas amplias, presencia de observaciones eficientes en los tres grupos y colas inferiores que evidencian unidades con brechas relevantes. En el grupo de gran escala/alta tecnología,

la media es levemente superior, pero la mediana no alcanza a 1 y la variabilidad sugiere comportamientos heterogéneos: no toda granja tecnificada es eficiente si la combinación de decisiones no acompaña. En el grupo de baja inversión y menor experiencia, la distribución es más “pesada” en niveles medios/bajos, aunque también aparecen casos sobre la frontera. El patrón refuerza la lectura de que la eficiencia es alcanzable desde distintos puntos de partida, siempre que se ajuste el mix de manejo hacia los referentes.

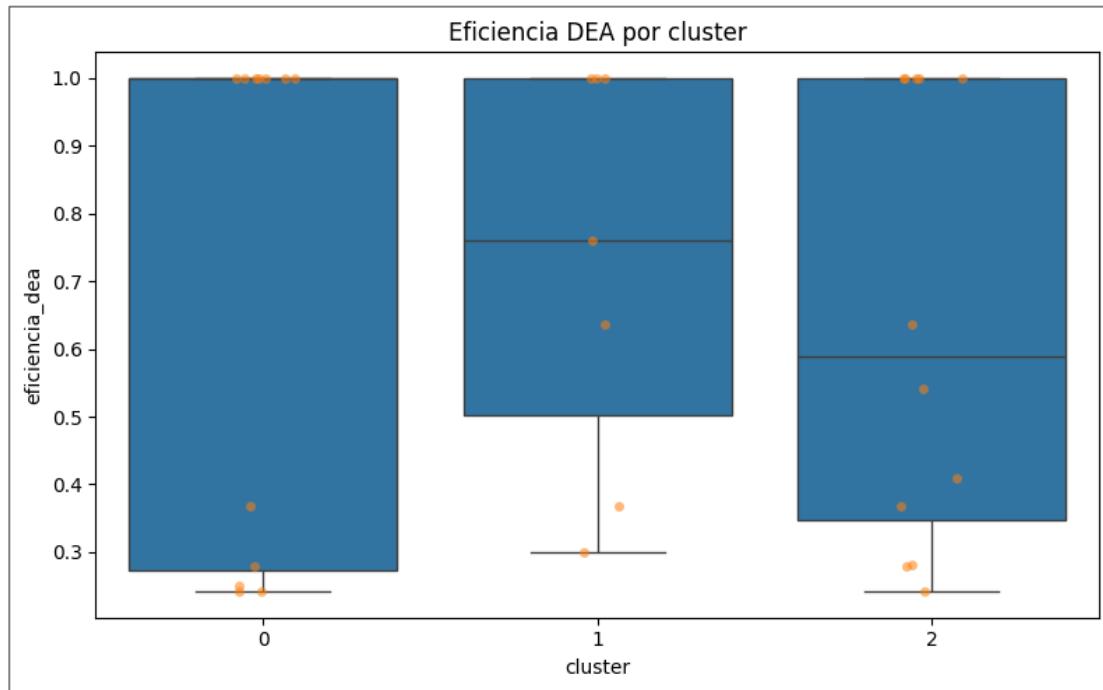


Figura 13. Eficiencia técnica (DEA) por cluster de productores. Boxplots de θ por grupo con puntos individuales. Tamaños muestrales del cruce: $n_0 = 12$, $n_1 = 7$, $n_2 = 12$. *Fuente: elaboración propia en Python.*

El contraste estadístico y el resumen ampliado por tipologías se incluyen en el Anexo IV (Tabla 4.5; Fig. 4.3).

Evaluación del Modelo de Tobit

Para indagar en los determinantes de la eficiencia técnica estimada por DEA, se aplicó un modelo de regresión Tobit con doble censura en los límites [0, 1]. Esta

especificación es apropiada cuando la variable dependiente se encuentra acotada y exhibe acumulación en los bordes (en este caso, particularmente en 1), situación en la que una regresión lineal ordinaria produciría estimaciones sesgadas tanto en los coeficientes como en la varianza de los errores. La forma funcional asume una variable latente de eficiencia, lineal en los parámetros y con errores normales homocedásticos; lo observado se censura en 0 si la variable latente cae por debajo de ese umbral y en 1 si lo supera. En este marco, la constante se interpreta como el valor esperado de la eficiencia latente en la categoría base de la explicativa (dummy=0) y el coeficiente como el cambio en ese valor latente al pasar de 0 a 1, siempre condicionado al mecanismo de censura.

La estrategia empírica fue deliberadamente parsimoniosa. Se estimó un Tobit univariado por cada explicativa de interés (modalidad societaria, condición de propietario, acceso a financiamiento, rotación de cultivos y uso de semilla propia) para preservar claridad interpretativa dado el tamaño muestral y evitar reciclar información de los insumos del DEA como regresores. La base surge del merge por ID entre la tabla de eficiencia y la de caracterización, restringiendo al subconjunto con datos completos; por seguridad, la eficiencia se recortó en [0, 1].

Los resultados para modalidad societaria son consistentes con la intuición y con la lectura cualitativa previa: en la categoría base (sociedad=0) la constante del Tobit se ubica en 0,913 (escala latente) y el coeficiente de “sociedad” es negativo y no significativo ($p \approx 0,43$). En propiedad de la tierra el patrón es simétrico: cuando propietario=0 el nivel latente esperado es 0,770; el coeficiente de ser propietario es positivo y no significativo ($p \approx 0,58$). En financiamiento aparece el caso más sugestivo: el coeficiente es negativo y no significativo al 10% ($p \approx 0,26$), pese a la expectativa de signo positivo si el crédito suaviza restricciones; esto abre hipótesis sobre uso y condiciones del crédito (monto, tasa, plazo, destino) o sobre composición muestral. En rotación de cultivos el coeficiente es positivo y no significativo ($p \approx 0,81$), lo que sugiere que, sin controles, su efecto marginal queda diluido entre prácticas que interactúan. En semilla propia se observa la señal más clara: el coeficiente es positivo y significativo al 5% ($p \approx 0,027$), con constante base 0,620, compatible con ganancias de eficiencia

asociadas a provisión/adaptación varietal o reasignaciones de costos; es una asociación, no una prueba causal.

En conjunto, no emerge un único atributo estructural u organizacional que, por sí mismo, determine la eficiencia. Sociedad y propiedad no ofrecen evidencia precisa; la rotación mantiene el signo esperado sin significancia; el financiamiento plantea una paradoja empírica que invita a desagregar “acceso” de “calidad del crédito”; y semilla propia destaca como la asociación positiva más consistente en este corte. La acumulación de observaciones censuradas en 1 —propia de los puntajes DEA— amplía la incertidumbre y obliga a leer los tamaños de efecto como relaciones en la escala latente, no como impactos porcentuales sobre la eficiencia observada. Para documentación, los coeficientes, errores estándar, estadísticos z, valores-p y medidas de ajuste, junto con conteos de censura, se sistematizan en *tobit_univariado_5vars.xlsx*.

Como línea de robustez y extensión, se sugiere un Tobit multivariado mínimo (tamaño, adopción tecnológica, experiencia) para verificar estabilidad de signos —en especial financiamiento y semilla— y, en paralelo, especificaciones alternativas para datos acotados (logit/probit fraccional o modelos de dos partes) que separen la probabilidad de estar en la frontera de la intensidad de eficiencia en (0,1).

Para contrastar determinantes de la eficiencia, se estimaron cinco Tobit univariados con doble censura en [0, 1] sobre la eficiencia DEA. La constante se interpreta como el nivel de eficiencia latente cuando la dummy vale 0; el coeficiente es el cambio latente al pasar de 0 a 1, condicionado a la censura. En la muestra (N=46) hubo 26 observaciones no censuradas y 20 censuradas en 1,0, y todas las corridas convergieron. El resultado más nítido corresponde a semilla propia (coeficiente positivo, $p \approx 0,03$). En sociedad el coeficiente es negativo y no significativo; en propietario positivo y no significativo; financiamiento muestra un signo negativo, pero no significativo ($p \approx 0,26$); y rotación es positivo y no significativo. En suma, solo semilla propia exhibe evidencia estadística de asociación positiva con la eficiencia en este corte; el resto no es concluyente bajo especificaciones univariadas.

La evidencia debe leerse con el prisma de la censura y del tamaño muestral. En el set focal de cinco explicativas, el patrón es claro: semilla propia es la única señal

positiva y significativa; sociedad y propietario no resultan significativas (con signos negativo y positivo, respectivamente); financiamiento exhibe signo negativo y no significativo; rotación sostiene un signo positivo sin precisión estadística. Al ampliar el barrido a otras variables de caracterización, los estadísticos z ordenan la intensidad relativa de los efectos, pero los valores-p rara vez bajan de 0,05; ello refleja, sobre todo, poder estadístico limitado (N moderado, masa en 1 por la frontera DEA y varianza no explicada por el univariado). Por eso, estas salidas deben entenderse como pistas para priorizar hipótesis, no como pruebas concluyentes.

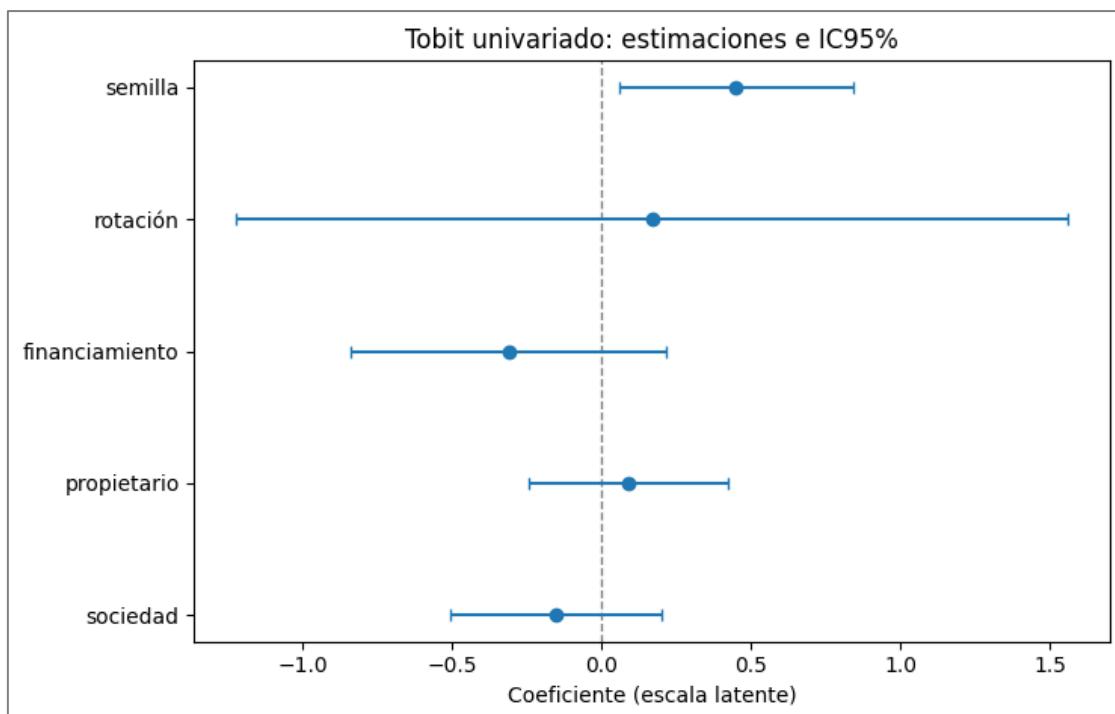


Figura 14. Coeficientes del Tobit univariado (eficiencia DEA $\in [0,1]$).
Puntos = estimaciones; barras = IC95%; línea vertical en 0 (sin efecto). *Fuente: elaboración propia en Python.*

En cada modelo Tobit se contrasta la hipótesis nula $H_0: \beta = 0$ (la variable no tiene efecto sobre la eficiencia) frente a la alternativa $H_1: \beta \neq 0$ (la variable sí tiene efecto). Se utiliza un nivel de significancia $\alpha = 0.10$ y se rechaza H_0 cuando $p < 0.10$. El coeficiente en dummies indica el cambio en la eficiencia latente al pasar de 0 a 1, condicionado a la censura; la constante es la eficiencia latente esperada en la categoría

base. Los coeficientes del Tobit no son porcentajes sobre la eficiencia observada y cualquier lectura causal requiere modelos con controles y atención a composición. En síntesis, el núcleo institucional (sociedad, propiedad) no muestra efectos precisos; financiamiento plantea una paradoja empírica que amerita abrir la caja de la calidad del crédito; rotación sostiene el signo esperado sin significancia; y semilla propia aparece como el determinante positivo más consistente.

Los resultados completos de estas estimaciones, junto con sus parámetros estadísticos, se presentan en el Anexo V.

Conclusiones

Conclusiones de los resultados

El análisis DEA bajo retornos variables a escala (VRS) y con orientación a insumos muestra una heterogeneidad notable entre las explotaciones de soja de la región de Tandil y zonas aledañas, aunque simultáneamente identifica un núcleo de explotaciones que operan consistentemente en la frontera de eficiencia.

El puntaje promedio observado fue 0,663, con mediana 0,643, y el 43,5 % de las unidades resultaron eficientes según la estimación realizada. Estas cifras implican que una proporción relevante de productores alcanza niveles de desempeño cercanos al óptimo técnico definido por la frontera estimada, mientras que el restante presenta un margen de mejora operativo cuantificable: en promedio, las explotaciones no eficientes podrían reducir alrededor de un 34 % del uso conjunto de insumos sin perder rendimiento si ajustaran sus patrones de decisión hacia los de sus referentes.

La distribución asimétrica de los puntajes (con acumulación en valores próximos a la eficiencia plena) y la repetida aparición de un subconjunto acotado de explotaciones como referentes sugieren que existen prácticas locales replicables que permiten cerrar brechas importantes sin necesidad de transformaciones estructurales radicales. Estos patrones empíricos guardan coherencia con la literatura metodológica sobre análisis de eficiencia (Farrell, 1957; Coelli et al., 2005; Fried, Lovell & Schmidt, 2008), y con estudios empíricos regionales que muestran que la heterogeneidad entre productores suele explicarse por diferencias en manejo y calidad de insumos más que por la sola dimensión de la explotación (Rodríguez Sperat, Brugiafreddo & Raña, 2017; FAUBA, 2023).

El cruce de los puntajes DEA con las tipologías obtenidas por conglomerados confirma que la eficiencia técnica no se distribuye de manera estrictamente categórica por tamaño, experiencia o nivel de inversión tecnológica: en cada cluster conviven explotaciones en la frontera con otras que exhiben márgenes de mejora. Este resultado pone de relieve que la estructura productiva por sí sola no determina la posición respecto de la frontera, y que la gestión interna (incluyendo la elección y calidad de la semilla, la

calibración y calendarización de las aplicaciones, la intensidad y oportunidad del laboreo y la organización operativa) actúa como factor decisivo para el desempeño.

Los modelos Tobit univariados, con doble censura entre 0 y 1, aportaron señales relevantes: la variable “semilla propia” mostró una asociación positiva y estadísticamente significativa con los puntajes de eficiencia; otras variables esperadas—modalidad societaria, condición de propietario, rotación de cultivos y acceso a financiamiento—mostraron signos compatibles con la teoría pero sin significancia robusta en este primer abordaje, lo que obliga a interpretar dichas asociaciones con prudencia y como insumo para hipótesis de trabajo posteriores.

La presencia de observaciones amontonadas en 1 (propia del enfoque DEA) y el tamaño muestral moderado explican, en parte, la limitada precisión del “segundo estadio” y justifican la recomendación metodológica de aplicar especificaciones alternativas y remuestreo en trabajos siguientes (Simar & Wilson, 2007; Simar & Wilson, 2011).

Los hallazgos locales no son una anomalía aislada: estudios argentinos y latinoamericanos evidencian patrones de heterogeneidad y margen de mejora similares en actividades agropecuarias diversas. Investigaciones sobre eficiencia en agricultura familiar y en producciones regionales han documentado diferencias en eficiencia técnica entre explotaciones que se explican por prácticas de manejo y calidad de insumos, más que por la mera escala productiva (Rodríguez Sperat et al., 2017). Informes sobre productividad total de factores en la agricultura nacional muestran que existe espacio para mejoras técnicas sostenidas que, acumuladas en el tiempo, resultan relevantes para la competitividad del sector (Baumann Fonay & Cohan, 2018; Grotz, 2020). Estudios agronómicos y de evaluación de cultivares del INTA respaldan la idea de que calidad de semilla, ajuste de fechas de siembra y rotación son determinantes concretos del rendimiento en soja, y por tanto plausibles explicaciones de las asociaciones observadas en este trabajo (INTA, 2024). A su vez, la literatura internacional sobre eficiencia productiva en cultivos agrícolas confirma que la adopción de mejores prácticas, la calidad del material de siembra y la integración institucional y comercial explican variaciones de eficiencia entre productores y regiones (Wang & Shi, 2020; Selorm et al., 2023). En conjunto, la evidencia empírica local y externa fortalece la validez de los

resultados: no son anomalías metodológicas sino síntomas de un patrón regional donde la gestión y la calidad operativa son vectores fundamentales para acercarse a la frontera de eficiencia.

Conclusión general del trabajo

Este trabajo constituye una línea de base cuantitativa rigurosa sobre la eficiencia técnica del cultivo de soja en Tandil, aportando resultados reproducibles y políticamente relevantes. Al integrar datos primarios de encuesta con técnicas no paramétricas (DEA VRS orientado a insumos) y análisis econométricos del segundo estadio (Tobit censurado), se logró cuantificar brechas operativas, identificar explotaciones referentes y ofrecer una evaluación preliminar de determinantes asociados a la eficiencia. El hallazgo central es doble: por un lado, una proporción sustantiva de explotaciones (43,5 %) opera con eficiencia plena en el corte analizado, lo que demuestra la existencia de referentes locales con prácticas efectivas; por otro, las explotaciones no eficientes disponen de un margen de mejora operacional importante —estimado en torno al 34 % de reducción de insumos potencial sin pérdida de rendimiento— que es alcanzable mediante cambios de gestión y adopción de prácticas ya presentes en el territorio. Esta constatación tiene efectos prácticos directos: permite diseñar estrategias de extensión y transferencia que prioricen prácticas replicables y de alto impacto relativo, en lugar de recomendaciones genéricas centradas exclusivamente en ampliar escala o elevar inversión.

Desde la perspectiva teórica y metodológica, el trabajo confirma la utilidad de la estrategia DEA + análisis del segundo estadio para estudios de eficiencia en agricultura (Coelli et al., 2005; Fried et al., 2008), pero también evidencia sus límites prácticos cuando la muestra es moderada y existe acumulación de observaciones en la frontera. Por ello, las inferencias sobre determinantes deben considerarse como evidencia asociativa y como guía para análisis multivariados más profundos; no como pruebas de causalidad concluyentes. La semilla propia aparece como un insumo operativo con asociación significativa, lo que coincide con la evidencia agronómica e institucional sobre la relevancia de la calidad y procedencia del germoplasma para el rendimiento (INTA, 2024; Morla, 2022). Asimismo, el hecho de que variables

estructurales clásicas (tamaño, inversión tecnológica, experiencia) no resulten por sí solas determinantes claros subraya la importancia de focalizar en prácticas y en procesos de gestión, así como en mecanismos institucionales que faciliten la adopción de dichas prácticas.

En términos de política y de extensión, los resultados señalan un camino pragmático: focalizar esfuerzos en asistencia técnica orientada a replicar las prácticas de los referentes, promover la mejora de la calidad de la semilla mediante análisis y certificación, fortalecer acceso a información de mercado y calendarios agronómicos, y diseñar esquemas de financiamiento y cadenas comerciales que reduzcan fricciones operativas. La evidencia local y regional sugiere que intervenciones así diseñadas pueden producir mejoras técnicas reales y sostenibles sin exigir transformaciones estructurales inmediatas, aunque la acción pública debe ser sensible al contexto y acompañada de monitoreo riguroso para evitar efectos no deseados. En suma, la tesis no sólo diagnostica una realidad medible, sino que también entrega insumos accionables para políticas y prácticas con retorno operativo directo.

Futuras direcciones de investigación

Para consolidar y extender las conclusiones aquí presentadas es imprescindible avanzar en tres ejes complementarios: mayor profundidad temporal, mayor riqueza de variables explicativas y mayor rigor en el segundo estadio metodológico. En primer lugar, un diseño longitudinal que recoja panel de explotaciones a lo largo de varias campañas permitirá evaluar la persistencia de la eficiencia, la dinámica de aprendizaje entre productores y el impacto duradero de intervenciones puntuales; solo con panel será posible distinguir cambios de eficiencia de desplazamientos de la frontera por innovación tecnológica (índices de Malmquist y análisis dinámicos son herramientas pertinentes para ese propósito). En segundo lugar, la incorporación sistemática de variables climáticas de alta resolución, calidad de suelo (textura, nutrientes, pH, materia orgánica), medidas de riesgo hídrico intra-campaña, y variables de infraestructura y mercado (distancia a puertos o acopios, costo de transporte, acceso a almacenamiento) aumentará considerablemente la capacidad explicativa de los modelos y reducirá sesgos por variables omitidas que hoy pueden afectar las asociaciones detectadas. Además, la

recolección de datos de manejo agronómico con granularidad —dosis y fechas exactas de fertilización y fitosanitarios, variedad y calidad de semilla, densidad de siembra, prácticas de conservación de suelo, tercerización de labores— es imprescindible para identificar cuáles medidas concretas ofrecen mayor retorno por unidad de insumo.

Metodológicamente, el “segundo estadio” debe fortalecerse mediante especificaciones multivariadas más adecuadas al problema de variables acotadas y a la heterogeneidad no observada. La estimación por Tobit puede complementarse o sustituirse por regresiones fraccionales, modelos de dos partes (hurdle models) que separen la probabilidad de alcanzar la frontera de la magnitud de la ineficiencia, modelos mixtos o jerárquicos que capturen efectos no observados a nivel de productor o zona, y técnicas de corrección para sesgo de selección cuando corresponda. Asimismo, resulta necesario aplicar procedimientos de remuestreo (bootstrap u otras variantes) para cuantificar la incertidumbre en los puntajes DEA y construir intervalos de confianza para las estimaciones del segundo estadio, tal como recomiendan Simar & Wilson (2007, 2011). Comparar estimaciones DEA bajo VRS y CRS permitirá, además, evaluar la presencia de economías o deseconomías de escala y orientar recomendaciones sobre tamaño óptimo de explotación o modalidades de agrupamiento productivo.

Finalmente, se sugiere complementar los estudios observacionales con intervenciones piloto y evaluaciones de impacto (ensayos experimentales o cuasi-experimentales) que prueben en campo el efecto de prácticas identificadas como prometedoras en este trabajo: mejora y certificación de semilla, paquetes de manejo estandarizados, asistencia técnica focalizada, instrumentos financieros diseñados para capital de trabajo o tecnología, y modelos de comercialización que reduzcan incertidumbre de precios. Paralelamente, es esencial integrar análisis de sostenibilidad: medir consumo de agua, uso energético y emisiones asociadas para garantizar que mejoras en eficiencia técnica no generen externalidades ambientales indeseadas. El avance por estas líneas consolidará la robustez de las recomendaciones y permitirá transformar el diagnóstico en programas de política y extensión con impacto verificable y medible.

Bibliografía

- Agroavances. (2024, marzo 19). *El cambio climático y su impacto en la producción agrícola*. <https://agroavances.com/noticias/detalle.php?idNot=5500>
- Agroempresario. (2022). *Enlist: La tecnología que lleva a un nuevo nivel el control de malezas en cultivos de soja*. <https://agroempresario.com>
- Banco Mundial. (2015). *Crecimiento y productividad total de factores en la agricultura: Evidencia empírica en América Latina*. <https://documents1.worldbank.org/curated/ar/970151468197997810/pdf/104000-WP-P155040-Crecimiento-y-Productividad-Total-de-Factores-en-la-Agricultura-Lema-PUBLIC-SPANISH.pdf>
- Banco Mundial. (2020). *Productividad agrícola y eficiencia en el uso de los recursos*. Washington, D.C.: Banco Mundial.
- Baumann Fonay, G., & Cohan, L. (2018). *Productividad total de factores en la agricultura argentina: una aproximación empírica*. Buenos Aires: Ministerio de Hacienda.
- Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20(2), 325–332.
- Berbel, J., & Rodríguez-Ocaña, A. (2007). Análisis envolvente de datos (DEA): Una herramienta para la evaluación de la eficiencia en la gestión agraria. *Revista de Estudios Agro-Sociales y Pesqueros*, 214, 95–116.
- Bolsa de Comercio de Rosario. (2024, abril 5). *La campaña de soja 2023/24 cierra con récord de exportaciones de aceite y más de USD 16.000 millones netos exportados por el complejo*. <https://www.bcr.com.ar/es/mercados/investigacion-y-desarrollo/informativo-semanal/noticias-informativo-semanal/la-campana-4>

Bolsa de Comercio de Rosario. (2024, abril 12). *El empleo en la cadena sojera: Más de 400 mil puestos de trabajo en Argentina.*
<https://www.bcr.com.ar/es/mercados/investigacion-y-desarrollo/informativo-semanal/noticias-informativo-semanal/empleo>

Bongiovanni, R., & Lowenberg-Deboer, J. (2004). Precision agriculture and sustainability. *Precision Agriculture*, 5(4), 359–387.

Cadena 103. (s.f.). *Estructura productiva del partido de Tandil.*
<https://www.cadena103.com.ar/itandil/estructura.htm>

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6), 429–444.

Cobb, C. W., & Douglas, P. H. (1928). A theory of production. *American Economic Review*, 18(1), 139–165.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis* (2.^a ed.). New York: Springer.

CONICET. (2018). *Productividad total de factores en la agricultura argentina: tendencias y desafíos*. Consejo Nacional de Investigaciones Científicas y Técnicas.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). *Data Envelopment Analysis: A comprehensive text with models, applications, references and DEA-Solver Software*. Springer.

Córdoba, F. (2014). *Análisis espacial de la productividad agrícola y características del suelo*. Universidad Nacional de Córdoba.

Corteva Agriscience. (2022). *Guía de uso de ENLIST para soja*.

CREA. (2024). *Agricultura de precisión: El camino hacia una producción más eficiente y sostenible.* <https://www.contenidoscrea.org.ar/agricultura/agricultura-precision-el-camino-una-produccion-mas-eficiente-y-sostenible-n5327449>

- Cruz Lauracio, J. (2019). *La función de producción Cobb-Douglas en la estimación de la productividad agrícola*. Universidad Nacional Autónoma de México.
- Debertin, D. L. (2012). *Agricultural production economics* (2nd ed.). Macmillan Publishing.
- Díaz, M., & Gómez, L. (2017). Barreras para la adopción de agricultura de precisión en pequeños y medianos productores. *Revista de Economía y Agronegocios*, 15(2), 45–62.
- Don Mario. (2022). *Nuevo portfolio de soja Don Mario y un sistema que supera las expectativas*. <https://www.donmario.com>
- EEAOC. (2020). *Eficiencia técnica en cultivos extensivos: prácticas agrícolas y productividad*. Estación Experimental Agroindustrial Obispo Colombres.
- El Eco de Tandil. (2018, febrero 17). *La mayor superficie de cultivo del partido se destina a la soja*. <https://www.eleco.com.ar/interes-general/la-mayor-superficie-de-cultivo-del-partido-se-destina-a-la-soja>
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester: Wiley.
- FAUBA. (2023). *Informe técnico sobre eficiencia productiva en sistemas agrícolas argentinos*. Facultad de Agronomía, Universidad de Buenos Aires.
- Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society: Series A (General)*, 120(3), 253–290.
- Fried, H. O., Lovell, C. A. K., & Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity change* (2.^a ed.). New York: Oxford University Press.
- Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science*, 327(5967), 828–831.

Google Research. (2020). *Welcome to Colaboratory: A research tool for machine learning education and research.* Google.
<https://research.google.com/colaboratory/>

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Grotz, A. (2020). *Eficiencia y productividad en la agricultura argentina: un análisis regional.* Buenos Aires: CEPAL.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Andover: Cengage Learning.

Hatfield, J. L., Boote, K. J., Kimball, B. A., Ziska, L. H., Izaurrealde, R. C., Ort, D., Thomson, A. M., & Wolfe, D. (2011). Climate impacts on agriculture: Implications for crop production. *Agronomy Journal*, 103(2), 351–370.

Instituto Nacional de Estadísticas y Censos. (2024). *Exportaciones del complejo oleaginoso argentino durante 2024.* INDEC.

Instituto Nacional de Semillas. (2023). *Informe de soja 2023/2024.*
https://www.argentina.gob.ar/sites/default/files/sisa_soja_23_24_final_web.pdf

INTA. (2023). *Eficiencia y rendimiento en soja y maíz en la región pampeana: resultados experimentales.* Estación Experimental Pergamino, Instituto Nacional de Tecnología Agropecuaria.

INTA. (2024). *Evaluación de cultivares y prácticas de manejo en soja: Informe anual.* Instituto Nacional de Tecnología Agropecuaria.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.

- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: Wiley.
- Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17(6), 441–458.
- Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11), 1443–1452.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley: University of California Press.
- McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56). Austin, TX: SciPy.
- Ministerio de Agricultura, Ganadería y Pesca de la Nación. (2023). *Estimaciones agrícolas: Campaña 2022/2023*. <https://datoestimaciones.magyp.gob.ar/>
- Morla, M. (2022). *Calidad de semilla y germoplasma: su impacto en la productividad agrícola argentina*. Buenos Aires: Editorial Facultad de Agronomía (UBA).
- Pérez, J., Martínez, R., & López, A. (2020). Elasticidad de los factores productivos en sistemas agrícolas mecanizados. *Revista Latinoamericana de Economía Agraria*, 27(1), 55–72.
- Perren, J. (2008). *Determinación de costos y análisis de viabilidad económica en cultivos extensivos*. Universidad Nacional del Litoral.
- ReporteAsia. (2024, junio 19). *Agricultura sostenible en Argentina: Iniciativas que impulsan el futuro del agro*. <https://reporteasia.com/economia/desarrollo-sostenible/2024/06/19/agricultura-sostenible-argentina/>

- Rodríguez Sperat, D., Brugiafreddo, C., & Raña, P. (2017). Eficiencia técnica en productores agropecuarios argentinos: un análisis empírico. *Revista de Economía y Sociología Rural*, 55(2), 285–305.
- Rodríguez Sperat, R., Brugiafreddo, M. P., & Raña, E. (2017). Eficiencia técnica en la agricultura familiar: Análisis envolvente de datos (DEA) versus aproximación de fronteras estocásticas (SFA). *Nova Scientia*, 9(1), 18–39.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Selorm, A., Mensah, J., & Boateng, K. (2023). Determinants of technical efficiency in soybean production: Evidence from smallholder farmers. *Agricultural Economics Review*, 24(1), 45–62.
- Selorm, A., Sarpong, D. B. S., Egyir, I. S., Mensah Bonsu, A., & An, H. (2023). Does contract farming affect technical efficiency? Evidence from soybean farmers in Northern Ghana. *Agricultural and Food Economics*, 11(1), Article 9.
- Sharma, R., & Baliyan, S. P. (2022). Impact of precision agriculture adoption on soybean productivity. *Journal of Agribusiness and Rural Development*, 63(2), 101–115.
- Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136(1), 31–64.
- Simar, L., & Wilson, P. W. (2011). Two-stage DEA: Caveat emptor. *Journal of Productivity Analysis*, 36(2), 205–218.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36.

Van Rossum, G. (2009). Python programming language. In J. K. Binstock (Ed.), *USENIX Annual Technical Conference* (pp. 1–36). Berkeley, CA: USENIX Association.

Wang, H., & Shi, X. (2020). Technical efficiency and productivity growth in agriculture: Evidence from developing countries. *World Development*, 135, 105089.

Wang, Y., & Shi, X. (2020). Analysis on efficiency and influencing factors of new soybean producers. *Agronomy*, 10(4), 568.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Anexo

Anexo I. Preguntas realizadas en la encuesta a productores agropecuarios

1. ¿Qué tamaño tiene su explotación agrícola?
 - Menos de 100 hectáreas
 - 100–500 hectáreas
 - 500–1.000 hectáreas
 - Más de 1.000 hectáreas
2. ¿En qué localidad se ubica el principal establecimiento productivo?
 - (Respuesta abierta)
3. ¿Cuántos años lleva trabajando en la producción agrícola?
 - Menos de 5 años
 - Entre 5 y 10 años
 - Entre 11 y 20 años
 - Más de 20 años
4. ¿Qué modalidad jurídica implementa en la producción?
 - Unipersonal
 - Societaria (SA, SRL, etc.)
 - Fideicomiso
 - Cooperativa
 - Otro
5. ¿Es propietario de los inmuebles utilizados en la producción agropecuaria?
 - Sí, soy propietario
 - No, arriendo
 - Otro
6. En la última campaña, ¿ha realizado una rotación de cultivos con soja?
 - Sí
 - No
7. ¿Con qué frecuencia realiza rotaciones de cultivos?
 - No realizo rotaciones
 - Cada año
 - Cada 2 años

- Cada 3 años o más
- 8. ¿Qué método de siembra utiliza?
 - Siembra directa
 - Siembra convencional
 - Otro
- 9. ¿Contrata servicios externos para siembra y cosecha?
 - Sí, cosecha
 - Sí, siembra
 - Sí, ambas
 - No, utilizo maquinaria propia
- 10. ¿Utiliza tecnologías de Agricultura de Precisión (AP)?
 - Sí
 - No
- 11. Si respondió “Sí” a la anterior pregunta: ¿Qué tecnologías emplea?
 - GPS
 - Sensores de suelo
 - Drones
 - Software de gestión agrícola
 - Otro
- 12. ¿Qué insumos fueron utilizados en la última campaña?
 - Semillas modificadas genéticamente
 - Semillas propias
 - Otro tipo de semillas
 - Herbicidas
 - Insecticidas
 - Fungicidas
 - Otro
- 13. ¿Cuál fue el rendimiento promedio de soja en la última campaña?
 - (Respuesta abierta, kg/ha)
- 14. Variación del rendimiento en comparación con años anteriores
 - Ha aumentado
 - Se ha mantenido igual
 - Ha disminuido

15. ¿Qué factores afectaron el rendimiento?

- Clima
- Calidad del suelo
- Manejo de insumos
- Tecnología
- Asesoramiento técnico
- Plagas y malezas resistentes
- Otro

16. ¿Cómo describiría las condiciones climáticas de la última campaña?

- Muy favorables
- Favorables
- Neutras
- Desfavorables
- Muy desfavorables

17. ¿Cuál fue el costo estimado de producción por hectárea?

- (Respuesta abierta, kg/ha)

18. ¿Qué insumo impactó más en su costo por hectárea?

- Fertilizante
- Otro

19. ¿Ha utilizado financiamiento externo para la producción en el último año?

- Sí
- No

20. Si respondió “Sí”: ¿Qué tipo de financiamiento utilizó?

- Préstamos bancarios
- Tarjeta de crédito agropecuaria
- Sociedad de Garantía Recíproca
- Canje de granos
- Cooperativas o asociaciones
- Comercial
- Otro

21. ¿Cómo comercializa su producción?

- Venta directa sin intermediarios
- A través de un acopiador o cooperativa

- Venta para exportación
- Venta en mercado futuro
- Otro

22. ¿Qué prácticas agrícolas sostenibles implementa?

- Conservación del suelo
- Reducción del uso de químicos
- Uso eficiente del agua
- Rotación de cultivos
- Otro

23. ¿Cuáles cree que fueron los principales desafíos en la producción de soja?

- Cambio climático
- Acceso a insumos
- Fluctuaciones de precios
- Plagas y enfermedades
- Acceso a financiamiento
- Otro

24. ¿Está dispuesto a adoptar nuevas tecnologías si mejoran la eficiencia?

- Sí
- No

25. ¿Qué tecnologías le interesan?

- Agricultura de Precisión
- Biotecnología
- Sistemas de riego avanzados
- Sensores y drones
- Internet de las cosas
- Otro

26. ¿Con qué frecuencia revisa y ajusta sus estrategias de producción?

- Antes de cada temporada
- Trimestralmente
- Solo en caso de emergencias

27. ¿Cuáles son los tres factores más importantes para lograr la máxima eficiencia en su producción?

- (Respuesta abierta)

28. ¿Recibe asesoramiento técnico de especialistas?

- Sí
- No

29. ¿Colabora con otros productores o instituciones?

- Sí
- No

30. ¿Participa en jornadas técnicas de semilleros, ensayos y exposiciones para mejorar su rendimiento?

- Sí
- No

31. ¿Qué porcentaje de su inversión anual se destina a tecnología (maquinaria, software, etc.)?

- (Respuesta abierta)

32. ¿Usa tecnología ENLIST para aumentar la eficiencia en el control de malezas?

- Sí
- No

Anexo II. Metodología y diagnósticos del clustering

Figura 2.1. Dendrograma (clustering jerárquico – Ward). Agrupamiento jerárquico con distancia euclídea sobre variables estandarizadas y criterio de Ward. *Fuente: elaboración propia en Python en base a datos de encuesta.*

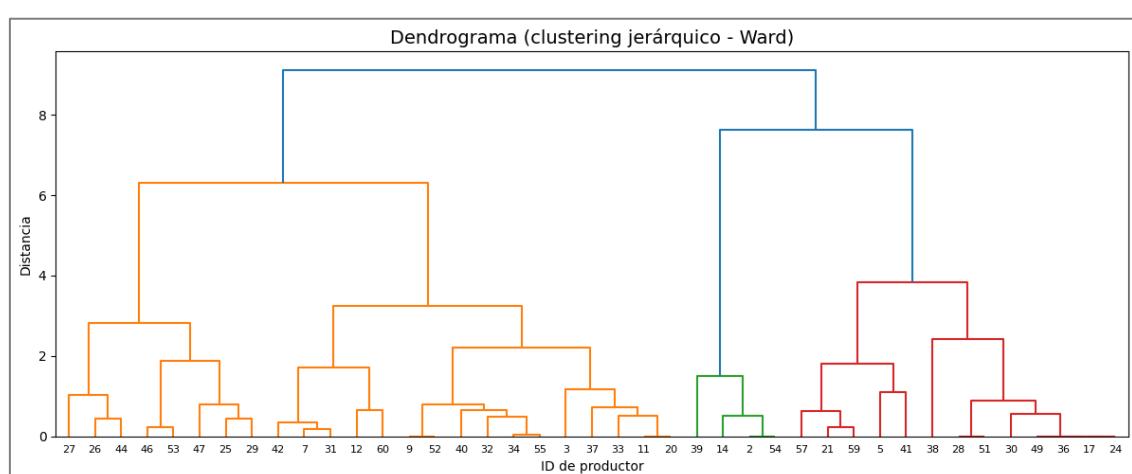


Figura 2.2a. Método del codo (inerzia) para la elección de K. Inercia total de K-means para $k \in [2,7]$. *Fuente: elaboración propia en Python en base a datos de encuesta.*

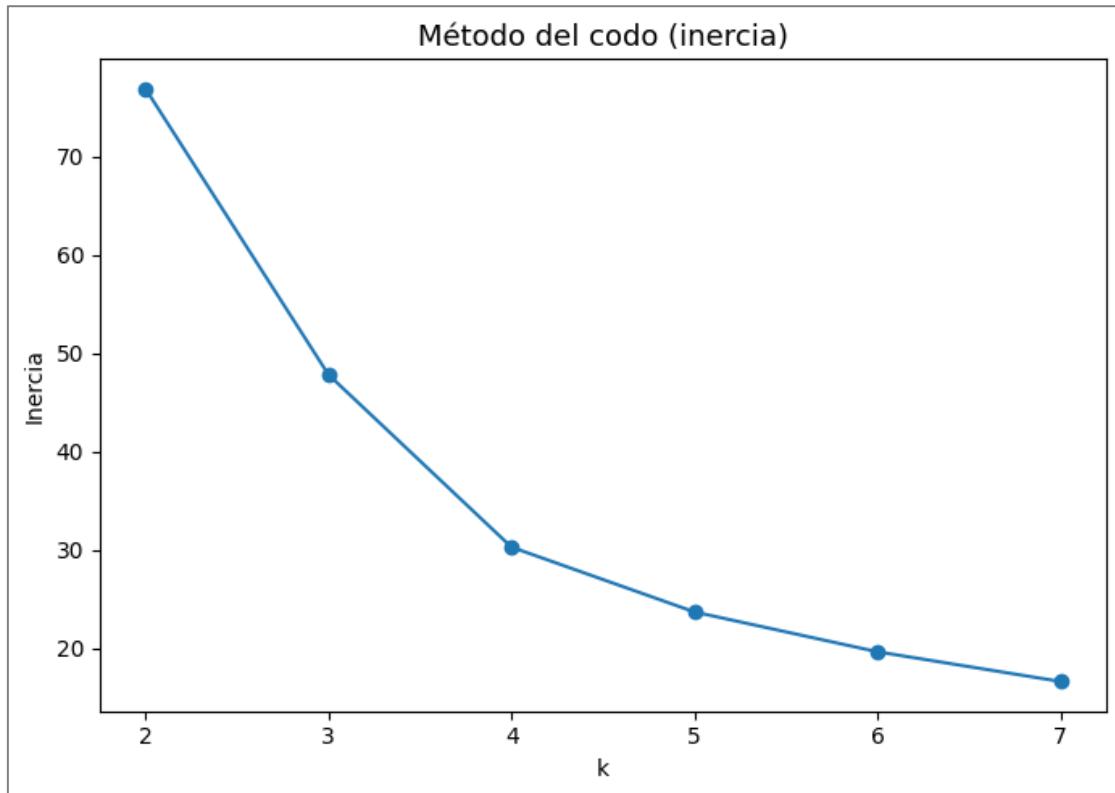
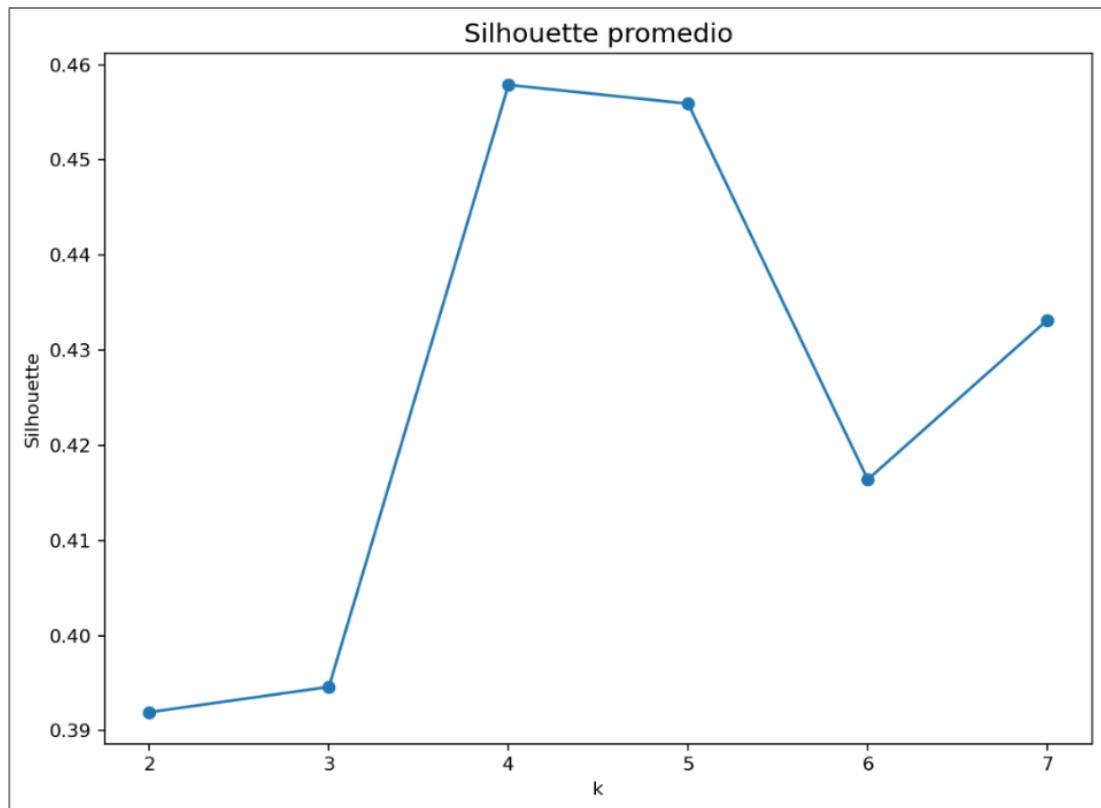


Figura 2.2b. Índice silhouette promedio por K. Índice silhouette promedio por k; máximo en $k=4$ ($\approx 0,458$). *Fuente: elaboración propia en Python en base a datos de encuesta.*



Nota: las variables seleccionadas para el clustering fueron superficie (ha), años de experiencia, % de inversión tecnológica. Todas las variables se estandarizaron (z-score) antes de calcular distancias. La visualización principal en el cuerpo (Figura 4) corresponde a PCA 2D coloreado por K-means ($k=3$) para facilitar lectura. La elección de k se guía por codo + silhouette; se reporta $k=3$ por parsimonia, sin cambios cualitativos frente a $k=4$.

Anexo III. Tablas y archivos de soporte del Clustering

Tabla 3.1. Estadísticas por conglomerado (K-means, $k=3$). Medias, medianas, desvíos y tamaño muestral para: *superficie (ha)*, *años de experiencia*, *porcentaje de inversión tecnológica*.

Cluster	Superficie				Experiencia				Porcentaje inversión tecnología			
	Media	Mediana	Desvio	Recuento	Media	Mediana	Desvio	Recuento	Media	Mediana	Desvio	Recuento
0	488.24	300.0	286.43	17	22.06	25.0	4.70	17	0.17	0.15	0.14	17
1	1250.00	1250.0	0.00	8	19.00	20.0	6.93	8	0.57	0.55	0.25	8
2	506.25	300.0	410.94	16	4.50	3.0	2.00	16	0.13	0.12	0.11	16

Nota: Los valores son medias por grupo. El “porcentaje de inversión tecnológica” está acotado entre 0 y 1. Tamaños muestrales: n_cluster0 = 17, n_cluster1 = 8, n_cluster2 = 16. Las variables se estandarizaron solo para hacer el clustering; la tabla muestra las medias en unidades originales. La elección de k se basó en el método del codo y el índice silhouette (máximo en k = 4; silhouette \approx 0.458). Se reporta k = 3 por parsimonia, sin cambios cualitativos en la interpretación.

Tabla 3.2. Asignación de productores a conglomerados (ID a cluster). Listado de IDs y su cluster correspondiente.

	CLUSTER 0	CLUSTER 1	CLUSTER 2
ID	7	2	3
12	5	9	
17	14		11
21	26		20
24	27		25
28	39		29
30	44		32
31	54		33
36	-		34
38	-		37
41	-		40
42	-		46
49	-		47
51	-		52
57	-		53
59	-		55
60	-		-

Anexo IV. Resultados del Análisis Envolvente de Datos (DEA)

Tabla 4.1. Resumen de la eficiencia técnica. Count, media, mediana, mínimo, máximo, % eficientes ($\theta=1$ en la muestra).

count	46
mean	0,6633
median	0,6434
min	0,2414
max	1
pct_eficientes	43,4783

Tabla 4.2. Unidades eficientes ($\theta=1$) y frecuencia como referentes. Listado de IDs eficientes y columna “veces_referencia”.

	veces_referencia
30	26
31	15
32	13
43	11
37	9
27	8
44	4
9	3
3	2
42	2

Tabla 4.3. Resultados completos por productor. ID, θ , referentes y (si aplica) metas por insumo.

ID	superficie	fertilizante	herbicida	insecticida	fungicida	siembra	cosecha	experiencia	soja	eficiencia_dea	peers_ids
1	1250	0	1	0	0	0	0	25	2800	0,340	[30, 31]
2	300	0	1	1	1	1	1	7	2800	0,636	[30, 32, 43]
3	300	0	0	0	0	1	1	7	3000	1,000	[3]
4	1250	1	1	1	1	1	1	25	4000	0,450	[27, 31, 42, 43, 44]
5	1250	1	1	1	1	0	0	25	3200	0,299	[30, 31, 37]
6	300	1	1	1	1	0	0	15	3200	0,435	[30, 31, 32, 37]
7	1250	0	1	1	1	1	1	25	1700	0,241	[30, 43]
8	300	0	1	1	1	1	0	15	3200	0,435	[30, 31, 32, 37]
9	300	0	1	1	1	0	1	3	3800	1,000	[9]
10	750	0	1	1	1	1	1	15	3500	0,399	[27, 30, 31, 43, 44]
11	1250	0	1	1	1	0	1	25	2900	0,241	[30, 32]
12	300	0	1	1	1	1	1	25	2500	0,250	[30, 32]
13	1250	1	1	1	0	1	1	25	3200	0,274	[3, 30, 31, 37]
14	300	0	1	0	0	0	0	25	3500	0,760	[30, 31]
15	1250	0	1	1	1	1	1	3	2000	1,000	[43]
16	750	0	1	1	0	0	1	7	3000	0,650	[30, 37]
17	300	0	1	1	1	1	1	3	4000	1,000	[27, 43]
18	750	0	1	1	0	0	1	25	2700	0,253	[30, 37]
19	75	1	1	1	0	0	0	7	3700	1,000	[19]
20	300	0	1	1	1	0	1	25	4500	1,000	[20]
21	300	0	1	1	1	0	0	25	3000	0,280	[30, 31, 32]
22	1250	0	1	1	0	1	1	3	3000	1,000	[43]
23	300	0	1	0	0	0	1	25	3500	0,760	[30, 31]
24	1250	0	1	1	1	0	0	3	3000	1,000	[32]
25	300	1	1	1	1	1	1	15	3600	0,541	[19, 27, 30, 31]
26	300	0	1	1	1	1	1	3	4200	1,000	[27]
27	300	0	1	1	1	1	1	3	4500	1,000	[27]

28	300	0	1	1	1	0	1	15	2300	0,368	[30, 32]
29	300	0	1	1	1	0	1	25	3000	0,280	[30, 31, 32]
30	75	0	0	0	0	0	0	7	2700	1,000	[30]
31	300	0	0	0	0	0	0	25	3900	1,000	[31]
32	75	0	1	1	1	0	0	3	3600	1,000	[32]
33	1250	0	1	1	1	1	1	25	3350	0,281	[27, 30, 31, 43, 44]
34	300	0	1	0	0	1	1	15	2000	0,410	[30, 43]
35	1250	1	1	1	1	0	0	15	4500	1,000	[35]
36	1250	0	1	1	1	0	0	25	2800	0,241	[30, 37]
37	750	0	1	1	0	0	0	3	4000	1,000	[37]
38	1250	0	1	1	1	0	1	7	4500	1,000	[38]
39	1250	0	1	1	1	0	1	15	2100	0,368	[9, 30]
40	300	0	1	1	1	1	1	7	3500	0,636	[27, 30, 32, 43]
41	750	0	1	1	1	0	1	3	3200	1,000	[9]
42	1250	0	1	1	1	0	1	25	5500	1,000	[42]
43	300	0	1	0	0	1	1	3	3600	1,000	[43]
44	1250	0	0	1	1	0	0	7	4000	1,000	[44]
45	750	0	1	1	1	0	0	25	3200	0,313	[30, 31, 32, 37]
46	750	0	1	1	1	0	0	15	3000	0,368	[30, 32]

Tabla 4.5. Eficiencia por tipologías (DEA × clusters) y contraste. Media, mediana, desvío, min, max, % eficientes por cluster; **Kruskal–Wallis (H, p-value)** y comparaciones pareadas (p-ajustado Bonferroni).

cluster	count	mean	median	std	min	max	pct_eficientes
0	12	0,698	1	0,374	0,241	1	58,333
1	7	0,723	0,76	0,301	0,299	1	42,857
2	12	0,647	0,589	0,33	0,241	1	41,667

pair	p_raw	p_adj_bonf
0 vs 1	0,890911139	1
0 vs 2	0,975371479	1
1 vs 2	0,597933357	1

ID	eficiencia_dea	cluster
2	0,636	1
3	1,000	2
5	0,299	1
7	0,241	0
9	1,000	2
11	0,241	2
12	0,250	0
14	0,760	1
17	1,000	0
20	1,000	2
21	0,280	0
24	1,000	0
25	0,541	2
26	1,000	1
27	1,000	1
28	0,368	0
29	0,280	2
30	1,000	0
31	1,000	0
32	1,000	2
33	0,281	2
34	0,410	2
36	0,241	0
37	1,000	2
38	1,000	0
39	0,368	1
40	0,636	2

41	1,000	0
42	1,000	0
44	1,000	1
46	0,368	2

Anexo V. Resultados del Modelo de Tobit

Variable	sociedad	propietario	financiamiento	rotacion	semilla
N	46	46	46	46	46
Coef(propietario)		0,09			
SE(propietario)		0,17			
z(propietario)		0,55			
p(propietario)		0,58			
Coef(financiamiento)			-0,31		
SE(financiamiento)			0,27		
z(financiamiento)			-1,13		
p(financiamiento)			0,26		
Coef(rotacion)				0,17	
SE(rotacion)				0,71	
z(rotacion)				0,24	
p(rotacion)				0,81	
Coef(semilla)					0,45
SE(semilla)					0,20
z(semilla)					2,21
p(semilla)					0,03
Uncens.	26	26	26	26	26
Left-cens.	0	0	0	0	0
Right-cens.	20	20	20	20	20
Constante (latente en v=0)	0,91	0,77	0,99	0,65	0,62
SE(Constante)	0,22	0,13	0,14	0,76	0,15
z(Constante)	4,19	5,91	6,88	0,86	4,24
p(Constante)	0,00	0,00	0,00	0,39	0,00
Coef(sociedad)	-0,15				
SE(sociedad)	0,18				
z(sociedad)	-0,79				
p(sociedad)	0,43				
sigma	0,52	0,52	0,50	0,52	0,47

Anexo VI. Script de Python ejecutado en Google Colab

```
# 0. Pasos previos al análisis

# 1. Librerías y configuraciones iniciales
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 2. Subida local del archivo
from google.colab import files
uploaded = files.upload()

# 3. Lectura del archivo
df = pd.read_excel("resultados_encuesta.xlsx")

----

# 1. Análisis descriptivo general

# 1. Vista general del DataFrame
df.info()
display(df.describe().T)

# 2. Sepación de variables numéricas y categóricas
id_series = df['ID'].copy()

# Solo para análisis de variables numéricas, excluimos ID
df_numericas = df.select_dtypes(include=['float64', 'int64']).drop(columns=['ID'],
errors='ignore')
df_categoricas = df.select_dtypes(exclude=['float64', 'int64'])

# 3. Valores únicos por variable categórica
valores_unicos_cat = df_categoricas.nunique().sort_values(ascending=False)
display(valores_unicos_cat)

# 4. Estadísticas descriptivas detalladas
desc_numericas = df_numericas.describe().T
desc_numericas["missing_pct"] = df_numericas.isnull().mean() * 100
display(desc_numericas)

# 5. Porcentaje de valores faltantes por variable (numéricas + categóricas)
faltantes = df.drop(columns=['ID'],
errors='ignore').isnull().mean().sort_values(ascending=False) * 100

# 6. Visualización de variables con faltantes
```

```

faltantes_filtrados = faltantes[faltantes > 0]

plt.figure(figsize=(10, 6))
faltantes_filtrados.plot(kind='barh', color='salmon')
plt.title("Porcentaje de valores faltantes por variable")
plt.xlabel("% de valores faltantes")
plt.ylabel("Variable")
plt.gca().invert_yaxis()
plt.tight_layout()
plt.show()

# 7. Variables seleccionadas para análisis gráfico
variables_interes = ['rendimiento_soja_kg_ha', 'costo_producción',
'años_experiencia', 'superficie_ha', 'porcentaje_inversión_tecnología']

# 8. Histogramas
df[variables_interes].hist(bins=15, figsize=(12, 8), color='skyblue',
edgecolor='black')
plt.suptitle("Distribución de variables clave (inputs y output DEA)", fontsize=14)
plt.tight_layout()
plt.show()

Nota: 'años_experiencia' y 'superficie_ha' fueron asignadas por criterios promedio
definidos en base a segmentación de respuestas (no reportadas directamente).

# 9. Análisis de conglomerados

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt
from matplotlib import patheffects as pe
import seaborn as sns
import numpy as np
import pandas as pd

try:
    from adjustText import adjust_text
    _ADJUST_OK = True
except Exception:
    _ADJUST_OK = False

# 9.1 Variables a usar
vars_cluster = ["superficie_ha", "años_experiencia",
"porcentaje_inversión_tecnología"]

```

```

req_cols = ["ID"] + vars_cluster
faltan = [c for c in req_cols if c not in df.columns]
if faltan:
    raise ValueError(f"Faltan columnas en df: {faltan}\nColumnas disponibles: {list(df.columns)}")

df_cluster = df[req_cols].copy()

for c in vars_cluster:
    df_cluster[c] = pd.to_numeric(df_cluster[c], errors="coerce")
df_cluster = df_cluster.dropna(subset=vars_cluster).reset_index(drop=True)

# 9.2 Normalización
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_cluster[vars_cluster])

# 9.3 K-Means
k = 3
kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
df_cluster["cluster"] = kmeans.fit_predict(X_scaled)

# 9.4 PCA 2D
pca = PCA(n_components=2, random_state=42)
coords = pca.fit_transform(X_scaled)
df_cluster["PC1"] = coords[:, 0]
df_cluster["PC2"] = coords[:, 1]

fig, ax = plt.subplots(figsize=(16, 10), dpi=160)
sns.scatterplot(
    x="PC1", y="PC2", hue="cluster", data=df_cluster,
    palette="Set2", s=140, alpha=0.9, ax=ax
)

# Etiquetas ID con halo blanco
texts = []
for _, r in df_cluster.iterrows():
    t = ax.text(
        r.PC1 + 0.03, r.PC2 + 0.03, str(int(r.ID)),
        fontsize=12, weight="bold", ha="center", va="center",
        path_effects=[pe.withStroke(linewidth=3, foreground="white")]
    )
    texts.append(t)

# Evitar solapamientos
if _ADJUST_OK:
    adjust_text(texts, ax=ax, arrowprops=dict(arrowstyle="-", lw=0.6, alpha=0.6))

```

```

ax.set_title("Clusters de productores (PCA 2D)", fontsize=20)
ax.set_xlabel("Componente principal 1", fontsize=14)
ax.set_ylabel("Componente principal 2", fontsize=14)
ax.tick_params(labelsize=12)
ax.legend(title="Cluster", fontsize=12, title_fontsize=12, markerscale=1.4,
frameon=True)
ax.grid(alpha=0.2)
ax.margins(x=0.12, y=0.15)
fig.tight_layout()
plt.show()
fig.savefig("Fig4_Clusters_PCA2D.png", dpi=300, bbox_inches="tight")

# 9.5 Caracterización de clusters
caract_clusters = (
    df_cluster.groupby("cluster")[vars_cluster]
    .agg(["mean", "median", "std", "count"])
    .round(2)
)
display(caract_clusters)

# 9.6 Clustering jerárquico (Ward)
Z = linkage(X_scaled, method="ward")
fig_d, ax_d = plt.subplots(figsize=(16, 6), dpi=160)
dendrogram(
    Z,
    labels=df_cluster["ID"].astype(str).values,
    leaf_rotation=0,
    leaf_font_size=9,
    color_threshold=None,
    ax=ax_d
)
ax_d.set_title("Dendrograma (clustering jerárquico - Ward)", fontsize=16)
ax_d.set_xlabel("ID de productor", fontsize=12)
ax_d.set_ylabel("Distancia", fontsize=12)
fig_d.tight_layout()
plt.show()
fig_d.savefig("Anexo2_Fig2_1_Dendrograma_Ward.png", dpi=300, bbox_inches="tight")

# 9.7 Diagnóstico de k: codo y silhouette
Ks = range(2, 8)
inercias, silhouettes = [], []
for ki in Ks:
    km_i = KMeans(n_clusters=ki, random_state=42, n_init=10).fit(X_scaled)
    inercias.append(km_i.inertia_)
    silhouettes.append(silhouette_score(X_scaled, km_i.labels_))

fig_k, ax = plt.subplots(1, 2, figsize=(16, 6), dpi=160)

```

```

ax[0].plot(list(Ks), inercias, marker="o")
ax[0].set_title("Método del codo (inercia)", fontsize=14)
ax[0].set_xlabel("k"); ax[0].set_ylabel("Inercia")

ax[1].plot(list(Ks), silhouettes, marker="o")
ax[1].set_title("Silhouette promedio", fontsize=14)
ax[1].set_xlabel("k"); ax[1].set_ylabel("Silhouette")

fig_k.tight_layout()
plt.show()
fig_k.savefig("Anexo2_Fig2_2_Codo_Silhouette.png", dpi=300, bbox_inches="tight")

best_k_by_sil = list(Ks)[int(np.argmax(silhouettes))]
print(f"Mejor k por silhouette ≈ {best_k_by_sil}
(silhouette={max(silhouettes):.3f})")

# 9.8 Exportación
!pip -q install openpyxl
with pd.ExcelWriter("Anexo3_clusters_output.xlsx", engine="openpyxl") as writer:
    df_cluster[["ID", "cluster", "PC1", "PC2"]].sort_values(["cluster", "ID"]).to_excel(
        writer, sheet_name="asignacion", index=False
    )
    caract_clusters.to_excel(writer, sheet_name="estadisticas")

df_cluster[["ID", "cluster"]].sort_values(["cluster", "ID"]).to_excel(
    "clusters_productores.xlsx", index=False
)

print("Guardados:")
print(" - Fig4_Clusters_PCA2D.png")
print(" - Anexo2_Fig2_1_Dendrograma_Ward.png")
print(" - Anexo2_Fig2_2_Codo_Silhouette.png")
print(" - Anexo3_clusters_output.xlsx (hojas: asignacion, estadisticas)")
print(" - clusters_productores.xlsx (ID → cluster)")

--- 

# 2. Análisis envolvente de datos

Usando un nuevo dataset, donde se reduce la cantidad de columnas y la cantidad de observaciones enfocándose en Buenos Aires.

# 1. Subida local del archivo
from google.colab import files
import pandas as pd
import numpy as np

```

```

uploaded = files.upload()
fname = next(iter(uploaded))
df_nm = pd.read_excel(fname)
df_nm.head()

# 2. Elección de variables para DEA
OUTPUT = "soja"
INPUTS = ["superficie", "fertilizante", "herbicida", "insecticida",
          "fungicida", "siembra", "cosecha", "experiencia"]

# Exógenas (no entran al DEA)
EXOG = ["sociedad", "propietario", "financiamiento", "rotacion", "semilla"]

if "ID" not in df_nm.columns:
    import numpy as np
    df_nm.insert(0, "ID", np.arange(1, len(df_nm)+1, dtype=int))

# Dataset para DEA (ID como índice para "ocultarlo" en vistas)
df_dea = df_nm[["ID", OUTPUT] + INPUTS].copy().set_index("ID").sort_index()

# Dataset con exógenas (para Tobit/descri)
df_exog = df_nm[["ID"] + EXOG].copy().set_index("ID").sort_index()

print("Shape df_dea (DEA):", df_dea.shape)
display(df_dea.head())

# 3. DEA VRS (input-oriented)
from scipy.optimize import linprog
import numpy as np
import pandas as pd

OUTPUT = "soja"
INPUTS = ["superficie", "fertilizante", "herbicida", "insecticida",
          "fungicida", "siembra", "cosecha", "experiencia"]

X = df_dea[INPUTS].to_numpy(dtype=float)      # n x m
Y = df_dea[[OUTPUT]].to_numpy(dtype=float)     # n x 1
ids = df_dea.index.to_numpy()
n, m = X.shape
s = Y.shape[1]      # 1

def dea_input_vrs(X, Y, tol=1e-9):
    """
    DEA input-oriented VRS:
        min θ
        s.a.  X λ <= θ x₀
              Y λ >= y₀
    """

```

```

1' λ = 1
λ >= 0, θ >= 0
Devuelve: eficiencias (θ), lambdas, status
"""

n, m = X.shape
s = Y.shape[1]
eff = np.zeros(n)
lambdas_all = np.zeros((n, n))
status = []

for o in range(n):
    x0 = X[o, :]
    y0 = Y[o, :]

    # Variables: [λ_1..λ_n, θ]
    c = np.zeros(n + 1); c[-1] = 1.0 # objetivo: min θ

    # X λ <= θ x0
    A_ub = np.zeros((m, n + 1)); b_ub = np.zeros(m)
    for j in range(m):
        A_ub[j, :n] = X[:, j]
        A_ub[j, -1] = -x0[j]

    # Y λ >= y0 → -Y λ <= -y0
    A_ub_y = np.zeros((s, n + 1)); b_ub_y = -y0
    A_ub_y[:, :n] = -Y.T

    # VRS: sum λ = 1
    A_eq = np.zeros((1, n + 1)); b_eq = np.array([1.0])
    A_eq[0, :n] = 1.0

    bounds = [(0, None)] * (n + 1)

    A_ub_total = np.vstack([A_ub, A_ub_y])
    b_ub_total = np.concatenate([b_ub, b_ub_y])

    res = linprog(c, A_ub=A_ub_total, b_ub=b_ub_total,
                  A_eq=A_eq, b_eq=b_eq, bounds=bounds, method="highs")

    if res.success:
        theta = res.x[-1]
        lambdas = res.x[:n]
        # Limpieza numérica
        theta = max(theta, 0.0)
        lambdas[lambdas < tol] = 0.0

        eff[o] = theta

```

```

        lambdas_all[o, :] = lambdas
        status.append("ok")
    else:
        eff[o] = np.nan
        status.append(f"fail: {res.message}")

    return eff, lambdas_all, status

eff, lambdas, status = dea_input_vrs(X, Y)

# Guardar resultados en el DF
df_dea["eficiencia_dea"] = eff
df_dea["status"] = status

# Peers/benchmarks por DMU (IDs con lambda>0)
peers = []
for i in range(n):
    peer_ids = ids[lambdas[i] > 0].tolist()
    peers.append(peer_ids)
df_dea["peers_ids"] = peers

# Vista rápida
display(df_dea[INPUTS + ["eficiencia_dea", "peers_ids"]].head(30))

# 4. Exportación y visualización

import matplotlib.pyplot as plt

# Export principal (recupero ID como columna)
df_out = df_dea[INPUTS + [OUTPUT, "eficiencia_dea", "status",
"peers_ids"]].reset_index()
df_out.to_excel("resultados_eficiencia_dea.xlsx", index=False)

# Lambdas (benchmarks) con ID
df_lambdas = pd.DataFrame(lambdas, index=ids, columns=[f"lambda_{i+1}" for i in
range(len(ids))]).reset_index()
df_lambdas = df_lambdas.rename(columns={"index": "ID"})
with pd.ExcelWriter("resultados_eficiencia_dea_completo.xlsx") as w:
    df_out.to_excel(w, index=False, sheet_name="DEA")
    df_lambdas.to_excel(w, index=False, sheet_name="Lambdas")

print("Archivos exportados: resultados_eficiencia_dea.xlsx y
resultados_eficiencia_dea_completo.xlsx")

# Histogram + línea de eficiencia plena
plt.figure(figsize=(9,5))
plt.hist(df_dea['eficiencia_dea'].dropna(), bins=10, edgecolor='black')

```

```

plt.axvline(1.0, color='red', linestyle='--', linewidth=1, label='Eficiencia plena  
(1.0)')
plt.title('Distribución de la eficiencia técnica (DEA) – VRS / Input-oriented')
plt.xlabel('Eficiencia técnica'); plt.ylabel('Frecuencia')
plt.legend(); plt.grid(axis='y', alpha=0.3); plt.tight_layout()
plt.show()

# Boxplot
plt.figure(figsize=(7,5))
plt.boxplot(df_dea['eficiencia_dea'].dropna(), vert=False)
plt.title('Boxplot de eficiencia técnica (DEA)'); plt.xlabel('Eficiencia técnica')
plt.grid(axis='x', alpha=0.3); plt.tight_layout()
plt.show()

# 5. KPIs

import numpy as np
import pandas as pd

# 5.1 KPIs generales
kpi = df_dea["eficiencia_dea"].agg(["count", "mean", "median", "min", "max"]).to_frame("valor")
kpi.loc["pct_eficientes"] = (df_dea["eficiencia_dea"].eq(1).mean() * 100)
display(kpi.round(4))

ids_eficientes = df_dea.index[df_dea["eficiencia_dea"].round(6) == 1.0].tolist()
ids_ineficientes = df_dea.index[df_dea["eficiencia_dea"] < 1].tolist()
print("IDs eficientes (muestra):", ids_eficientes[:15])

targets_raw = df_dea[INPUTS].multiply(df_dea["eficiencia_dea"], axis=0)
targets = targets_raw.add_suffix("_target")

ahorro_abs = (df_dea[INPUTS] - targets_raw).add_suffix("_ahorro") # (n,8)
base_inputs = df_dea[INPUTS].replace(0, np.nan) # evitar
/0

ahorro_pct_vals = (ahorro_abs.values / base_inputs.values) * 100
ahorro_pct = pd.DataFrame(ahorro_pct_vals, index=df_dea.index,
                           columns=[f"{c}_ahorro_pct" for c in INPUTS])

# 5.2 Peers más citados
peer_counts = pd.Series([pid for sub in df_dea["peers_ids"] for pid in sub],
                        dtype="Int64").value_counts()
peers_top = peer_counts.head(10).to_frame("veces_referencia")
display(peers_top)

# 5.3 Export para documentación

```

```

res_full = pd.concat(
    [df_dea[INPUTS + [OUTPUT, "eficiencia_dea", "peers_ids"]], 
     targets, ahorro_abs, ahorro_pct],
    axis=1
).reset_index()

with pd.ExcelWriter("DEA_kpis_objetivos.xlsx") as w:
    kpis.round(4).to_excel(w, sheet_name="KPIs", header=False)
    res_full.round(4).to_excel(w, index=False, sheet_name="Resultados+Objetivos")
    peers_top.to_excel(w, sheet_name="Peers_top")

print("Exportado: DEA_kpis_objetivos.xlsx")

# 6. Cruce Clusters-DEA

# 6.1 Carga de clusters
import pandas as pd
import numpy as np

if "df_cluster" in globals() and {"ID","cluster"}.issubset(df_cluster.columns):
    cl = df_cluster[["ID","cluster"]].copy()
else:
    cl = pd.read_excel("clusters_productores.xlsx")[["ID","cluster"]].copy()

# tipos
cl["ID"] = pd.to_numeric(cl["ID"], errors="coerce").astype("Int64")
cl = cl.dropna(subset=[ "ID"]).copy()
cl["ID"] = cl["ID"].astype(int)

# 6.2 Merge con eficiencia DEA
dea_merge = df_dea.reset_index()[["ID","eficiencia_dea"]].merge(cl, on="ID",
how="inner")

# 6.3 Resumen por cluster
res_cluster =
dea_merge.groupby("cluster")["eficiencia_dea"].agg(["count","mean","median","std","mi
n","max"])
res_cluster["pct_eficientes"] =
dea_merge.groupby("cluster")["eficiencia_dea"].apply(lambda x: (x.eq(1).mean()*100))
display(res_cluster.round(3))

# 6.4 Gráfico: eficiencia por cluster
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8,5))
sns.boxplot(x="cluster", y="eficiencia_dea", data=dea_merge)

```

```

sns.stripplot(x="cluster", y="eficiencia_dea", data=dea_merge, alpha=0.55)
plt.title("Eficiencia DEA por cluster")
plt.tight_layout()
plt.show()

# 6.5 Tests no paramétricos (Kruskal y pares con Bonferroni)
from scipy.stats import kruskal, mannwhitneyu
import itertools

groups = [g["eficiencia_dea"].values for _, g in dea_merge.groupby("cluster")]
H, p_kw = kruskal(*groups)
print(f"Kruskal-Wallis: H={H:.3f}, p-value={p_kw:.4f}")

pairs = []
for a, b in itertools.combinations(sorted(dea_merge["cluster"].unique()), 2):
    x = dea_merge.loc[dea_merge["cluster"]==a, "eficiencia_dea"]
    y = dea_merge.loc[dea_merge["cluster"]==b, "eficiencia_dea"]
    stat, p = mannwhitneyu(x, y, alternative="two-sided")
    pairs.append({"pair": f"{a} vs {b}", "p_raw": p})

m = len(pairs)
pairs_df = pd.DataFrame(pairs)
if m > 0:
    pairs_df["p_adj_bonf"] = (pairs_df["p_raw"] * m).clip(upper=1.0)
display(pairs_df)

# 6.6 Export para el informe
with pd.ExcelWriter("DEA_vs_Clusters.xlsx") as w:
    dea_merge.to_excel(w, index=False, sheet_name="merge_ID_cluster_eff")
    res_cluster.round(3).to_excel(w, sheet_name="resumen_cluster")
    pairs_df.to_excel(w, index=False, sheet_name="tests_posthoc")

print("Exportado: DEA_vs_Clusters.xlsx")

# 3. Análisis de multicolinealidad

```

Antes de realizar el análisis, se va a evaluar la multicolinealidad. También se tendrá por criterio el 35% de valores nulos para eliminar o no una columna.

```

# 1. Carga de librerías necesarias
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
from IPython.display import display

```

```

# 2. Función para calcular coeficiente Phi
def phi_coefficient(x, y):
    cm = confusion_matrix(x, y, labels=[0, 1])
    if cm.shape != (2, 2):
        return np.nan
    a, b, c, d = cm[0, 0], cm[0, 1], cm[1, 0], cm[1, 1]
    denominator = np.sqrt((a + b)*(c + d)*(a + c)*(b + d))
    return (a*d - b*c) / denominator if denominator != 0 else np.nan

# 3. Crear copia de trabajo
df_modelo = df.copy()

# 4. Eliminar variables utilizadas en el DEA (excepto eficiencia_dea)
variables_dea = [
    'rendimiento_soja_kg_ha',
    'costo_producción',
    'porcentaje_inversión_tecnología',
    'años_experiencia',
    'superficie_ha'
]
df_modelo.drop(columns=[v for v in variables_dea if v in df_modelo.columns],
inplace=True)

# 5. Definir variables numéricas de interés (manteniendo ID fuera del análisis)
variables_numericas = [
    'frecuencia_rotación_cult', 'variación_rendimiento', 'clima_percibido',
    'frecuencia_rev_estrategias', 'eficiencia_dea'
]
variables_numericas = [v for v in variables_numericas if v in df_modelo.columns]

# 6. Definir variables categóricas tipo dummy (sin incluir variables nominales ni
numéricas)
excluir_dummies = ['ID', 'partido', 'provincia', 'reg_productiva']
variables_dummies = [
    v for v in df_modelo.columns
    if v not in variables_numericas and v not in excluir_dummies
]

# 7. Separar subconjuntos
df_num = df_modelo[variables_numericas].copy()
df_dum = df_modelo[variables_dummies].copy()

# 8. Matriz de correlación de Pearson
umbral_pearson = 0.6
corr_matrix = df_num.corr()

plt.figure(figsize=(10, 8))

```

```

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Matriz de Correlación de Pearson (Variables Numéricas)")
plt.tight_layout()
plt.show()

# 9. Identificación de pares con alta correlación
correlaciones_altas_pearson = []
for i in range(len(corr_matrix.columns)):
    for j in range(i + 1, len(corr_matrix.columns)):
        var1 = corr_matrix.columns[i]
        var2 = corr_matrix.columns[j]
        r = corr_matrix.iloc[i, j]
        if abs(r) >= umbral_pearson:
            correlaciones_altas_pearson.append((var1, var2, r))

# 10. Mostrar resultados
print("Pares de variables con correlación alta ( $|r| \geq 0.6$ ):")
for var1, var2, r in correlaciones_altas_pearson:
    print(f"{var1} - {var2}: r = {r:.2f}")

# 11. Correlación Phi (para dummies)
phi_matrix = pd.DataFrame(index=df_dum.columns, columns=df_dum.columns)

for col1 in df_dum.columns:
    for col2 in df_dum.columns:
        if col1 == col2:
            phi_matrix.loc[col1, col2] = 1.0
        else:
            try:
                phi = phi_coefficient(df_dum[col1], df_dum[col2])
                phi_matrix.loc[col1, col2] = phi
            except:
                phi_matrix.loc[col1, col2] = np.nan

phi_matrix = phi_matrix.astype(float)

# 12. Matriz de correlación PHI (visualización)
plt.figure(figsize=(12, 10))
sns.heatmap(phi_matrix, annot=False, cmap='YlGnBu', fmt=".2f", linewidths=0.5)
plt.title("Matriz de Correlación Phi (Variables Dummy)")
plt.tight_layout()
plt.show()

# 13. Identificar pares altamente correlacionados
threshold_phi = 0.6
correlaciones_altas_phi = []

```

```

for i in range(len(phi_matrix.columns)):
    for j in range(i + 1, len(phi_matrix.columns)):
        var1 = phi_matrix.columns[i]
        var2 = phi_matrix.columns[j]
        corr = phi_matrix.iloc[i, j]
        if pd.notnull(corr) and abs(corr) >= threshold_phi:
            correlaciones_altas_phi.append((var1, var2, corr))

# 14. Mostrar en formato tabla
tabla_phi = pd.DataFrame(correlaciones_altas_phi, columns=["Variable 1", "Variable 2", "Coeficiente Phi"])
tabla_phi = tabla_phi.sort_values(by="Coeficiente Phi",
                                   ascending=False).reset_index(drop=True)

print(f"\n\N{circled question mark} {len(tabla_phi)} pares de variables dummy con |Phi| ≥ {threshold_phi}:\n")
display(tabla_phi)

# 15. Variables candidatas a excluir (por multicolinealidad)

# Variables numéricas con |r| ≥ umbral_pearson
vars_excluir_pearson = set()
for v1, v2, r in correlaciones_altas_pearson:
    vars_excluir_pearson.add(v2)

# Variables dummy con |phi| ≥ threshold_phi
vars_excluir_phi = set()
for v1, v2, phi in correlaciones_altas_phi:
    vars_excluir_phi.add(v2)

# 16. Mostrar resultados
print("\n\N{circled question mark} Variables numéricas sugeridas para excluir (Pearson ≥ 0.6):")
print(sorted(vars_excluir_pearson))

print("\n\N{circled question mark} Variables dummy sugeridas para excluir (Phi ≥ 0.6):")
print(sorted(vars_excluir_phi))

# 17. Variables dummy con alta correlación Phi a eliminar
variables_phi_altas = ['desafío_cambio_clim', 'fact_asesoramiento', 'fact_insumos',
                       'insumos_fungicida']

# Eliminar del DataFrame
df_modelo.drop(columns=variables_phi_altas, inplace=True)

print("Variables eliminadas por alta correlación Phi:", variables_phi_altas)

# 18. Análisis Phi post depuración

```

```

from sklearn.metrics import confusion_matrix
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Función para calcular coeficiente Phi
def phi_coefficient(x, y):
    cm = confusion_matrix(x, y, labels=[0, 1])
    if cm.shape != (2, 2):
        return np.nan
    a, b, c, d = cm[0, 0], cm[0, 1], cm[1, 0], cm[1, 1]
    denominator = np.sqrt((a + b)*(c + d)*(a + c)*(b + d))
    return (a*d - b*c) / denominator if denominator != 0 else np.nan

# 19. Reconstruir DataFrame de dummies actualizado
variables_dummies_actualizadas = [v for v in df_modelo.columns if
df_modelo[v].nunique() == 2]
df_dum_actualizado = df_modelo[variables_dummies_actualizadas].copy()

# Calcular matriz Phi
phi_matrix = pd.DataFrame(index=variables_dummies_actualizadas,
columns=variables_dummies_actualizadas)

for col1 in df_dum_actualizado.columns:
    for col2 in df_dum_actualizado.columns:
        if col1 == col2:
            phi_matrix.loc[col1, col2] = 1.0
        else:
            try:
                phi = phi_coefficient(df_dum_actualizado[col1],
df_dum_actualizado[col2])
                phi_matrix.loc[col1, col2] = phi
            except:
                phi_matrix.loc[col1, col2] = np.nan

phi_matrix = phi_matrix.astype(float)

# 20. Visualizar matriz
plt.figure(figsize=(12, 10))
sns.heatmap(phi_matrix, cmap="YlGnBu", annot=False, fmt=".2f", linewidths=0.5)
plt.title("Matriz de Correlación Phi (Variables Dummy - Revisada)")
plt.tight_layout()
plt.show()

# 21. Listar pares con Phi ≥ 0.6 o ≤ -0.6

```

```

threshold_phi = 0.6
correlaciones_altas_phi = []

for i in range(len(phi_matrix.columns)):
    for j in range(i + 1, len(phi_matrix.columns)):
        var1 = phi_matrix.columns[i]
        var2 = phi_matrix.columns[j]
        corr = phi_matrix.iloc[i, j]
        if pd.notnull(corr) and abs(corr) >= threshold_phi:
            correlaciones_altas_phi.append((var1, var2, corr))

# 22. Mostrar resultados en tabla
tabla_phi = pd.DataFrame(correlaciones_altas_phi, columns=["Variable 1", "Variable 2", "Coeficiente Phi"])
tabla_phi = tabla_phi.sort_values(by="Coeficiente Phi",
                                   ascending=False).reset_index(drop=True)

from IPython.display import display
print(f"\n{len(tabla_phi)} pares de variables dummy con |Phi| ≥ {threshold_phi}:")
display(tabla_phi)

# 4. Eliminación de dummies con baja variabilidad

# 1. Carga de excel original
df = pd.read_excel("resultados_encuesta.xlsx")
df_modelo = df.copy()

# 2. Detectar columnas dummy (0 y 1)
columnas_dummy = [col for col in df_modelo.columns if sorted(df_modelo[col].dropna().unique()) == [0, 1]]

# 3. Calcular proporciones y cantidades
frecuencias_dummy = pd.DataFrame({
    'proporcion_1s': df_modelo[columnas_dummy].mean(),
    'proporcion_0s': 1 - df_modelo[columnas_dummy].mean(),
    'cantidad_1s': df_modelo[columnas_dummy].sum(),
    'cantidad_0s': df_modelo[columnas_dummy].apply(lambda x: (x == 0).sum())
})

# 4. Umbral de baja variabilidad
umbral = 0.85
dummies_extremos = frecuencias_dummy[
    (frecuencias_dummy['proporcion_1s'] > umbral) |
    (frecuencias_dummy['proporcion_0s'] > umbral)
]

```

```

# 5. Exportar la tabla de dummies a eliminar a Excel
dummies_extremos.to_excel("dummies_baja_variabilidad.xlsx")
print("☒ Archivo 'dummies_baja_variabilidad.xlsx' exportado correctamente.")

# 6. Eliminar dummies sesgados (solo si aún existen)
columnas_a_eliminar = [col for col in dummies_extremos.index if col in
df_modelo.columns]
df_modelo.drop(columns=columnas_a_eliminar, inplace=True)

# 7. Confirmación
print(f"\n☒ Variables eliminadas por baja variabilidad: {len(columnas_a_eliminar)}\ncolumnas.")
print("Columnas eliminadas:")
print(columnas_a_eliminar)

---


# 5. Modelo de Tobit

# 1. Tobit univariado con censura [0,1] y errores robustos

import os, glob, numpy as np, pandas as pd
from scipy.stats import norm
from scipy.optimize import minimize

def find_first_file(patterns):
    """Devuelve el primer archivo que existe que matchee cualquiera de los
patrones."""
    for pat in patterns:
        hits = sorted(glob.glob(pat))
        if hits:
            return hits[0]
    return None

def find_col(cols, candidates):
    for c in candidates:
        if c in cols: return c
    return None

def to_numeric_binary(s):
    """Intenta mapear a 0/1; si no puede, aplica códigos de categoría como float."""
    if pd.api.types.is_numeric_dtype(s):
        return pd.to_numeric(s, errors='coerce')
    s2 = s.astype(str).str.strip().str.lower()
    map_yes = {"si":1,"sí":1,"true":1,"1":1,"alta":1,"enlist":1,"sd":1,"siembra
directa":1}

```

```

map_no = {"no":0,"false":0,"0":0,"convencional":0}
out = s2.map(map_yes).fillna(s2.map(map_no))
if out.isna().any():
    # fallback: codificación ordinal solo para que el modelo corra
    out = s2.astype("category").cat.codes.astype(float)
return out.astype(float)

# 2. Base para Tobit (eficiencia + exógenas)

# Preparar base Tobit: eficiencia + exógenas
EXOG = ["sociedad","propietario","financiamiento","rotacion","semilla"]
LEFT, RIGHT = 0.0, 1.0 # censura doble en [0, 1]

def preparar_bases():
    # 1) eficiencia desde el Excel de DEA
    try:
        ef = df_dea[["eficiencia_dea"]].copy()
        ef.index.name = "ID"
    except:
        ef =
pd.read_excel("resultados_eficiencia_dea.xlsx")[["ID","eficiencia_dea"]].set_index("ID")

    # 2) exógenas desde el modelo_DEA (
    try:
        ex = df_exog[EXOG].copy()
        ex.index.name = "ID"
    except:
        ex = pd.read_excel("modelo_DEA.xlsx")
        if "ID" not in ex.columns:
            ex.insert(0, "ID", np.arange(1, len(ex)+1, dtype=int))
        ex = ex.set_index("ID")[EXOG]

    # 3) merge + limpieza
    tb = ef.merge(ex, left_index=True, right_index=True, how="inner").copy()
    tb["eficiencia_dea"] = pd.to_numeric(tb["eficiencia_dea"],
errors="coerce").clip(LEFT, RIGHT)
    return tb

df_tb = preparar_bases()
print("N total:", len(df_tb), "| masa en 1.0:",
(df_tb["eficiencia_dea"]==1).mean().round(3))
display(df_tb.head())

# 3. Saneamiento de exógenas

import numpy as np

```

```

import pandas as pd
from scipy.stats import norm
from scipy.optimize import minimize
import warnings
warnings.filterwarnings("ignore")

def to_numeric_binary(s):
    if pd.api.types.is_numeric_dtype(s):
        return pd.to_numeric(s, errors="coerce")
    s2 = s.astype(str).str.strip().str.lower()
    map_yes = {"si":1,"sí":1,"true":1,"1":1,"alta":1,"enlist":1,"sd":1,"siembra":1,"directa":1}
    map_no = {"no":0,"false":0,"0":0,"convencional":0}
    out = s2.map(map_yes).fillna(s2.map(map_no))
    # fallback: si aún quedan NaN, codifico categorías
    if out.isna().any():
        out = s2.astype("category").cat.codes.astype(float)
    return out.astype(float)

for v in EXOG:
    if v not in df_tb.columns:
        raise ValueError(f"Falta la columna '{v}' en df_tb.")
    df_tb[v] = to_numeric_binary(df_tb[v])

# 4. Tobit dos límites (0 y 1) por MLE
def _tobit_loglike(params, y, X, left=0.0, right=1.0):
    beta = params[:-1]
    sigma = np.exp(params[-1]) + 1e-12 # asegurar σ>0
    xb = X @ beta

    L = y <= left + 1e-12
    R = y >= right - 1e-12
    U = ~(L | R)

    ll = np.zeros_like(y, dtype=float)
    if U.any():
        z = (y[U] - xb[U]) / sigma
        ll[U] = np.log(norm.pdf(z)) - np.log(sigma)
    if L.any():
        zL = (left - xb[L]) / sigma
        ll[L] = np.log(np.maximum(norm.cdf(zL), 1e-300))
    if R.any():
        zR = (right - xb[R]) / sigma
        ll[R] = np.log(np.maximum(1 - norm.cdf(zR), 1e-300))
    return ll.sum()

def fit_tobit_double_censor(y, x, left=0.0, right=1.0):

```

```

X = np.column_stack([np.ones(len(x)), x]) # constante + x
init = np.zeros(X.shape[1] + 1)           # betas=0, log_sigma=0
init[-1] = np.log(0.5)                   # σ inicial

def nll(p): return -_tobit_loglike(p, y, X, left, right)
res = minimize(nll, init, method="L-BFGS-B")

params = res.x
try:
    Hinv = res.hess_inv.todense() if hasattr(res.hess_inv, "todense") else
np.array(res.hess_inv)
    se = np.sqrt(np.diag(Hinv))
except Exception:
    se = np.full_like(params, np.nan, dtype=float)

z = params / se
from scipy.stats import norm
p = 2 * (1 - norm.cdf(np.abs(z)))

L = int((y <= left + 1e-12).sum())
R = int((y >= right - 1e-12).sum())
U = len(y) - L - R
ll = -res.fun
k = len(params)
aic = 2*k - 2*ll
bic = k*np.log(len(y)) - 2*ll

return {
    "params": params, "se": se, "z": z, "p": p,
    "sigma": float(np.exp(params[-1])),
    "loglike": ll, "AIC": aic, "BIC": bic,
    "n": len(y), "uncensored": U, "left_cens": L, "right_cens": R,
    "converged": res.success, "message": res.message
}

# 5. Estimación univariada para cada variable en EXOG
rows, detalles = [], {}
for v in EXOG:
    sub = df_tb[["eficiencia_dea", v]].dropna().copy()
    y = pd.to_numeric(sub["eficiencia_dea"], errors="coerce").clip(LEFT,
RIGHT).values
    x = pd.to_numeric(sub[v], errors="coerce").values.reshape(-1,1)

    if len(sub) < 10:
        print(f"⚠️ Pocas observaciones para {v} (n={len(sub)}). Se estima igual, pero
interpretá con cautela.")
```

```

fit = fit_tobit_double_censor(y, x, left=LEFT, right=RIGHT)
b0, b1, logsig = fit["params"][0], fit["params"][1], fit["params"][2]
se0, se1, se_log = fit["se"][0], fit["se"][1], fit["se"][2]
z0, z1 = fit["z"][0], fit["z"][1]
p0, p1 = fit["p"][0], fit["p"][1]

rows.append({
    "Variable": v,
    "N": fit["n"], "Uncens.": fit["uncensored"],
    "Left-cens.": fit["left_cens"], "Right-cens.": fit["right_cens"],
    "Constante (latente en v=0)": b0, "SE(Constante)": se0,
    "z(Constante)": z0, "p(Constante)": p0,
    f"Coef({v})": b1, f"SE({v})": se1, f"z({v})": z1, f"p({v})": p1,
    "sigma": fit["sigma"], "logLik": fit["loglike"],
    "AIC": fit["AIC"], "BIC": fit["BIC"], "Convergió": fit["converged"]
})
detalles[v] = sub

res_tobit = pd.DataFrame(rows)

# Orden simple (primero sociedad, luego resto)
orden_vars = [v for v in
["sociedad", "propietario", "financiamiento", "rotacion", "semilla"] if v in
res_tobit["Variable"].values]
res_tobit["orden"] = res_tobit["Variable"].map({v:i for i,v in
enumerate(orden_vars)})
res_tobit = res_tobit.sort_values(["orden"]).drop(columns=["orden"])

pd.set_option("display.float_format", lambda x: f"{x:0.4f}")
display(res_tobit)

# 6. Exportar a Excel
out_xlsx = "tobit_univariado_5vars.xlsx"
with pd.ExcelWriter(out_xlsx) as w:
    res_tobit.to_excel(w, sheet_name="Resumen", index=False)
    for v, d in detalles.items():
        d.to_excel(w, sheet_name=f"data_{v[:25]}", index=False)

print(f"☑ Archivo exportado: {out_xlsx}")

# 7. Gráficos

import numpy as np
import matplotlib.pyplot as plt

# Coeficientes y errores estándar
vars_ = ["sociedad", "propietario", "financiamiento", "rotación", "semilla"]

```

```

coef  = np.array([-0.15, 0.09, -0.31, 0.17, 0.45])
se    = np.array([0.18, 0.17, 0.27, 0.71, 0.20])

z = 1.96 # IC 95%
lo = coef - z*se
hi = coef + z*se

y = np.arange(len(vars_))

plt.figure(figsize=(7.5, 4.8))
plt.axvline(0, ls="--", lw=1, color="gray")
plt.errorbar(coef, y, xerr=[coef-lo, hi-coef], fmt="o", capsize=3)
plt.yticks(y, vars_)
plt.xlabel("Coeficiente (escala latente)")
plt.title("Tobit univariado: estimaciones e IC95%")
plt.tight_layout()
plt.show()

```