



Colección Análisis y Reflexiones en torno al Sector Agropecuario

Construir tamaños de muestra a partir de datos del Tercer Censo Nacional Agropecuario (3^{er} CNA): una alternativa ante la ausencia de marco muestral

Yajaira Romero Barrera

Jacqueline Ávila Cárdenas

Pedro David Suárez Villota

Juan Carlos Fuentes Benítez

AGROSAVIA
EDITORIAL

Construir tamaños de muestra a partir de datos del Tercer Censo Nacional Agropecuario (3^{er} CNA): una alternativa ante la ausencia de marco muestral

Yajaira Romero Barrera
Jacqueline Ávila Cárdenas
Pedro David Suárez Villota
Juan Carlos Fuentes Benítez

AGROSAVIA
EDITORIAL

Mosquera, Colombia, 2021

Construir tamaños de muestra a partir de datos del Tercer Censo Nacional Agropecuario (3er CNA): una alternativa ante la ausencia de marco muestral / Yajaira Romero Barrera [y otros tres] – Mosquera (Colombia) : AGROSAVIA, 2021.

82 páginas (Colección Análisis y Reflexiones en torno al Sector Agropecuario)

Incluye ilustraciones, gráficos, tablas

ISBN E-book: 978-958-740-468-5

1. Muestreo estadístico 2. Variación estadística 3. Planificación del trabajo 4. Análisis de datos 5. Encuestas. I. Romero Barrera, Yajaira II. Jacqueline Ávila Cárdenas III. Pedro David Suárez Villota IV. Juan Carlos Fuentes Benítez.

Palabras clave normalizadas según Tesauro Multilingüe de Agricultura Agrovoc

Catalogación en la publicación – Biblioteca Agropecuaria de Colombia

Corporación Colombiana de Investigación Agropecuaria – AGROSAVIA

Centro de Investigación Tibaitatá. Kilómetro 14 vía Mosquera-Bogotá, Mosquera, Cundinamarca. Código postal 250047, Colombia.

Esta publicación es resultado del proyecto de la Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA "Caracterización de la oferta y demanda de semilla de calidad de las especies de interés".

Colección Análisis y Reflexiones en torno al Sector Agropecuario

Fecha de recepción: 25 de noviembre de 2019

Fecha de evaluación: 16 de abril de 2020

Fecha de aceptación: 19 de enero de 2021

Publicado: junio de 2021

Preparación editorial

Editorial AGROSAVIA

editorial@agrosavia.co

Editora: Liliana Gaona García

Corrección de estilo: Andrés Castillo Brieva

Diagramación: María Paula Berón Ramírez

DOI:

<https://doi.org/10.21930/agrosavia.analisis.7404685>

Citación sugerida: Romero Barrera, Y., Ávila Cárdenas, J., Suárez Villota, P. D., & Fuentes Benítez, J. C. (2020). *Construir tamaños de muestra a partir de datos del Tercer Censo Nacional Agropecuario (3er CNA): una alternativa ante la ausencia de marco muestral*. Corporación Colombiana de Investigación Agropecuaria-AGROSAVIA. <https://doi.org/10.21930/agrosavia.analisis.7404685>

Cláusula de responsabilidad: AGROSAVIA no es responsable de las opiniones e información recogidas en el presente texto. Los autores asumen de manera exclusiva y plena toda responsabilidad sobre su contenido, ya sea este propio o de terceros, y declaran, en este último supuesto, que cuentan con la debida autorización de terceros para su publicación; igualmente, declaran que no existe conflicto de interés alguno en relación con los resultados de la investigación propiedad de tales terceros. En consecuencia, los autores serán responsables civil, administrativa o penalmente, frente a cualquier reclamo o demanda por parte de terceros relativa a los derechos de autor u otros derechos que se hubieran vulnerado como resultado de su contribución.

Línea de atención al cliente: 018000121515

atencionalcliente@agrosavia.co

<http://www.agrosavia.co>



https://co.creativecommons.org/?page_id=13

Los autores	11
Glosario de símbolos matemáticos	15
Presentación	17
Introducción	21

Capítulo I

Marcos de referencia	25
Marco teórico	25
Breve introducción al muestreo	26
Marco conceptual	30

Capítulo II

Diseño estadístico	33
Marco estadístico	33
Definición de variables	33
Universo	34
Población objetivo	34

Capítulo III

Diseño muestral	37
Tipo de muestreo	37
Estimación de la media y el total poblacional	38
Definición del tamaño de la muestra	39
Afijación igual	40
Afijación proporcional	40

Afijación óptima de costo	41
Afijación óptima	42
Algunas consideraciones sobre la planeación de estudios por medio de aplicación de encuestas	43

Capítulo IV

Proceso metodológico desarrollado para el sector agropecuario	49
Preparación y procesamiento de bases de datos	49
Elaboración de estratos mediante la variable Area_Sembrada_Sum	52
Selección de una proporción de la población para estimar parámetros por estratos	53
Determinar varianzas, desviación estándar y medias por estratos	53

Capítulo V

Aplicación de la metodología desarrollada para el sector agropecuario	57
Criterios para la priorización de zonas de estudio	59
Cálculo para determinar el tamaño de muestra y estimar el promedio de la demanda de semilla por unidad de área	61
Estimación de número de UPA y área sembrada en otros cultivos	65
Sobre el trabajo de campo	66

Referencias	69
--------------------	----

Anexos	75
---------------	----

Anexo 1. Diseño del registro y descripción de las variables	75
Anexo 2. Código en SAS 9.4	76

Figura 1	Esquema general para la decisión del método de muestreo	26
Figura 2	Métodos de extracción de elementos de una muestra	29
Figura 3	Tipos de afijación para calcular el tamaño de la muestra	40
Figura 4	Proceso para el diseño, ejecución y análisis de datos de un estudio mediante encuesta	44
Figura 5	Códigos empleados en los filtros, en la preparación y procesamiento del archivo S06A(Cultivos)	61
Figura 6	Proceso para determinar los parámetros para el cálculo del tamaño de la muestra.	63

Tabla 1	Predominancia étnica	51
Tabla 2	Criterios de priorización de zonas para validar la metodología	59
Tabla 3	Área mínima de referencia	62
Tabla 4	Población por especie y municipio	62
Tabla 5	Distribución de la muestra a nivel UPA para tres sistemas productivos	64
Tabla 6	Aplicación de la metodología en otras especies	65





Los autores

Yajaira Romero Barrera

Correo: yromero@agrosavia.co

Orcid: <https://orcid.org/0000-0002-6512-7307>

Ingeniera de producción biotecnológica de la Universidad Francisco de Paula Santander y máster en Estadística Aplicada de la Universidad del Norte. Trabajó durante seis años en el Centro de Investigación Caribia de la Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA, donde participó en proyectos de investigación en hortalizas, frutales y métodos de muestreo de poblaciones biológicas. Se encuentra vinculada a la Dirección de Investigación y Desarrollo desde 2014 y es miembro del nodo central del Equipo Extendido de Estadística; además, participa en proyectos pertenecientes a diferentes redes en los componentes de muestreo de poblaciones biológicas y estudios socioeconómicos, así como en la planificación, ejecución y análisis de experimentos. También se desempeña como profesional de apoyo de la Red de Innovación de Hortalizas y plantas aromáticas de AGROSAVIA.

Jacqueline Ávila Cárdenas

Correo: javilac@agrosavia.co

Orcid: <https://orcid.org/0000-0001-8610-9194>

Profesional en mercadeo graduada de la Jorge Tadeo Lozano y máster en Mercadeo Agroindustrial de la misma universidad. Trabajó durante seis años como profesora catedrática y miembro del grupo de investigación Redes Agroempresariales y Territorio (RAET) de la Universidad Jorge Tadeo Lozano, donde desarrolló varios trabajos relacionados con el sector hortofrutícola, en especies como ají y hortalizas de clima frío. En el año 2014 ingresó a la Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA como investigadora máster del Departamento de Transferencia de Tecnología, donde realizó trabajos enfocados en conocer los modelos de negocio desarrollados por las asociaciones de productores de semilla y en estudiar las dinámicas de la oferta y demanda de semilla en Colombia de especies como papa, cacao, plátano, aguacate y guayaba, entre otras. Desde el año 2020 está adscrita a la Red de Innovación de Frutales y se desempeña como analista de mercados en el Departamento de Mercadeo de Tecnología, Productos y Servicios.

Pedro David Suárez Villota

Correo: psuarezv@agrosavia.co

Orcid: <https://orcid.org/0000-0001-7389-5798>

Ingeniero agrónomo de la Universidad Nacional de Colombia. Trabajó en la Asociación Colombiana para el Avance de la Ciencia (ACAC) y el Ministerio de Agricultura y Desarrollo Rural (MADR). En el año 2014 ingresó a la Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA como profesional de investigación del Centro de Investigación Tibaitatá, donde participó en proyectos de investigación cualitativa para la caracterización cultural en el manejo de sistemas de producción y en la caracterización física, química, morfológica y nutraceútica de frutas para el manejo de cosecha y postcosecha. Ha desarrollado diversos proyectos relacionados con el fortalecimiento de capacidades de asistentes técnicos, extensionistas y productores agropecuarios y en caracterizaciones socioeconómicas con enfoque de investigación mixta. Desde 2019 se encuentra vinculado como profesional al Departamento de Mercadeo de Tecnologías, Productos y Servicios, donde participa en la formulación y ejecución de estrategias de vinculación de la oferta tecnológica y como enlace de la Red de innovación de Cacao.

Juan Carlos Fuentes Benítez

Correo: jfuentes@agrosavia.co

Orcid: <https://orcid.org/0000-0002-3132-7992>

Profesional en economía, máster en Ciencias Económicas de la Universidad Santo Tomás. Desde 2009 hasta 2014 trabajó como docente catedrático en la misma universidad. En 2014, ingresó a la Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA como investigador máster del área socioeconómica de la Dirección de Investigación y Desarrollo. Cuenta con una sólida formación en microeconomía aplicada y una amplia experiencia en trabajo comunitario, en la formulación de emprendimientos agropecuarios con población en condición de vulnerabilidad. También tiene fortalezas en líneas de investigación de agricultura familiar, campesina y comunitaria, y capacidad para entender la dinámica del sector agropecuario optimizando la aplicabilidad de conceptos económicos en los sistemas productivos agropecuarios. Posee conocimientos en métodos y modelos de evaluación de impacto ex-ante y ex-post de políticas y programas. Actualmente se encuentra adscrito a la Red de Innovación de Hortalizas y plantas aromáticas.



Glosario de símbolos matemáticos

N_h :	Número de unidades de la población en el h-ésimo estrato.
n_h :	Número de unidades de la muestra en el h-ésimo estrato.
C_0 :	Costo fijo independiente del número de elementos por seleccionar.
C_h :	Costo de muestrear un elemento en el estrato h.
y_{hi} :	Valor observado en la i-ésima unidad perteneciente al h-ésimo estrato.
$W_h = \frac{N_h}{n}$:	Ponderación del h-ésimo estrato.
$\bar{Y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$:	Estimador de la media muestral en el h-ésimo estrato.
$\hat{Y}_h = N_h \bar{Y}_h$:	Estimador del total en el h-ésimo estrato.
$s_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2}{N_{h-1}}$:	Cuasivarianza verdadera en el h-ésimo estrato.
$s_h = \sqrt{s_h^2}$:	Desviación estándar en el h-ésimo estrato.



Presentación

La presente metodología de muestreo se desarrolló y validó en el proyecto “Caracterización de la oferta y demanda de semilla de calidad de las especies de interés”. Este proyecto formó parte del macroproyecto “Programa nacional semillas: producción de semillas de calidad de variedades mejoradas y materiales regionales, para disponibilidad de los pequeños productores agrícolas”, ejecutado por la Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA entre los años 2013 y 2018, como una estrategia para apoyar la toma de decisiones en la planeación y ejecución de estudios socioeconómicos sobre el sector agropecuario con rigurosidad estadística.

La propuesta surgió en respuesta a la ausencia de un marco muestral, el cual es necesario para realizar estudios de este tipo. Por tal razón, se buscaron alternativas que permitieran identificar el número de productores o el número de unidades productoras agropecuarias dedicadas al cultivo de ciertas especies en las zonas de estudio. Se recurrió al Departamento Administrativo Nacional de Estadística (DANE), la autoridad nacional de estadística en Colombia, para revisar las fuentes de información oficiales —las Evaluaciones Agropecuarias Municipales (EVA), la Encuesta Nacional Agropecuaria (ENA) y el Tercer Censo Nacional Agropecuario (3^{er} CNA)—, y se llegó a la conclusión de que el censo contenía las bases de datos con mayor cantidad de información sobre cobertura geográfica, identificación de especies y características de producción. Además, el censo es una base de datos abierta, cuyos registros están anonimizados y cualquier ciudadano puede consultarlos en cualquier momento desde la página del Archivo Nacional de Datos (ANDA) del DANE (<http://andacna.dane.gov.co/index.php/home>).

El lector encontrará aquí una metodología que incluye el diseño estadístico, el diseño muestral, la definición de la muestra y el proceso metodológico para la preparación y procesamiento de las bases de datos reportadas en el 3^{er} CNA. También encontrará consideraciones referentes a la planeación de estudios por medio de aplicación de encuestas y por último algunas experiencias de uso.







Introducción

El objetivo de una encuesta realizada por muestreo es la estimación de parámetros de una población de interés. Dichos parámetros usualmente son características numéricas de la población, como la media, la varianza o el total. Estas medidas descriptivas de la población se denominan *parámetros* (Scheaffer et al., 1987). La importancia de la representatividad estadística de una muestra radica en que permite generalizar las estimaciones y extrapolarlas a una población a partir de un número reducido de observaciones (Otzen & Manterola, 2017).

Un inconveniente que se presenta con frecuencia para calcular el tamaño de una muestra es la insuficiente información respecto a la varianza poblacional, lo cual dificulta la elección de un método apropiado de muestreo. La teoría del muestreo ayuda a la selección de una muestra de una población mediante un mecanismo probabilístico. El mecanismo más frecuente es el muestreo aleatorio simple, en el cual se asigna igual probabilidad a cada unidad de la población de la muestra o probabilidad proporcional al tamaño (se asignan probabilidades distintas con base en algún otro criterio en una medida de tamaño) (Sánchez et al., 2010).

En un trabajo sobre costos de producción en Colombia, Perfetti et al. (2012) plantean que es necesario desarrollar una metodología que asegure calidad, robustez y veracidad estadística de la información del sector agropecuario, ya que las metodologías utilizadas por los diferentes actores tienen serios problemas de accesibilidad y representatividad estadística. Los datos y la información por lo general son tomados de fuentes secundarias para construir estimaciones teóricas. En casos muy particulares se acude a información primaria (por medio de visitas a fincas o a productores seleccionados de manera

aleatoria), pero sin emplear métodos de muestreo, debido posiblemente a la ausencia de un censo agropecuario actualizado que permita definir y acceder a un marco muestral de unidades productoras agropecuarias.

El anterior planteamiento no solo es aplicable a los estudios de tipo socioeconómico, sino también a otras investigaciones en las que se recolectan datos y se construye información del sector agropecuario, independientemente de la especie productiva (Gutiérrez López, 2009).

La rigurosidad estadística es crucial para dirigir las investigaciones y los retos económicos y sociales de un país (Perfetti et al., 2012). Es por esto que las investigaciones socioeconómicas deben ir de la mano de la correcta aplicación de métodos estadísticos de muestreo que contribuyan a hacer frente a las necesidades y desafíos sociales de un país, abordando aspectos científicos y técnicos. Con esto, además, las investigaciones ayudan a generar conocimiento; desarrollar avances tecnológicos; mejorar la competitividad, equidad y sostenibilidad, e innovar en procesos y productos. También contribuyen a crear nuevos modelos de negocios que soporten el desempeño económico y ayuden a enfrentar los desafíos de la sociedad, sin perder de vista la relevancia que tiene el sector agropecuario en el país tanto en temas de seguridad alimentaria como en la mejora de la calidad de vida de la población (Wallerstein, 1999).

En consecuencia, las investigaciones socioeconómicas son de vital importancia para entender las estructuras, el comportamiento y las necesidades de la población en un territorio. El sector agropecuario no es ajeno a esta necesidad de acceder a una información veraz y oportuna que sirva de soporte para la toma de decisiones y permita crear valor agregado en los sistemas productivos para competir en mercados cada día más sofisticados.

El territorio y su delimitación no pueden depender solamente de criterios geográficos o de la división administrativa del país, ya que con esto se corre el riesgo de sobrevalorar o subvalorar los alcances reales de diferentes regiones (Kalmanovitz & López López, 2006). Es fundamental entender y analizar la distribución espacial de la población, su conformación social, sus condiciones de vida, su cultura y costumbres, las dinámicas de las comunidades, sus sistemas de producción y las zonas de influencia del área que habitan. Es en estos factores donde las técnicas de muestreo desempeñan un papel importante.

De otro lado, como lo registra Soto Agudelo (2016), la disponibilidad de información estadística oportuna y de calidad en el sector agropecuario es determinante para la toma de decisiones por parte de los actores de la política pública y de otros relacionados con el sector. Sin embargo, en Colombia existen serias dificultades con la calidad de la información disponible. Estas dificultades se deben a las metodologías utilizadas en la planeación y la captura de información, a la falta de robustez estadística, a la existencia de múltiples y dispersas fuentes de información, a la duplicidad de esfuerzos en los mismos temas de investigación, al limitado acceso a la información recopilada por otras instituciones y al uso inefectivo de los recursos.

Ante la dificultad de acceder a listados de productores y unidades productoras, y la imposibilidad de conocer las áreas por sistemas de producción del sector agropecuario, se diseñó una estrategia para calcular las varianzas que requiere el cálculo del tamaño de muestra. Para dicha estrategia se empleó la base de datos anonimizados del 3^{er} CNA, disponible en el ANDA, el cual es un catálogo que puede explorarse en la web para buscar, comparar, solicitar acceso y descargar información de censos y encuestas por muestreo.

Estos registros se emplearon porque abarcaban todos los sitios de Colombia y porque se recolectaron bajo tres modalidades: 1) *Modalidad de barrido*: para regiones de alta densidad de población, municipios con formación catastral y buena infraestructura vial, en los que se desarrolla mayoritariamente la actividad agropecuaria. 2) *Modalidad ruta*: para regiones con baja densidad de población, zonas apartadas, baja infraestructura vial, baja actividad agropecuaria, alta cobertura de bosques naturales y predominio de territorios étnicos y parques nacionales. 3) *Grandes productores*: empresas agropecuarias y no agropecuarias, con sedes administrativas en las áreas urbanas, con presencia nacional, unas especializadas y otras diversificadas, y tamaños grandes en área o en número de cabezas de animales (Departamento Administrativo Nacional de Estadística [DANE], 2016b).

AGROSAVIA, al ejecutar los proyectos de investigación de la agenda dinámica corporativa, requiere información oficial para determinar el tamaño de las muestras que se usarán en los procesos de investigación. Con la presente metodología se facilita el uso de los datos del 3^{er} CNA para determinar los parámetros poblacionales y estimar el tamaño de las muestras.



La importancia de la selección del método de muestreo y del tamaño óptimo de muestra en cualquier investigación debe ser entendida como una decisión estratégica que requiere ser abordada con mayor profundidad. Es por esta razón que se presenta un breve marco conceptual, el cual incluye una revisión de los diferentes tipos de muestreo, así como las definiciones de algunos términos y variables con los que se desarrolla la propuesta metodológica entregada.

Marco teórico

El objetivo de toda investigación social, observacional y de tipo cuantitativo o cualitativo es estimar, explicar, predecir e incluso, en algunos casos, controlar una variable o una interrelación de variables. El estado del arte y la literatura especializada sobre las metodologías para definir métodos de muestreo y estimar tamaños de muestra con diversas variables, y sobre los métodos de análisis estadísticos apropiados para los datos, no son de fácil manejo por parte de los investigadores y tomadores de decisiones (Rositas Martínez, 2014).

Varios autores coinciden en que una decisión importante en cualquier proceso de investigación es la selección adecuada del método y por ende del tamaño muestral. Sin embargo, en la práctica no es fácil determinar dicho valor, debido a que es necesario disponer de datos e información para estimar parámetros poblacionales como varianza poblacional o desviación poblacional. En algunos casos, dicha estimación se puede hacer a partir de datos históricos, recurriendo a reportes en literatura especializada o basándose

en la experiencia de un experto en la materia (Valdivieso Taborga et al., 2011). En algunos casos muy específicos, se puede llevar a cabo un estudio piloto para que los datos recolectados permitan estimar los parámetros, pero esto significa incurrir en mayores costos.

En Colombia, con anterioridad se han empleado bases de datos de censos y encuestas nacionales para estimar la varianza y la desviación poblacional que se necesitan para determinar el tamaño de una muestra. En estudios como los realizados por Zapata-Ossa et al. (2011), los tamaños de muestra se calcularon mediante información obtenida de la Encuesta de Calidad de Vida de 2008. Acevedo González y Palacio Martínez (2016) emplearon el 3^{er} CNA como fuente primaria para estimar el número de productores en el oriente Antioqueño, y Bonilla Sebá (2018) utilizó la misma base de datos para estudiar la concentración de la tenencia de tierra según áreas y unidades productoras.

Breve introducción al muestreo

Los métodos para seleccionar una muestra se ven afectados por factores como aleatoriedad, costos de recogida y desplazamiento, tiempo y habilidad para la toma de la muestra o aplicación del instrumento de recolección, así como por la naturaleza de los elementos que conforman una población. El muestreo se divide en dos tipos, dependiendo de si se conocen o no las características de la muestra y del valor de probabilidad de elegir un elemento de la población (figura 1).

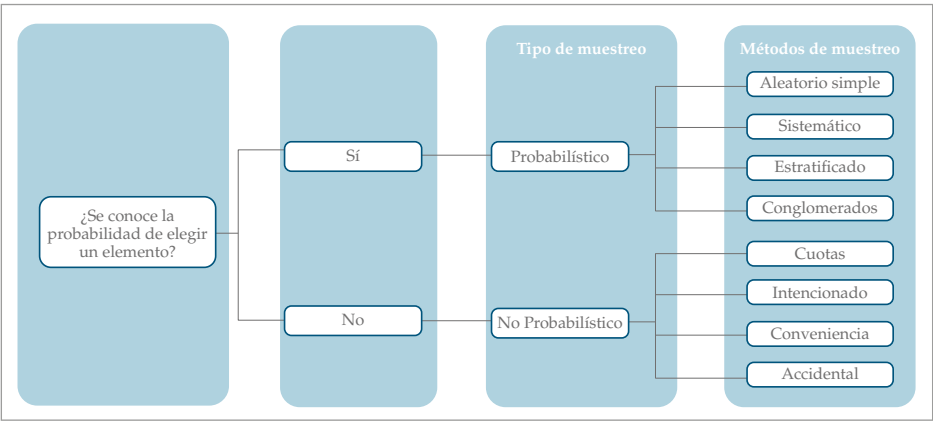


Figura 1. Esquema general para la decisión del método de muestreo.

Fuente: Elaboración propia con base en Berenson y Levine (1996)

1. *Muestreo no probabilístico*: las muestras no son representativas en términos estadísticos. Influyen las personas que toman la muestra y esto puede afectar el resultado. En situaciones muy concretas, los errores cometidos no son grandes, debido a la homogeneidad de la población, pero en general no es un tipo de muestreo riguroso ni científico, dado que no todos los elementos de la población pueden formar parte de la muestra (ausencia de marco muestral). Además, no se conoce la probabilidad de elegir un determinado elemento de la población. Existen diversos métodos de muestreo no probabilísticos y su principal diferencia radica en el método de elección de los elementos a muestrear (Cochran, 1977; Kish, 1965). A continuación, se mencionan los métodos de uso más frecuente:
 - a. *Muestreo por cuotas*: consiste en dividir la población en grupos basándose en ciertas características. El tamaño de muestra de cada grupo debe ser proporcional al tamaño de la población. Este tipo de muestreo es de amplio uso en los estudios de opinión.
 - b. *Muestreo intencionado*: se concentra en ciertas características seleccionadas de una población¹ de acuerdo con el objetivo del estudio. En general, el tamaño de la muestra es pequeño. Este tipo de muestreo es el empleado en los estudios de caso.
 - c. *Muestreo de conveniencia*: en este método la muestra se compone de elementos seleccionados por su disponibilidad.
 - d. *Muestreo accidental*: método en el que se toman unidades de muestreo hasta completar un número de unidades con una característica deseada. Los elementos son elegidos casualmente. Este tipo de muestreo siempre es sesgado.
2. *Muestreo probabilístico*: su característica más notable es que, de cada elemento de la población, se conoce el valor de probabilidad de ser incluida en la muestra (el valor de probabilidad no tiene por qué ser igual para todos los elementos). Las muestras probabilísticas exigen métodos que escojan los elementos de manera aleatoria (Cochran, 1977; Kish, 1965; Scheaffer et al., 2007; Särndal et al., 2003). Las siguientes son las ventajas del proceso probabilístico:

¹ La diferencia con el muestreo por cuotas radica en que para este último se debe guardar la proporcionalidad del número de elementos originales en la población, de acuerdo con una variable categórica seleccionada (sexo, religión, estrato, etc.) que permita agrupar a esos elementos. En un muestreo intencionado se busca que los elementos cumplan una característica de interés relacionada con el objetivo del estudio (personas con una enfermedad específica, propietarios de un tipo de móvil, deportistas de élite) y solo se muestrea un pequeño grupo de elementos.

- Permite conocer de antemano la probabilidad de que cada elemento de la población pertenezca a la muestra (probabilidad de inclusión) y este valor es diferente de cero (Särndal et al., 2003).
- Permite estimar el grado de error muestral.
- Permite determinar matemáticamente el tamaño óptimo de la muestra, un grado de confianza y un error máximo permitido.
- Proporciona estimaciones que son esencialmente insesgadas y tienen una precisión medible.

Las principales desventajas son que requiere mayor tiempo para planificar y ejecutar el muestreo, demanda mayor tiempo para realizar los cálculos y costos más altos para recolectar la información y exige más complejidad en comparación con los métodos no probabilísticos. Según Cochran (1977), Kish (1965), Scheaffer et al. (2007) y Särndal et al. (2003), los siguientes son algunos tipos de muestreo probabilístico de amplio uso:

- a. *Muestreo aleatorio simple*: es el sistema más conocido y usado y en él cada elemento de la población tiene igual probabilidad de formar parte de la muestra. Teóricamente equivale a convertir cada elemento de la población en un número y escoger, aleatoriamente y sin reemplazo, tantos números como se requieran para completar el tamaño de la muestra calculado. La selección debe ser cuidadosa para garantizar que todos los elementos de la población tengan la misma probabilidad de aparecer en la muestra.

Este método es especialmente útil en casos de poblaciones pequeñas, cuando se dispone de una lista (marco de lista) de todos los elementos que conforman la población y cuando la dispersión geográfica no constituye un problema para recoger los datos y hacer las estimaciones.

Existen varios procedimientos para seleccionar la muestra (Särndal et al., 2003; Vivanco, 2005) (figura 2), pero el punto clave es tener un marco de lista actualizado de todos los elementos de la población que permita su identificación. Este marco de lista actualizado permite recolectar la información que requiere la encuesta (Särndal et al., 2003), de forma tal que si se extraen varias veces las muestras, estas representarán a la población y generalmente producirán estimaciones razonables de los parámetros. Cuando hay problemas con los listados, se presentan sesgos en los resultados y en las inferencias respecto a la población.

Poblaciones finitas	Intervalos regulares	Números aleatorios
- Sin reemplazo - Los elementos solo podrán ser elegidos una vez. - El valor de probabilidad para un elemento cambia a medida que se extraen elementos. - Con reemplazo - Una vez medido u observado el elemento, este es devuelto al marco y mantiene la misma probabilidad de ser elegido. - Es aplicado cuando se usa muestreo por conglomerados o existe submuestreo.	- Todos los elementos deben estar enumerados. - Se divide el tamaño total de la población por el tamaño de la muestra. - En el primer intervalo se selecciona un elemento aleatoriamente y luego se suma el valor del intervalo y se continúa sucesivamente.	- Es análogo al extraer números, por ejemplo de una balotera. - Su uso es sencillo dado que existen tablas y los programas de computador pueden generar diferentes series de números para la selección de los elementos enumerados previamente. - Es el de mayor uso.

Figura 2. Métodos de extracción de elementos de una muestra.

Fuente: Elaboración propia

- b. *Muestreo sistemático*: es una variante del muestreo aleatorio simple. Consiste en seleccionar, de los elementos ordenados de la población, un punto de partida aleatorio, escogiendo después cada k-ésimo elemento de la lista. Es de amplio uso porque permite simplificar la elección de la muestra.
- c. *Muestreo estratificado*: la población se divide en subpoblaciones denominadas *estratos*, los cuales deben ser heterogéneos en cuanto a la característica de interés, pero los elementos dentro de cada estrato deben ser homogéneos. Dentro de cada estrato se selecciona una muestra de forma independiente, aplicando muestreo aleatorio simple, sistemático u otra técnica de muestreo probabilístico. El tamaño de la muestra final será la suma de los tamaños de estas submuestras en cada estrato. Los estratos pueden estar conformados naturalmente o se pueden crear teniendo en cuenta las características de la población
- d. *Muestreo por conglomerados*: para obtener una muestra por conglomerados, se divide la población en subpoblaciones generalmente heterogéneas, de tamaños desiguales. Se selecciona una muestra aleatoria de conglomerados y se toman todos los elementos o parte de ellos (por ejemplo, empleando muestreo aleatorio simple) para obtener la muestra definitiva. Con este método, no todos los conglomerados son muestreados, debido a que cada uno puede tener diferentes

probabilidades de ser seleccionado. Este tipo de muestreo supone un método de trabajo opuesto al estratificado.

- e. *Muestreo por etapas*: es una forma de muestreo menos eficiente que el aleatorio simple, pero facilita la elección de los elementos de la muestra aleatoria, especialmente cuando no se dispone de un marco de lista. En cada etapa, la unidad de muestreo no se corresponde con aquella en la cual se observan o se colectan los datos. Se pueden incluir tantas etapas como sean necesarias. Entre las principales ventajas, está el que no necesita un listado completo de todos los posibles elementos, sino que es suficiente con los listados de los conglomerados que fueron elegidos en las sucesivas etapas. El principal inconveniente radica en que aumenta el número de errores muestrales (uno por cada etapa) (Abascal Fernández & Grande Esteban, 2005).

Marco conceptual

A continuación se presentan los conceptos más relevantes para comprender la temática, que se emplean a lo largo de este documento para facilitar la aplicación de la presente metodología.

Área cosechada: es la superficie de la cual se obtiene la producción. Esta variable se genera a partir del momento en que se inicia la recolección, la cual se puede hacer en una sola ocasión o en varios cortes o recolecciones.

Área sembrada: es la superficie agrícola en la cual se deposita la semilla de cualquier cultivo, previa preparación del suelo.

Censo: procedimiento estadístico mediante el cual se mide la totalidad del universo de estudio. Según las unidades de observación se puede hacer un censo de población y vivienda, un censo económico, un censo agropecuario, entre otros.

Censo agropecuario: es una operación estadística que recoge, procesa y difunde datos sobre la estructura del sector agropecuario de todo un país o de una parte importante de este. Los datos estructurales recogidos dan

información sobre tamaño de la explotación agropecuaria, tenencia y uso de la tierra, área agrícola, riego, población ganadera, mano de obra y otros insumos agropecuarios. En un censo agropecuario, los datos se recogen en las explotaciones agrícolas, pero a veces también se pueden recolectar en el ámbito comunitario (Organización de las Naciones Unidas para la Alimentación y la Agricultura [FAO], 2016b).

Muestra: colección de unidades de muestreo seleccionadas con el fin de estimar parámetros de una población. Las unidades pueden seleccionarse de diversas formas (Vivanco, 2005; Scheaffer et al., 2007).

Unidad Productora Agropecuaria (UPA): unidad de organización de la producción agropecuaria. Debe cumplir con las siguientes condiciones: 1) producir bienes agrícolas, forestales, pecuarios, acuícolas, 2) tener un único productor que puede ser una persona natural o jurídica que asuma la responsabilidad y los riesgos, 3) utilizar mínimo un medio de producción en los predios que la integran, como construcciones, maquinaria, equipo y/o mano de obra.

Productor: persona natural o jurídica que dirige la UPA y toma las decisiones principales sobre el cultivo de plantas, la cría de animales, las prácticas agropecuarias, el uso de los medios de producción y la venta de productos agropecuarios.

Tamaño de muestra: número de elementos que componen una muestra extraída de una población. Se representa con la letra n .

Tercer Censo Nacional Agropecuario (3^{er} CNA): operación estadística que apunta a proporcionar y actualizar información estratégica georreferenciada sobre el sector agropecuario, forestal, acuícola y pesquero y sobre aspectos ambientales. También proporciona información sobre productores agropecuarios y población residente en el área rural dispersa, para la toma de decisiones de política pública y la estructuración del Sistema Estadístico Agropecuario (SEA), componente del Sistema Estadístico Nacional (SEN) (DANE, 2016a).



La definición del diseño estadístico incluye todas las actividades que están relacionadas con la selección de la muestra, tales como el esquema de muestreo, el diseño de la muestra y la selección de los elementos en la población. Generalmente, este diseño debe ser desarrollado en conjunto por personas con conocimientos en muestreo y personal con amplia experiencia en el campo u objetivo de estudio.

Marco estadístico

La estimación de los parámetros necesarios para determinar el tamaño de la muestra se realizó empleando todos los registros de las UPA contenidos en el 3^{er} CNA que cumplían los siguientes criterios:

1. Registro con cédula catastral, es decir, identificación con código de encuesta diferente a 999999999.
2. Ubicación en territorio con predominancia de grupos étnicos diferentes a comunidades indígenas, negras, afrocolombianas, raizales o palenqueras.
3. Producción de cultivos agrícolas de interés.

Definición de variables

Área sembrada: zona ocupada por un cultivo o por una asociación de cultivos en forma compacta. En el caso de cultivos asociados, el área sembrada es la misma para todas las especies que forman parte del policultivo.

Área cosechada: área total en la cual se recolecta un cultivo. Se excluye el área destruida por sequías, inundaciones, plagas o cualquier otra razón (FAO, 2016b).

Rendimiento: cantidad de productos que se obtiene por unidad de superficie cosechada. Su fórmula equivale a la cantidad cosechada en 2013 dividida por el área cosechada (ecuación 1).

$$\text{Rendimiento 2013 (t/ha)} = \frac{\text{Cantidad cosechada 2013 (t)}}{\text{Área cosechada en hectáreas (ha)}} \quad \text{Ecuación 1.}$$

Universo

Está compuesto por la totalidad del área rural del país que figura en los registros catastrales del Instituto Geográfico Agustín Codazzi (IGAC). Lo conforman 3.946.522 registros prediales, 773 resguardos indígenas (31.569.990 ha) y 181 tierras de comunidades negras (5.322.982 ha), con lo cual se cubren 113.008.623,9 ha, equivalentes al 99 % del área rural continental e insular del país y distribuidas en 1.101 municipios, 20 corregimientos departamentales y el Archipiélago de San Andrés, Providencia y Santa Catalina, así como en la parte continental de los 56 Parques Nacionales Naturales existentes hasta 2013 (DANE, 2016b).

Población objetivo

Por *población objetivo* se entiende el conjunto de todos los elementos en los cuales se desea estudiar un fenómeno. Tales elementos están acotados en el tiempo y el espacio y poseen una característica común que puede ser medida u observada.

La población debe ser definida de acuerdo con los objetivos del estudio y puede delimitarse en términos geográficos (país, municipio, localidad o cualquier otra categoría intermedia) o sectoriales (población urbana, industrias de cerámica, productores de leña). También se deben establecer

límites temporales en la definición del universo, ya que su composición y sus características pueden cambiar con el correr del tiempo. Se recomienda que los límites espaciales del universo coincidan con agrupaciones estándares (de uso común) u oficiales del país (políticas, administrativas, naturales, etc.), para poder estimar sus dimensiones a partir de bases de información ya existentes (FAO, 2003).





El conjunto de estrategias y procedimientos que permiten seleccionar una muestra con características deseables a partir de una población objetivo se denomina *diseño muestral*. Este diseño permite especificar la forma como se elegirán los elementos sobre los cuales se llevará a cabo el estudio.

Tipo de muestreo

Se determinó que el mejor método de muestreo es el aleatorio estratificado porque dentro de la población objetivo (productores con un cultivo determinado entre los años 2013 y 2014), existe una alta heterogeneidad en términos de área sembrada y cosechada. El conjunto de elementos se dividió en subpoblaciones o estratos con el objetivo de reducir la variabilidad y alcanzar mayor precisión (Kish, 1965). Este tipo de muestreo es potente y flexible, por lo cual es de amplio uso en la práctica (Särndal et al., 2003).

La definición del número de elementos de un estrato que formará parte del tamaño total de la muestra depende de varios factores:

1. Tamaño relativo del estrato en la población (afijación proporcional).
2. Variabilidad de la variable de interés dentro de cada estrato (afijación óptima).
3. Costos del muestreo de los elementos dentro de cada estrato (afijación óptima de costo).

Las siguientes son las principales razones para emplear el muestreo aleatorio estratificado en lugar del tradicional aleatorio simple (Scheaffer et al., 1987; Särndal et al., 2003):

1. La estratificación produce un error de estimación más pequeño que el generado por una muestra irrestricta aleatoria del mismo tamaño. El resultado es particularmente cierto si las mediciones dentro de los estratos son homogéneas.
2. El costo de medir u observar un elemento puede reducirse mediante la conformación de estratos de los elementos de la población
3. Si el propósito es obtener datos de precisión conocida, es necesario considerar cada estrato como una población (Cochran, 1977).

En el momento de realizar la estratificación deben tenerse en cuenta los principios que se emplean para decidir si es necesario crear estratos dentro de la población y cómo se delimitan. Los estratos pueden definirse considerando una o más variables y estas deben tener una alta correlación con la variable de interés (objetivo del estudio), con el fin de generar estratos lo más homogéneos posibles. La construcción de los estratos debe incluir algunos requisitos, como precisión de la estimación para los diferentes dominios, y debe contemplar variables de estudios, costos y restricciones administrativas para la recolección de la información (Särndal et al., 2003).

En la elección de los elementos que conformarán la muestra, es fundamental que los estratos se identifiquen claramente en el marco muestral para que cada elemento de la población pertenezca a un solo estrato. También es importante que cada estrato sea independiente, ya que así se pueden realizar inferencias de cada una de las poblaciones que originaron la muestra (Cochran, 1977).

Estimación de la media y el total poblacional

El estimador de la media poblacional utilizado en el muestreo estratificado es la media estratificada. Este estimador tiene propiedades deseables (ser insesgado y tener mínima varianza) y se expresa mediante la ecuación 2:

$$\bar{Y}_{st} = \frac{\sum_{h=1}^L N_h \bar{Y}_h}{N} = \sum_{h=1}^L W_h \bar{Y}_h \quad \text{Ecuación 2,}$$

donde $\bar{Y}_{st}$ es la media estratificada, N_h el número de elementos de la población en el h-ésimo estrato, $\bar{Y}_h$ la media muestral en el h-ésimo estrato, N el número de elementos de la población, y W_h la ponderación del h-ésimo estrato.

En un muestreo estratificado, la varianza del estimador de la media poblacional ($\bar{Y}_{st}$) se formula como se indica en la ecuación 3:

$$V(\bar{Y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{s_h^2}{n_h} \quad \text{Ecuación 3,}$$

donde $V(\bar{Y}_{st})$ es la varianza de la media estratificada, N el número de elementos de la población, N_h el número de elementos de la población en el h-ésimo estrato, n_h número de elementos de la muestra en el h-ésimo estrato, y s_h^2 la cuasivarianza verdadera en el h-ésimo estrato.

La estimación del total poblacional Y se expresa mediante la ecuación 4:

$$\hat{Y}_{st} = N\bar{Y}_{st} \quad \text{Ecuación 4,}$$

donde $\hat{Y}_{st}$ es la estimación del total poblacional, N el número de elementos de la población, y $\bar{Y}_{st}$ la media estratificada.

Y la varianza del total poblacional se formula con la ecuación 5:

$$V(\hat{Y}_{st}) = \sum_{h=1}^L N_h \frac{s_h^2}{n_h} (N_h - n_h) \quad \text{Ecuación 5,}$$

donde $V(\hat{Y}_{st})$ es la varianza de la media estratificada, N_h el número de elementos de la población en el h-ésimo estrato, s_h^2 la cuasivarianza verdadera en el h-ésimo estrato, y n_h el número de elementos de la muestra en el h-ésimo estrato.

Definición del tamaño de la muestra

El tamaño de la muestra para la media y el total equivale al número de elementos que se deben muestrear de la población, esto con el fin de estimar un parámetro con determinados niveles de confianza y error de estimación.

Existen varios métodos o tipos de afijación para repartir los elementos dentro de los estratos, los cuales se explican a continuación (figura 3) (Cochran, 1977; Särndal et al., 2003; Scheaffer et al., 1987; Kish, 1965):

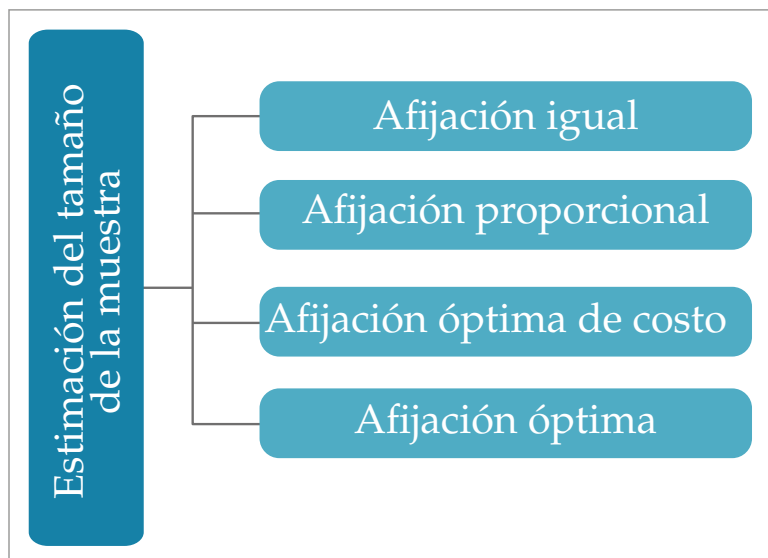


Figura 3. Tipos de afijación para calcular el tamaño de la muestra.

Fuente: Elaboración propia

Afijación igual

El tamaño de la muestra se divide entre el número total de estratos para que se pueda seleccionar el mismo número de elementos en cada uno, independientemente del peso del tamaño del estrato en la población. De todos los métodos para calcular el tamaño de la muestra, este es el de menor uso.

Afijación proporcional

Consiste en asignar a cada estrato un número de unidades muestrales proporcional a su tamaño. Las n unidades de la muestra se distribuyen de manera proporcional a los tamaños de los estratos, los cuales se expresan en número de unidades (ecuación 6).

$$n = \frac{\sum_{h=1}^L W_h s_h^2}{\frac{e^2}{z_{\alpha/2}^2} + \frac{1}{N} \sum_{h=1}^L (W_h s_h^2)} \quad \text{Ecuación 6,}$$

donde n es el número de elementos de la muestra, W_h la ponderación del h -ésimo estrato, s_h^2 la cuasivarianza verdadera en el h -ésimo estrato, e el error máximo admisible, $z_{\alpha/2}$ el valor tabulado del coeficiente de confianza, y N el número de elementos de la población.

El tamaño de muestra por estrato se expresa por medio de la ecuación 7:

$$n_h = n W_h \quad \text{Ecuación 7,}$$

donde n_h es el número de elementos de la muestra en el h -ésimo estrato, n el número de elementos de la muestra, y W_h la ponderación del h -ésimo estrato.

Afijación óptima de costo

La afijación óptima de costo consiste en determinar los valores de n_h (número de unidades que se extraen del h -ésimo estrato para la muestra), de forma que para un coste fijo C , la varianza de los estimadores sea mínima. En este tipo de afijación, se tiene en cuenta la variabilidad dentro de cada estrato, así como el costo de recolectar la información. La función de costo es una función lineal (ecuación 8):

$$\text{Costo (C)} = C_0 + \sum_{h=1}^L C_h n_h \quad \text{Ecuación 8,}$$

donde C es la función de costo, C_0 el costo fijo independiente del número de elementos por seleccionar, C_h el costo de muestrear un elemento en el h -ésimo estrato, y n_h el número de elementos de la muestra en el h -ésimo estrato.

El tamaño de la muestra se determina mediante la ecuación 9:

$$n = \frac{(\sum_{h=1}^L W_h s_h \sqrt{C_h}) \sum_{h=1}^L (W_h s_h / \sqrt{C_h})}{\frac{e^2}{z_{\alpha/2}^2} + \frac{1}{N} \sum_{h=1}^L (W_h s_h^2)} \quad \text{Ecuación 9,}$$

donde n es el número de elementos de la muestra, W_h la ponderación del h -ésimo estrato, s_h la desviación estándar en el h -ésimo estrato, C_h el costo de muestrear un elemento en el h -ésimo estrato, e el error máximo admisible, $z_{\alpha/2}^2$ el valor tabulado del coeficiente de confianza, y N el número de elementos de la población.

El tamaño de muestra por estrato se formula tal como se indica en la ecuación 10:

$$n_h = n \frac{W_h s_h / \sqrt{C_h}}{\sum_{h=1}^L (W_h s_h / \sqrt{C_h})} \quad \text{Ecuación 10,}$$

donde n_h es el número de elementos de la muestra en el h -ésimo estrato, n el número de elementos de la muestra, W_h la ponderación del h -ésimo estrato, s_h la desviación estándar en el h -ésimo estrato, y C_h el costo de muestrear un elemento en el h -ésimo estrato.

El costo afecta de forma inversa el tamaño de la muestra. A mayor costo por estrato, menor el tamaño de la muestra, mientras que la variabilidad afecta de manera directa el número de muestras que deben ser tomadas (a mayor varianza, mayor número de muestras).

Afijación óptima

También llamada *afijación de Neyman*, consiste en determinar los valores del número de unidades extraído del estrato, de forma tal que, con un tamaño de muestra fijo igual a n , la varianza de los estimadores sea mínima. Si el costo por unidad de muestreo es el mismo en cada estrato y si se fija el valor de la varianza, la fórmula para determinar el tamaño de la muestra sería la siguiente (ecuación 11):

$$n = \frac{(\sum_{h=1}^L W_h s_h)^2}{\frac{e^2}{z_{\alpha/2}^2} + \frac{1}{N} \sum_{h=1}^L (W_h s_h^2)} \quad \text{Ecuación 11,}$$

donde n es el número de elementos de la muestra, W_h la ponderación del h -ésimo estrato, s_h la desviación estándar en el h -ésimo estrato, s_h^2 el error máximo admisible, $z_{\alpha/2}$ el valor tabulado del coeficiente de confianza, y N el número de elementos de la población.

El tamaño de muestra por estrato se expresa mediante la siguiente fórmula (ecuación 12):

$$n_h = n \frac{W_h s_h^2}{\sum_{h=1}^L W_h s_h^2} \quad \text{Ecuación 12,}$$

donde n_h es el número de elementos de la muestra en el h -ésimo estrato, W_h la ponderación del h -ésimo estrato, n el número de elementos de la muestra, y s_h la desviación estándar en el h -ésimo estrato.

El tamaño de la muestra aumenta si se cumple alguno de los siguientes criterios:

1. El estrato es más grande.
2. El estrato es más variable (heterogéneo).
3. El muestreo tiene menor costo en el estrato.

Algunas consideraciones sobre la planeación de estudios por medio de aplicación de encuestas

Una encuesta es una técnica que permite obtener información de forma objetiva, en la que los datos recogidos pueden ser analizados empleando métodos cuantitativos y los resultados pueden ser extrapolables a la población. Entre sus múltiples ventajas, está la posibilidad de estandarizar las preguntas que se formulan a los elementos de la población, con lo cual se hacen fácilmente administrables, se simplifica la recogida y el tratamiento de los datos y se puede obtener información que a simple vista no está disponible (Abascal Fernández & Grande Esteban, 2005). Planear y hacer una encuesta con rigor estadístico exige surtir una serie de etapas (figura 4), algunas de las cuales se exponen a continuación.

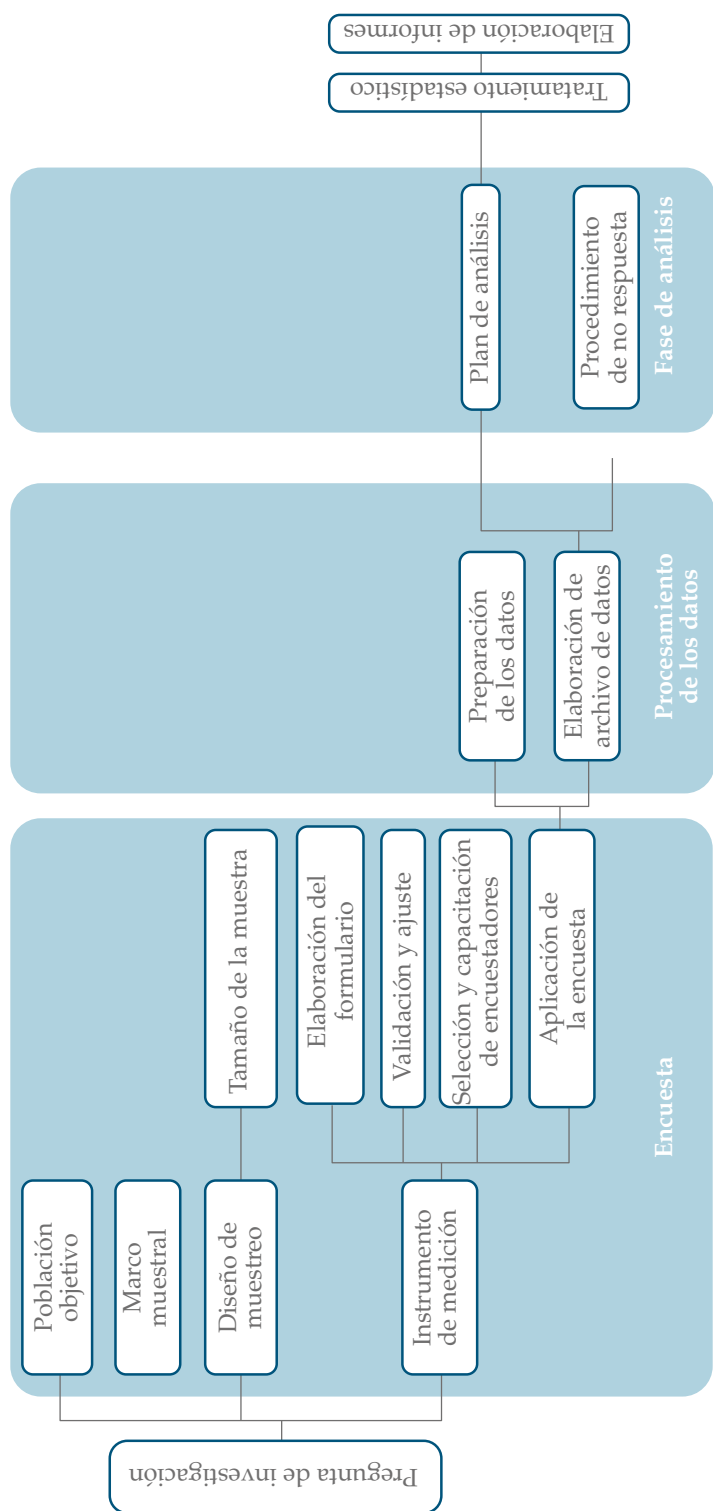


Figura 4. Proceso para el diseño, ejecución y análisis de datos de un estudio mediante encuesta.

Fuente: Elaboración propia

1. *Identificación del problema:* el problema surge de la pregunta de investigación. En esta etapa se deben definir claramente los objetivos y revisar estudios similares. En casos en los que se tenga poco conocimiento sobre el problema o la población (o sobre ambos), es necesario hacer una revisión de los temas relacionados y recurrir a otras técnicas para recabar la información que no se puede obtener por otros medios.
2. *Definición de la población objetivo:* se considera la UPA como predio, como parte de un predio o como conjunto de predios (sin importar su tamaño, régimen de tenencia y ubicación), y se considera también la actividad agropecuaria de interés que se desarrolla en la unidad, bajo la dirección de un productor agropecuario. Se establecen como segundo objeto de estudio las personas naturales o jurídicas, residentes habituales o no en el sector rural del territorio nacional, bajo cuya responsabilidad se toman las principales decisiones sobre la utilización de los recursos disponibles, el uso de los factores de producción (trabajo, capital, maquinaria, etc.) y el control administrativo de la UPA.
3. *Levantamiento del marco muestral:* enumeración de todos elementos del universo o población para hacer una selección aleatoria dentro de ellos. Para el levantamiento de los listados se recomienda visitar asociaciones, gremios, cooperativas, secretarías de agricultura municipales, Unidades Municipales de Asistencia Técnica Agropecuaria (Umata), entre otros. Es importante recalcar que no existe el marco muestral perfecto y que la actualización de dicho marco es fundamental en el proceso.
4. *Diseño de muestreo:* aunque existen diferentes tipos de muestreo, todos se basan en el principio de aleatoriedad. Esto significa que todos los elementos del universo deben tener oportunidad de ser elegidos para componer la muestra. Si no se eligieran aleatoriamente, se correría el grave riesgo de que los resultados obtenidos fueran representativos, no de toda la población, sino de un solo sector, algo a lo cual se le llama *sesgo*. Para poder hacer inferencias válidas de lo que ocurre, es necesario que la muestra sea representativa del universo, y esto se logra con la aleatoriedad y con un tamaño adecuado de la muestra. El tamaño de la muestra se define según la variabilidad del fenómeno a estudiar y según el parámetro de interés, el nivel de confianza fijado y el error admisible. Un error muy común es pensar que la muestra, para que sea representativa, debe tener un tamaño directamente proporcional al tamaño del universo, es decir, que a mayor tamaño del universo, mayor debería ser el tamaño de la muestra.

5. *Instrumentos de medición:* el instrumento de recogida de información es quizás el elemento más importante de una investigación basada en la aplicación de encuestas. La calidad con la cual se construya el instrumento será determinante para recoger con exactitud la información que se desea obtener (Martínez Medrano, 2004). Las siguientes son algunas recomendaciones para el diseño de cuestionarios según la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) (2003):
 - a. El tipo y número de preguntas dependerá de las variables para analizar. No se deben incorporar preguntas que no tengan relación directa con el objeto de estudio. Es importante que las preguntas sean breves, usen vocabulario sencillo y obedezcan a una secuencia acorde con la lógica del entrevistado.
 - b. En lo posible, las preguntas deben ser de respuesta cerrada, dado que así su procesamiento es más sencillo. Las preguntas abiertas son muy útiles para detectar opiniones, percepciones y preferencias.
 - c. Para formular el cuestionario definitivo, es necesario poner a prueba uno preliminar con una pequeña muestra de la población destino. Así se puede conocer la variabilidad de respuestas posibles o situaciones no consideradas originalmente.
 - d. Cuando se trabaje con muestras muy grandes y con preguntas complejas, es conveniente que estas últimas tengan instrucciones precisas de aplicación y que dichas instrucciones se resalten en el cuestionario.
6. *Selección y capacitación de los encuestadores:* los encuestadores son las personas encargadas de aplicar el instrumento y recolectar los datos, de ahí la importancia de seleccionarlos y capacitarlos adecuadamente. Estas personas deben conocer los objetivos del estudio, el diseño de muestreo, la técnica de obtención de datos y su procesamiento.
7. *Validación y ajuste del formulario:* esta etapa se denomina en algunos casos *pretest* o *prueba piloto de la encuesta* y pretende averiguar si el lenguaje es apropiado y si las preguntas del formulario son claras y guardan un orden coherente. En esta etapa también se estiman el tiempo y el costo que requiere la aplicación del instrumento. Los encuestados deben ser personas con características similares a los elementos de la población objetivo.
8. *Preparación de los datos colectados:* una vez recogidos los datos, es necesario revisar los cuestionarios (edición) e inspeccionar, corregir y en ocasiones precodificar las respuestas de acuerdo con un conjunto de

reglas fijas. Tras la inspección y corrección de las respuestas, se elabora el archivo de datos (Díaz de Rada Igúzquiza, 2009). En algunos casos, la codificación se puede hacer en esta etapa, aunque por lo general se debe hacer en el momento de recoger la información. En cualquier caso, es fundamental contar con un diccionario de datos para el análisis y el posterior desarrollo del informe.

El archivo de datos se elabora siguiendo unas pautas sencillas. Los nombres de las variables se escriben en la primera fila y el identificador de la encuesta en la primera columna, seguido de las variables de identificación geográfica y las variables de interés particular del estudio. El archivo, en lo posible, debe estar en un formato fácilmente recuperable (extensiones de archivo .xls, .xlsx, .txt., .csv) y en ningún caso se deben utilizar archivos propios de programas específicos que no permitan consultas *a posteriori*.

Las siguientes se consideran malas prácticas en la generación de bases de datos:

1. Dejar celdas vacías en la primera fila.
 2. Consignar registros sin código de identificación.
 3. Combinar celdas.
 4. Incluir textos en las celdas de las variables donde solo deben ir números.
 5. En ausencia de un dato, diligenciar el campo con el valor cero (0).
9. *Análisis de los datos y elaboración de informes*: el análisis involucra una serie de cálculos basados en los datos obtenidos. La idea es que de este análisis salgan reportes de resultados presentados en forma sintética y en diferentes escalas (unidad muestral, estratos y universo). Esta etapa se puede llevar a cabo mediante análisis univariado, análisis bivariado y, en caso de que el instrumento lo permita, también se pueden aplicar técnicas de análisis multivariado.

Una vez analizados los datos, se generan los reportes de resultados, que no son más que un resumen de los datos de las unidades muestrales. Lo habitual es elaborar dos informes, uno relacionado con el diseño de la muestra y los resultados de las encuestas, y otro compuesto de un informe técnico que plasme las conclusiones del estudio.



La estimación del tamaño de muestra se puede hacer teniendo en cuenta diversos métodos estándar, para los cuales se requiere un conocimiento previo sobre la población de interés del estudio. Según Caballero Sánchez de Puerta (2016), en el momento de seleccionar una muestra, se deben garantizar dos condiciones esenciales:

1. *Representatividad*: la muestra debe reflejar las características de la población de estudio. Esto significa que se deben seleccionar aquellas variables que permiten describir apropiadamente la población. Si una muestra no es representativa, se dice entonces que está sesgada y que no es posible realizar inferencias a partir de los datos colectados.
2. *Tamaño*: depende del objetivo de la investigación y de la población de estudio y debe ser suficiente para garantizar la representatividad. El número mínimo depende del número de elementos que conforman la población, la heterogeneidad de las variables de interés, el nivel de confianza (probabilidad de que la estimación de la muestra se ajuste a la realidad) y el error máximo con el que se decide realizar el estudio (error de muestreo permitido).

Preparación y procesamiento de bases de datos

La preparación es la etapa en la cual los datos sin procesar se limpian, organizan y estructuran de forma tal que se facilite el procesamiento. En esta etapa, por lo general se requiere gran cantidad de tiempo, aunque en la mayoría de casos no se presenta el paso a paso desarrollado previamente, el cual incluye la

unificación de bases de datos de diferentes fuentes en estructuras interoperables y la definición del formato del archivo y el sitio de almacenamiento final. La etapa de procesamiento se refiere a la transformación de la base de datos para que permita obtener información que sirva de insumo en posteriores análisis. A continuación se relaciona el paso a paso del proceso para el uso de los datos del 3er CNA como fuente de información:

1. *Importación de archivos S06A(Cultivos) y S01_15(Unidad_productora)*: el primer paso en el análisis de una base de datos es el proceso de revisión, depuración y preparación de datos, con el fin de determinar si la base de datos se ajusta a las necesidades de la investigación. Los archivos base con los que se determinan los parámetros para estimar el tamaño de la muestra se identifican como Total_nacional(csv) y Diccionario de datos del CNA 2014, disponibles en el ANDA.

Del archivo comprimido Total nacional(csv), son de interés dos de las once bases de datos contenidas allí. Estos dos archivos son los siguientes: 1) S01_15(Unidad_productora).csv, el cual contiene, entre otras variables, las UPA, las Unidades Productoras No Agropecuarias (UPNA), el tipo de territorialidad (comunidades indígenas, negras, afrocolombianas, raizales y palenqueras), el uso del suelo, la naturaleza jurídica, el área de cada UPA, su uso agropecuario y la predominancia de la tenencia de la tierra, y 2) S06A(Cultivos).csv, el cual corresponde al inventario, las prácticas agrícolas y pecuarias de la UPA, una descripción básica de cultivos acuícolas y la sustentabilidad de la actividad agropecuaria, tanto en la gestión de los recursos naturales como en la gestión de la producción (DANE, 2014b).

El proceso de preparación consistió en seleccionar cuatro de las 16 variables que contiene el archivo S06A y hacer filtros de dos de las variables (Encuesta y P_S6P46) que se describen en el Anexo 1, esto con el objetivo de:

- a. Eliminar registros de la variable ENCUESTA cuando tomen el valor o código 999999999. Estos registros denotan áreas no municipalizadas o municipios que no están incluidos en su totalidad en la modalidad de rutas operativas del DANE. Corresponden a predios sin identificación predial y sin cédula catastral.
- b. Filtrar el archivo por el o los códigos del o los cultivos de interés (variable P_S6P46).

1. *Procesamiento del archivo S06A_Cultivos*: una vez preparadas las bases de datos, es necesario calcular para cada unidad productiva agropecuaria el total del área sembrada, total del área cosechada y el promedio del rendimiento del cultivo. El siguiente es el procedimiento que se debe seguir:
 - a. Para cada UPA, se deben sumar las áreas sembradas y cosechadas por cultivo de interés.
 - b. Para cada UPA, calcular el promedio de rendimiento por cultivo de interés.
2. *Unión de base de datos del cultivo S06A(Cultivos) e información general S01_15(Unidad_productora)*: se une la tabla de datos del área total sembrada y cosechada con el rendimiento promedio de cultivo de cada UPA y la tabla resultante se une con el archivo de encabezado, para poder identificar la predominancia étnica del territorio donde se ubica el predio.

La segunda fase de preparación se lleva a cabo con la tabla de datos resultante y consiste en seleccionar las UPA que cumplan las siguientes características:

- a. UPA cuya área sembrada sea igual o mayor al valor de referencia del cultivo de interés, es decir, que tenga el área mínima de cultivo para que su productor sea considerado usuario de semilla o de material de siembra o propagación.
- b. UPA presentes en territorios cuya predominancia étnica no esté bajo las categorías de comunidades indígenas, negras, afrocolombianas, raizales y palenqueras (tabla 1).

Tabla 1. Predominancia étnica

PRED_ETNICA	Predominancia étnica
1	Resguardo indígena
2	Asentamiento indígena
3	Parcialidad indígena
4	Territorio indígena diferente a los anteriores
5	Territorios colectivos de comunidades negras titulados
6	Territorios de ocupación colectiva de comunidades negras sin titulación
7	Ninguno de los anteriores
9	Territorio ancestral raizal

Fuente: DANE (2014a)

Es importante aclarar que el procedimiento anteriormente descrito puede emplearse para preparar y procesar bases con diferentes fines, entre ellos definir parámetros para estimar tamaño de muestra con otros métodos de muestreo.

Elaboración de estratos mediante la variable `Area_Sembrada_Sum`

La variable de estratificación es auxiliar. Permite identificar una o varias características de la población (Särndal et al., 2003) para construir grupos homogéneos de elementos. Una condición para construir estos grupos es que la variable de estratificación esté estrechamente relacionada con la variable de interés o con la información recopilada durante la ejecución de la encuesta (Lavallée & Hidirolou, 1988; Vivanco, 2005). En el presente trabajo, se empleó la variable de área sembrada para generar los estratos, variable que a su vez se relaciona con la variable de área cosechada, la cual se utilizó posteriormente para determinar las varianzas por estratos.

La definición de los estratos se realizó teniendo en cuenta los siguientes criterios:

- La varianza debe ser mínima dentro de cada estrato, es decir, los elementos que la conforman deben ser lo más homogéneos posibles.
- Cada estrato debe tener medias heterogéneas, lo cual se traduce en varianzas grandes entre las medias de los estratos.
- Los estratos deben ser mutuamente excluyentes y colectivamente exhaustivos, es decir, un elemento solo puede estar asignado a un estrato poblacional.
- El número de estratos no puede ser ni muy grande ni muy pequeño para que no se desvirtúen las ventajas de hacer el agrupamiento (Vivanco, 2005). Una alternativa objetiva para definir el número de estratos es la propuesta por Dalenius y Hodges (1959), que no se profundiza en este documento.

Una vez conformados los estratos, se generó una nueva variable (`ar_sem`) para identificarlos dentro de la base de datos. En el Anexo 2 se pueden consultar todos los códigos generados en el *software* estadístico SAS 9.4 para

importar, preparar y unir archivos, y para estratificar variables, seleccionar elementos de la población y calcular parámetros por estrato.

Selección de una proporción de la población para estimar parámetros por estratos

En cada uno de los estratos conformados, se tomó una muestra con un tamaño igual a un porcentaje dado mediante muestreo aleatorio simple. Para fines prácticos, en cada estrato se seleccionó el 10 % del total de elementos.

Determinar varianzas, desviación estándar y medias por estratos

La población se divide en subpoblaciones con el objetivo de ganar precisión, es decir, para que la variable de interés tenga un comportamiento más homogéneo dentro de cada estrato. El tamaño final de la muestra depende de las cuasivarianzas por estrato.

La varianza de una muestra es un valor estimado de la varianza verdadera y desconocida de la población. Para disminuir el error y estimar con mayor precisión el valor verdadero, se emplea un estimador denominado *cuasivarianza* (ecuación 13).

$$s_{hi}^2 = \frac{1}{n_{hi}-1} \sum_{i=1}^L (y_{hi} - \bar{y}_{hi})^2 \quad \text{Ecuación 13,}$$

donde s_{hi}^2 es la cuasivarianza verdadera en el h-ésimo estrato, n_{hi} el número de elementos de la muestra en el h-ésimo estrato, y_{hi} el valor observado en la i-ésima unidad perteneciente al h-ésimo estrato, y $\bar{y}_{hi}$ la media muestral por estrato.

En la mayoría de los programas estadísticos, la varianza se estima mediante la fórmula de la cuasivarianza sin hacer diferencia entre las dos, es decir, varianza y cuasivarianza se usan de forma indistinta sin ningún criterio previamente definido (Guisande González et al., 2006). Esta es la razón por la cual en el código se solicita la varianza y la desviación estándar de los datos,

ya que el *software* SAS 9.4 calcula por defecto la cuasivarianza de los datos observados. Además, en el código se incluyó el cálculo del valor promedio simple de las variables para cada estrato.

Después de obtener estos tres parámetros, se calcula el tamaño de la muestra mediante las fórmulas presentadas en el capítulo III. Existen otras versiones, que varían según los diferentes autores, como la presentadas en los trabajos de Kish (1965), Cochran (1977) y Scheaffer et al. (1987, 2007), y todas ellas están bien documentadas en la extensa bibliografía sobre métodos de muestreo.







En los procesos de investigación existe una marcada tendencia a simplificar conceptos y procedimientos para calcular el tamaño de la muestra, pero este paso no es sencillo ni puede ser tomado a la ligera (Kerlinger & Lee, 2002). Algunos autores prefieren emplear métodos intuitivos, sin ningún soporte estadístico, como cuando estiman el tamaño de la muestra asignando un valor proporcional al total de la población o simplemente eligiendo un valor muy grande (Kerlinger & Lee, 2002; Valdivieso Taborga et al., 2011). Generalizan al momento de hacer inferencias o extrapolar y sacar conclusiones sobre la población de la cual se ha extraído la muestra (Otzen & Manterola, 2017), y con esto limitan la posibilidad de hacer una descripción más aproximada de la realidad de la población (Perfetti et al., 2012).

Una práctica frecuente en los documentos científicos consiste en decir que se empleó un determinado tamaño de muestra, pero sin detallar la metodología empleada para el cálculo o la descripción de las variables de estudio y su dispersión, y sin aclarar si se emplearon los parámetros de la variable con mayor coeficiente de variación. Otro aspecto relevante que se omite es el valor de la varianza, el cual se desconoce la mayoría de las veces. Existen alternativas que permiten superar este inconveniente, como usar los valores reportados en la literatura, hacer una prueba piloto o emplear un valor estimado por expertos en el tema (García-García et al., 2013).

Por otro lado, es fundamental decidir el diseño de recolección de datos, ya que cada uno tiene ventajas y desventajas. El diseño garantiza la validez estadística, el control y la precisión de las estimaciones y los costos de recolección de los datos (FAO, 2016a). Existen diversos métodos de muestreo para recopilar información agropecuaria, entre los cuales se encuentran el aleatorio simple,

el sistemático, el estratificado y el muestreo por conglomerados. Los métodos que realizan agrupamientos de individuos son los más empleados cuando no existen marcos de lista, de área o múltiples que permitan identificar a los productores (FAO, 2018). Las investigaciones sobre el sector agropecuario en Colombia suelen llevarse a cabo con estas carencias. Básicamente deben enfrentarse a las mismas dificultades para acceder a los marcos de lista y los datos históricos. Se desconocen los criterios empleados para el levantamiento de la información y falta representatividad estadística en los estudios previos (Perfetti et al., 2012).

Durante la ejecución del proyecto “Caracterización de oferta y demanda de semilla de calidad de las especies de interés”, desarrollado por AGROSAVIA entre 2015 y 2017, se identificaron las mismas limitantes descritas por Perfetti et al. (2012). Fue necesario, por tanto, buscar alternativas que permitieran acceder a un marco muestral de referencia en el que los datos estuvieran disponibles y abiertos y a los que fuera posible acceder mediante los diferentes convenios que ha suscrito la corporación con otras entidades de investigación relacionadas con el sector agropecuario.

Después de varias consultas, se identificó que los datos recolectados y presentados en el 3^{er} CNA eran la mejor opción. Cualquier investigador puede emplear estos datos como muestra piloto, debido principalmente a que la información se presenta en el nivel de las UPA y debido también a que los datos se colectaron en todo el territorio nacional y la organización de los registros permite identificar las actividades agropecuarias en los niveles de vereda, municipio y departamento. El acceso a la información, así como su disponibilidad en el tiempo, se ratificó durante la ejecución del proyecto, y el censo se empleó para determinar el tamaño de muestra de las tres especies de interés: papa, cacao y plátano.

La implementación de este tipo de metodologías permite a los diferentes actores tener una base de información confiable, para conocer el comportamiento de las poblaciones y apoyar la toma de decisiones bajo diferentes escenarios. Lo anterior se refleja en un conocimiento de las comunidades y los sistemas productivos y posibilita generar estrategias para adoptar tecnologías como semillas de calidad y nuevos materiales y técnicas de manejo en la producción, que apoyen el desarrollo económico de los productores agropecuarios en el corto, mediano y largo plazo.

La metodología permite definir tamaños de muestra para una determinada especie de interés en una zona de estudio. Además, sirve de insumo para que cada entidad pueda realizar planes de trabajo dirigidos a recolectar información primaria que permita dar respuesta a las problemáticas del sector agropecuario que se derivan de la deficiencia de información estadística confiable. El proyecto contempló varias etapas, entre las cuales se encuentran la priorización de zonas de estudio, el cálculo del tamaño de la muestra y la selección del número de UPA. A continuación, se detalla cada una de las etapas.

Criterios para la priorización de zonas de estudio

Las zonas productoras se seleccionaron y priorizaron empleando criterios cuantitativos y cualitativos. Algunos de tales criterios fueron área total sembrada del cultivo de interés, presencia del eslabón primario de la cadena en la zona y variedades sembradas. También se tuvo en cuenta la disponibilidad de información para determinar zonas agroecológicas homogéneas situadas en la zona de influencia de AGROSAVIA. Se seleccionaron los municipios de Une y Villapinzón (Cundinamarca) para el sistema productivo de papa, San Pablo de Borbur (Boyacá) para cacao y Fuentedeoro (Meta) para plátano. En la tabla 2 se resumen los criterios más importantes empleados en cada sitio.

Tabla 2. Criterios de priorización de zonas para validar la metodología

Sistema productivo	Municipio	Criterios de priorización
Papa	Villapinzón, Cundinamarca	Es el segundo municipio de Cundinamarca en área sembrada (6.549 ha).*
		Existen diferentes tipos de productores, semilleristas y viveristas (con y sin registro del Instituto Colombiano Agropecuario [ICA]), así como centros de acopio de la industria, comercializadores regionales y centrales de abastos.
		AGROSAVIA trabaja en la zona en temas relacionados con semilla; algunas organizaciones de productores están inscritas en el Plan Semilla.

(Continúa)

(Continuación tabla 2)

Sistema productivo	Municipio	Criterios de priorización
Papa	Une, Cundinamarca	Es el quinto municipio de Cundinamarca en área sembrada con papa (3.174 ha).*
		Cuenta con diferentes tamaños de productores; tradicionalmente ha sido una zona proveedora de semilla sin registro ICA.
		El nivel de intervención de AGROSAVIA ha sido bajo.
		Fue una zona con graves problemas de orden público.
Cacao	San Pablo de Borbur, Boyacá	El occidente de Boyacá tuvo en promedio 3.937 ha sembradas de cacao, con una participación de 2 % en el área sembrada nacional.*
		Es un área nueva en la producción de cacao, con alto potencial de crecimiento.
		Tiene un área sembrada de 520 ha y es el cuarto productor de cacao en Boyacá.
		Hay productores de diferentes tamaños, proveedores de material de siembra y propagación con y sin registro ICA.
		AGROSAVIA trabaja en la zona desde hace varios años en temas de semilla; existen organizaciones inscritas en el Plan Semilla.
Plátano	Fuentedeoro, Meta	Es el primer municipio productor de plátano (5.566 ha) en el departamento del Meta.*
		El municipio forma parte de un gran corredor de producción de plátano.
		Existen productores de diferentes tamaños y tecnologías de producción.
		AGROSAVIA trabaja en la zona desde hace varios años en temas de semilla; algunas organizaciones de productores están inscritas en el Plan Semilla.

* Datos tomados de *Evaluaciones Agropecuarias Municipales (EVA)*, 2007-2014 (Red de información y comunicación del sector Agropecuario Colombiano [Agronet], 2014).

Fuente: Elaboración propia

Cálculo para determinar el tamaño de muestra y estimar el promedio de la demanda de semilla por unidad de área

Una vez definidas las zonas, se procedió a determinar los parámetros necesarios para definir el tamaño de la muestra. En el sitio Archivo Nacional de Datos del DANE (<http://andacna.dane.gov.co/index.php/home>), se descargaron los registros del archivo S06A(Cultivos) del 3er CNA y el Diccionario de datos. Dado que cada productor puede estar más de una vez en la base de datos porque los registros por filas hacen referencia a cada uno de los lotes presentes en el predio, se sumaron todos los registros por encuesta de las variables de área sembrada y área cosechada y se calculó el promedio de los rendimientos por cultivo. De este modo se creó una nueva tabla de datos para las variables calculadas Area_Sembrada_Sum, Area_Cosechada_Sum y Rendimiento_Promedio. Los datos se prepararon tal como se describió en los capítulos anteriores. Se filtró la base por identificadores de la encuesta (municipio y cultivo), empleando los códigos que se presentan en la figura 5.

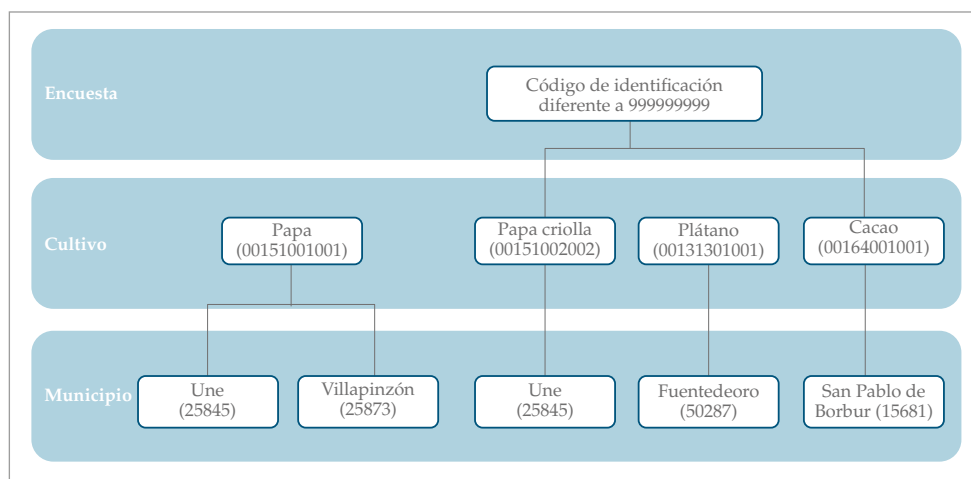


Figura 5. Códigos empleados en los filtros, en la preparación y procesamiento del archivo S06A(Cultivos).

Fuente: Elaboración propia

Con el fin de establecer un área mínima de cultivo que permitiera considerar si un productor demanda semilla para el establecimiento del cultivo, el equipo de trabajo se reunió con expertos de cada uno de los tres sistemas productivos. Con ellos se definieron los valores mínimos que se emplearon como valores de referencia (tabla 3).

Tabla 3. Área mínima de referencia

Cultivo	Área sembrada mínima (ha)
Papa	0,05
Cacao	0,5
Plátano	0,5

Fuente: Elaboración propia

El archivo se filtró eliminando los registros con áreas inferiores al valor de referencia. También se examinaron los datos para determinar si alguna UPA se ubicaba dentro de comunidades indígenas, negras, afrocolombianas, raizales y palenqueras, y se encontró que ninguna formaba parte de estos territorios especiales. En total, para las tres especies en las cuatro zonas priorizadas, se obtuvieron 2.107 registros únicos (tabla 4). A continuación, se crearon estratos por medio de la variable *Area_Sembrada_Sum*, se tomó de forma aleatoria aproximadamente el 10 % de los registros por estrato y sistema productivo, y estos registros se emplearon como sustituto de la muestra piloto de los productores.

Tabla 4. Población por especie y municipio

Departamento	Municipio	Especie	Número de Unidades Productoras Agropecuarias (UPA)
Boyacá	San Pablo de Borbur	Cacao	407
Meta	Fuentedeoro	Plátano	544
Cundinamarca	Une	Papa	103
	Villapinzón	Papa	1.053

Fuente: Elaboración propia con base en DANE (2016b)

Con esa porción de registros se calcularon la varianza, la desviación estándar y la media. Estos valores se emplearon en la fórmula que determina el tamaño de muestra por afijación óptima en el método de muestreo estratificado, con un error relativo de estimación del 5 % y un nivel de confianza del 90 %. Todos los procedimientos descritos se hicieron con el *software* estadístico SAS 9.4. En la figura 6 se presenta un esquema general de los pasos seguidos en la preparación y procesamiento de la base de datos.

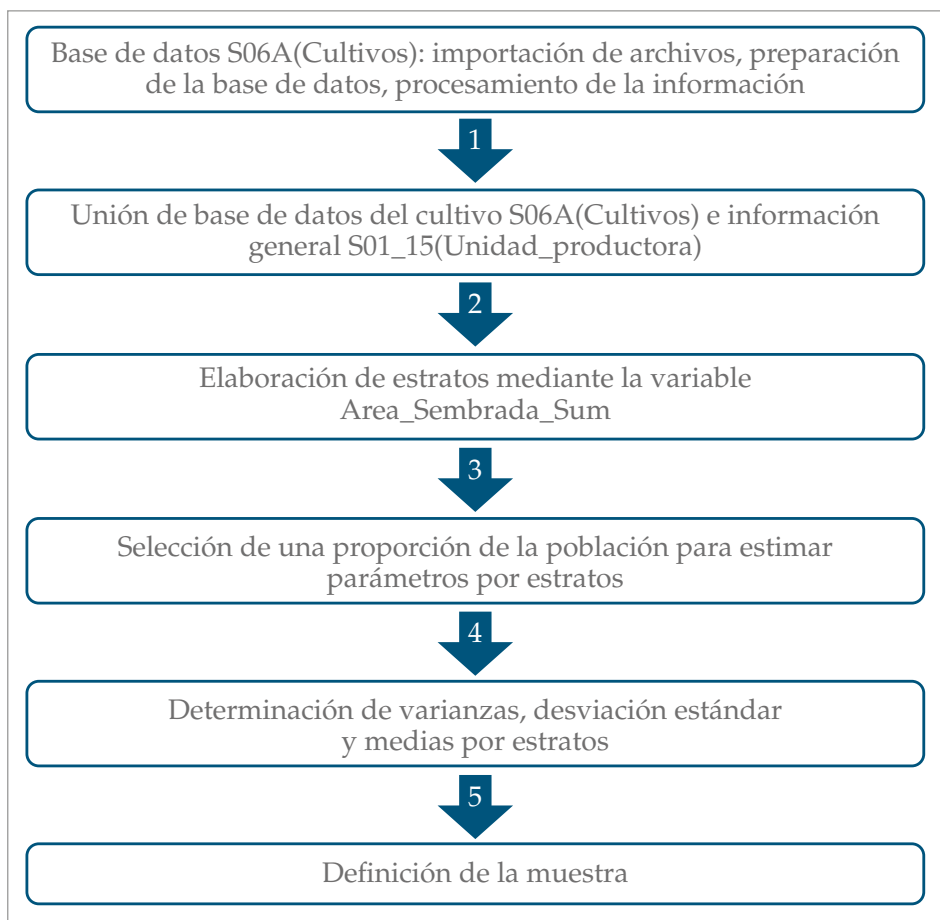


Figura 6. Proceso para determinar los parámetros para el cálculo del tamaño de la muestra.
Fuente: Elaboración propia

Los tamaños óptimos de la muestra por sistema productivo fueron los siguientes: 134 UPA para papa en los municipios de Une (20) y Villapinzón (114); 45 UPA para cacao en el municipio de San Pablo de Borbur, y 55 UPA para plátano en el municipio de Fuentedeoro (tabla 5).

Tabla 5. Distribución de la muestra a nivel UPA para tres sistemas productivos

Sistema productivo	Estratos	Tamaño de muestra	Número de veredas
Papa (Une y Villapinzón)	0,05-1,0	10	7
	1,01-2,0	8	6
	2,01-5,0	50	15
	5,01-10,0	31	11
	10,01-50,0	26	16
	50,01 y más	9	7
Cacao (San Pablo de Borbur)	0,5-1,0	9	9
	1,01-2,0	17	13
	2,01-3,0	7	5
	3,01 y más	12	10
Plátano (Fuentedeoro)	0,5-0,99	2	2
	1-1,99	1	1
	2-4,99	8	8
	5-9,99	12	9
	10-19,99	8	8
	20-39,99	11	6
	40-59,99	2	2
	60-79,99	6	6
	80,00 y más	5	4

Fuente: Elaboración propia con base en DANE (2016b)

En cada municipio se realizó una consulta con la respectiva secretaría de agricultura, la Umata, y asociaciones y cooperativas, para elaborar los marcos de lista que requería la selección de la muestra. Las listas contenían información sobre nombres de productores, datos de contacto, ubicación geográfica y tamaño de predios. Dados los aspectos socioculturales de los productores y la dinámica de producción de cada sistema, y ante posibles

rotaciones de cultivo o cambios en los lotes de producción, cada listado se verificó para determinar la coherencia de la información reportada por las diferentes entidades y la presencia de las unidades productivas en los sitios de interés, así como para identificar de manera definitiva los individuos muestrales a quienes se aplicaría la encuesta.

Estimación de número de UPA y área sembrada en otros cultivos

Con la metodología descrita en este documento, se estimó el número de unidades productoras y la cantidad de área sembrada y cosechada de las especies algodón, chontaduro, frijol, lima, limón, maíz amarillo, maíz blanco, mango, naranja y quinua. En estas especies, además de que se usaron para determinar sitios de cultivo en 2014 y número de UPA, permitieron a los colaboradores de AGROSAVIA desarrollar muestreos de material vegetal y suelos, así como caracterizaciones de productores y sistemas productivos. En la tabla 6 se presentan las especies y el uso que se le dio a la información procesada en el 3^{er} CNA.

Tabla 6. Aplicación de la metodología en otras especies

Especie	Usos
Algodón	Número de UPA; caracterización de productores
Chontaduro	Número de UPA
Frijol	Caracterización de productores
Lima	Número de UPA
Limón	Número de UPA para muestreo de material vegetal y suelos
Maíz amarillo	Caracterización de productores
Maíz blanco	Caracterización de productores
Mango	Número de UPA para muestreo de material vegetal y suelos; caracterización de productores
Naranja	Número de UPA
Quinua	Número de UPA

Fuente: Elaboración propia

La metodología descrita en el presente documento y el uso de preguntas a manera de filtros permitieron calcular los tamaños de muestra y estimar la media y el total poblacional de las variables de interés de los diferentes estudios. Con la variable de cultivos (Pregunta P_S6P46) se identificó la distribución de las especies cultivadas a nivel de vereda en los territorios de interés, se determinó si estaban presentes en territorios especiales y se hizo una aproximación a la caracterización de los sistemas productivos. La metodología también se adaptó a otros archivos del 3^{er} CNA disponibles en el sitio Datos abiertos, para identificar características personales y de vivienda de los productores agropecuarios del país.

Sobre el trabajo de campo

La base de datos del 3^{er} CNA permitió definir la muestra en los niveles de municipio y vereda y delimitar el tamaño con las UPA, las cuales se usaron como variable de estratificación. Hubo, sin embargo, un alto nivel de incertidumbre dada la dinámica de la producción agrícola, ya que en ella se dan movimientos de productores de una zona a otra, reducciones o aumentos de áreas de siembra, abandono de los productores de la actividad y aparición o reaparición de otros productores. En consecuencia, esta muestra con distribución de estratos por vereda debió ajustarse a la realidad actual del cultivo en cada municipio, y para ello fue necesario, tal como lo formula Díaz de Rada Igúzquiza (2008), contar con una fuente que entregara información detallada del dominio objeto de estudio. Esto representaba un desafío en aquellos casos en los que las muestras se habían construido con reportes de población de los censos, dado que estos últimos no registran en detalle la estructura de territorios con alta movilidad debido a la baja frecuencia con que se actualizan sus datos.

Para subsanar lo anterior, se acudió a las Umata, las cuales son el principal actor de la estadística y la organización productiva de los municipios, así como a organizaciones de productores y otros actores clave con información sobre productores, unidades productoras, cultivos, veredas y ubicación geográfica de los predios. La información suministrada por estas entidades permitió verificar si las UPA de cada vereda cumplían con el criterio de área sembrada (estratos), y con ello se construyeron los marcos de lista de los productores que conformaban las unidades estadísticas (los individuos de la población

estudiada). Así se pudo reconfigurar y ajustar la muestra por estrato a nivel de vereda en cada municipio, esto bajo el supuesto de que este conocimiento previo mejoraría la efectividad en la selección de los elementos muestrales. Al comprobar previamente cuáles individuos cumplían los requisitos definidos, se pudieron identificar los candidatos que formarían parte de la muestra.

El trabajo de campo inició con la planeación de rutas para acceder a los elementos muestrales definidos en cada estrato. En este proceso se detallaron los traslados y movimientos de los encuestadores, con el fin de optimizar el tiempo de desplazamiento, disminuir la fatiga física, reducir gastos, cubrir la totalidad de zonas de cada estrato y facilitar la localización tanto de los encuestados como de los encuestadores.

Con los datos de contacto, se estableció comunicación con los productores de las veredas que conformaban la muestra para concertar una cita en el predio y aplicar la encuesta. Quienes rechazaron la invitación a participar en el estudio fueron reemplazados por otros de la misma vereda que tenían cultivos con las mismas características de área. Cuando no fue posible encontrar sustituto, se buscó su reemplazo en otra vereda cercana. En estos últimos casos, se buscó el acompañamiento de quien los había referenciado, para tener mayor posibilidad de ser atendidos, o se visitaron los predios sin previo aviso. En estas visitas se presentó el proyecto y se solicitó participación en el estudio luego de hacer preguntas filtro para asegurar que los nuevos candidatos cumplieran con el perfil. De esta manera se completaron las rutas y se cumplió con el número total de encuestados que demandaba cada estrato.

Durante la ejecución del estudio, se evaluó diariamente el cumplimiento de la meta de encuestas del estrato de cada ruta. Las rutas debían repetirse cuando los productores reagendaban la cita o cuando, durante el recorrido, no se obtenía la respuesta esperada. En este último caso, se debió consultar de nuevo a los actores clave para indagar por nuevos productores que tuvieran el perfil, concertar una cita con ellos y seguir con el proceso hasta completar el tamaño del respectivo estrato. Si, pese al esfuerzo, no se lograba completar el número de encuestas esperadas para cada estrato, se discutía la pertinencia de encuestar productores con cultivos correspondientes al perfil de otros estratos. Se aprovechaba así la presencia en el territorio y se consideraba la pérdida de información como parte del proceso de aplicación de la encuesta.



Referencias

- Abascal Fernández, E., & Grande Esteban, I. (2005). *Análisis de encuestas*. Esic Editorial.
- Acevedo González, G., & Palacio Martínez, J. A. (2016). Evaluación de factores que afectan la comercialización de agro-productos de pequeños y medianos productores del oriente antioqueño. *Journal of Agriculture and Animal Sciences*, 5(2), 60-69. <https://doi.org/10.22507/jals.v5n2a5>
- Berenson, M. L., & Levine, D. M. (1996). *Estadística básica en administración: conceptos y aplicaciones*. Pearson Educación.
- Bonilla Sebá, E. C. (2018). Bajos impuestos al suelo e inoperancia del instrumento tributario. *Revista de Economía Institucional*, 20(39), 211-229. <https://doi.org/10.18601/01245996.v20n39.09>
- Caballero Sánchez de Puerta, P. (2016). *Investigación y recogida de información de mercados*. Paraninfo.
- Cochran, W. G. (1977). *Sampling techniques* (3 ed.). John Wiley & Sons.
- Dalenius, T., & Hodges, J. L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54(285), 88-101. <https://doi.org/10.2307/2282141>
- Departamento Administrativo Nacional de Estadística [DANE]. (2014a). *Colombia - Tercer Censo Nacional Agropecuario - 2014 - 3er CNA: diccionario de datos del CNA 2014*. http://andacna.dane.gov.co/index.php/catalog/669/related_materials
- Departamento Administrativo Nacional de Estadística [DANE]. (2014b). *Colombia - Tercer Censo Nacional Agropecuario - 2014 - 3er CNA: guía de diligenciamiento*. http://andacna.dane.gov.co/index.php/catalog/669/related_materials

- Departamento Administrativo Nacional de Estadística [DANE]. (2016a). *Metodología general: Tercer Censo Nacional Agropecuario (3^{er} CNA)*. https://www.dane.gov.co/files/investigaciones/fichas/agropecuario/metodologia_CNA-01_V1.pdf
- Departamento Administrativo Nacional de Estadística [DANE]. (2016b). *Tercer Censo Nacional Agropecuario: hay campo para todos* (Tomo 2). <https://www.dane.gov.co/files/images/foros/foro-de-entrega-de-resultados-y-cierre-3-censo-nacional-agropecuario/CNATomo2-Resultados.pdf>
- Díaz de Rada Igúzquiza, V. D. (2008). La selección de los entrevistados últimos en encuestas presenciales: un análisis de la utilización conjunta del método de rutas y el método de cuotas. *Revista Española de Investigaciones Sociológicas (Reis)*, (123), 209-247. <http://www.unavarra.es/personal/vidaldiaz/pdf/entrevistados08.pdf>
- Díaz de Rada Igúzquiza, V. D. (2009). *Análisis de datos de encuestas: desarrollo de una investigación completa utilizando SPSS*. Editorial UOC.
- García-García, J. A., Reding-Bernal, A., & López-Alvarenga, J. C. (2013). Cálculo del tamaño de la muestra en investigación en educación médica. *Investigación en Educación Médica*, 2(8), 217-224. [https://doi.org/10.1016/S2007-5057\(13\)72715-7](https://doi.org/10.1016/S2007-5057(13)72715-7)
- Guisande González, C., Barreiro Felpeto, A. B., Maneiro Estraviz, I., Riveiro Alarcón, I., Vergara Castaño, A. R., & Vaamonde Liste, A. (2006). *Tratamiento de datos*. Ediciones Díaz de Santos. <https://issuu.com/maiquim.floresm/docs/269990057-tratamiento-de-datos-esta>
- Gutiérrez López, J. A. (2009). La estadística estratégica del sector agropecuario en Colombia: un nuevo modelo de oferta. *Revista de la Información Básica*, 3(2). https://sitios.dane.gov.co/revista_ib/html_r6/articulo3_r6.html
- Kalmanovitz, S., & López López, E. (2006). *La agricultura colombiana en el siglo XX*. Fondo de Cultura Económica USA.
- Kerlinger, F., & Lee, H. (2002). *Investigación del comportamiento*. McGraw Hill.
- Kish, L. (1965). *Survey sampling* (1^a ed.). John Wiley & Sons.
- Lavallée, P., & Hidiroglou, M. A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14(1), 33-43. https://www.researchgate.net/publication/261873319_On_the_Stratification_of_Skewed_Populations
- Martínez Medrano, C. (2004). *Técnicas e instrumentos de recogida y análisis de datos*. UNED.

- Organización de las Naciones Unidas para la Alimentación y la Agricultura [FAO]. (2003). *Guía para encuestas de demanda, oferta y abastecimiento de combustible de madera*.
- Organización de las Naciones Unidas para la Alimentación y la Agricultura [FAO]. (2016a). *Manual de estadísticas sobre costos de producción agrícola*. <http://www.fao.org/3/ca6411es/ca6411es.pdf>
- Organización de las Naciones Unidas para la Alimentación y la Agricultura [FAO]. (2016b). *Programa mundial del censo agropecuario 2020: Vol. 1. Programa, definiciones y conceptos*. <http://www.fao.org/3/i4913s/i4913s.pdf>
- Organización de las Naciones Unidas para la Alimentación y la Agricultura [FAO]. (2018). *Programa mundial del censo agropecuario 2020: Vol. 2. Directrices operativas*. <http://www.fao.org/3/CA1963ES/ca1963es.pdf>
- Otzen, T., & Manterola, C. (2017). Técnicas de muestreo sobre una población a estudio. *International Journal of Morphology*, 35(1), 227-232. <http://dx.doi.org/10.4067/S0717-95022017000100037>
- Perfetti, J. J., Escobar, D., Castro, F., Cuervo, B., Rodríguez, M., & Vargas, J. (2012). *Costos de producción de doce productos agropecuarios*. Centro de Investigación Económica y Social (Fedesarrollo); IQuartil. <http://hdl.handle.net/11445/378>
- Red de información y comunicación del sector Agropecuario Colombiano [Agronet]. (2014). *Evaluaciones Agropecuarias Municipales EVA 2007-2014*. <https://www.agronet.gov.co/estadistica/paginas/home.aspx?cod=59>
- Rositas Martínez, J. (2014). Los tamaños de las muestras en encuestas de las ciencias. *Innovaciones de Negocios*, 11(22), 235-268. <http://eprints.uanl.mx/12605/>
- Sánchez, A., Solanilla, L., Clavijo, J., & Zambrano, A. (2010). Afijación óptima de tamaños de muestra en muestreo aleatorio estratificado vía programación matemática. *Comunicaciones en Estadística*, 3(1), 7-23. <https://doi.org/10.15332/s2027-3355.2010.0001.01>
- Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1987). *Elementos de muestreo* (G. R. Sánchez & J. R. Gómez Aguilar, Trans., 3ª ed.). Grupo Editorial Iberoamericana.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (2007). *Elementos de muestreo* (I. Sánchez García, Trad., 6ª ed.). Thomson Editores.

- Soto Agudelo, A. (2016). *La deficiencia de información estadística para el sector agrario en Colombia* [Tesis de maestría, Universidad de los Andes]. Repositorio Uniandes. <https://repositorio.uniandes.edu.co/bitstream/handle/1992/13853/u729557.pdf?sequence=1&isAllowed=y>
- Valdivieso Taborga, C. E., Valdivieso Taborga, O. Á, & Valdivieso Castellón, R. (2011). Determinación del tamaño muestral mediante el uso de árboles de decisión. *Investigación & Desarrollo* (0311), 148-176. <http://www.upb.edu/RePEc/iad/wpaper/0311.pdf>
- Vivanco, M. (2005). *Muestreo estadístico: diseño y aplicaciones*. Editorial Universitaria.
- Wallerstein, I. (1999). The heritage of sociology, the promise of social science. In *XIVth World Congress of Sociology, Montreal, 26 July 1998* (pp. 1-37). *Current Sociolog.* <https://doi.org/10.1177/0011392199047001002>
- Zapata-Ossa, H. J., Cubides-Munévar, Á. M., López, M. C., Pinzón-Gómez, E. M., Filigrana-Villegas, P. A., & Cassiani-Miranda, C. A. (2011). Muestreo por conglomerados en encuestas poblacionales. *Revista de Salud Pública*, 13(1), 142-151. <https://revistas.unal.edu.co/index.php/revsaludpublica/article/view/33543/33506>





Anexos

Anexo 1. Diseño del registro y descripción de las variables

Variable	Atributos				
	Tabla de datos	Tipo	Longitud	Descripción	Tabla de referencia (valores)
Encuesta	ALL-LLAVE	Alfanumérico	9	Código de encuesta	
P_S6P46	S06A(Cultivos)	Alfanumérico	11	¿Cuál cultivo o plantación forestal tiene en el lote?	Tabla de referencia 13*
AREA_SEMBRADA	S06A(Cultivos)	Numérico		Área sembrada	Hectáreas
AREA_COSECHADA	S06A(Cultivos)	Numérico		Área cosechada	Hectáreas
P_S6P59_UNIF	S06A(Cultivos)	Numérico		Rendimiento (t/ha)	
P_DEPTO	ALL-LLAVE	Numérico	2	Código de departamento	Tabla de referencia 2*
P_MUNIC	ALL-LLAVE	Numérico	2	Código de municipio	Tabla de referencia 3*
COD_VEREDA	ALL-LLAVE	Numérico	8	Código de vereda	Tabla de referencia 4*
PRED_ETNICA	S01_15(Unidad_productora)	Numérico	1	Predominancia étnica de la unidad	Tabla de referencia 6*

* Archivo TEMATICA_DISENO DE REGISTRO CNA2014.xlsx, hoja TABLAS_REFERENCIA

Fuente: Elaboración propia

Anexo 2. Código en SAS 9.4

El signo de equivalencia en los códigos (~) hace referencia a la ruta o las carpetas donde se ha almacenado el archivo en la computadora.

*/*Importar archivos*/*

a. Encabezado

```
PROC IMPORT OUT= WORK.Encabezado  
DATAFILE="~\S01_15(Unidad_productora).csv"  
DBMS=csv REPLACE;  
RUN;
```

b. S06A_Cultivos

```
PROC IMPORT OUT= WORK.S06A_Cultivos  
DATAFILE="~ \S06A(Cultivos).csv"  
DBMS=csv REPLACE;  
RUN;
```

*/*Preparación del archivo*/*

Código en SAS

```
PROC SQL;  
CREATE TABLE WORK.Preparacion AS  
SELECT t1.P_DEPTO, t1.P_MUNIC, t1.ENCUESTA, t1.COD_VEREDA, t1.P_  
S6P46, t1.P_S6P59_UNIF, t1.AREA_SEMBRADA, t1.AREA_COSECHADA  
FROM WORK.S06A_CULTIVOS t1  
WHERE t1.ENCUESTA NOT = 99999999 AND t1.P_S6P46 = /*Código del  
cultivo de interés*/;  
QUIT;
```

*/*Procesamiento del archivo S06A_Cultivos*/*

```
PROC TABULATE  
DATA= WORK.Preparacion OUT= WORK.Tabla_Resumen  
(LABEL="Tablas de sumalización para WORK.Preparacion") ;  
VAR P_S6P59_UNIF AREA_SEMBRADA AREA_COSECHADA;  
CLASS P_DEPTO / ORDER=UNFORMATTED MISSING;  
CLASS P_MUNIC / ORDER=UNFORMATTED MISSING;  
CLASS P_S6P46 / ORDER=UNFORMATTED MISSING;  
CLASS ENCUESTA / ORDER=UNFORMATTED MISSING;  
CLASS COD_VEREDA / ORDER=UNFORMATTED MISSING;
```

TABLE

/* Dimensión de la fila */

P_DEPTO* P_MUNIC* COD_VEREDA* ENCUESTA* P_S6P46,

/* Dimensión de la columna */

P_S6P59_UNIF* Mean AREA_SEMBRADA* Sum AREA_COSECHADA*

Sum;

RUN;

/*Unión de base de datos Cultivos y Encabezado*/

PROC SQL;

CREATE TABLE WORK.Combinacion_Cultivos_Encabezado **AS**

SELECT t1.P_DEPTO, t1.P_MUNIC, t1.COD_VEREDA, t1.ENCUESTA,
t2.PRED_ETNICA, t1.P_S6P46, t1.P_S6P59_UNIF_Mean, t1.AREA_
SEMBRADA_Sum, t1.AREA_COSECHADA_Sum

FROM WORK.Tabla_Resumen t1, WORK.Encabezado t2

WHERE (t1.P_DEPTO = t2.P_DEPTO **AND** t1.P_MUNIC = t2.P_MUNIC
AND t1.ENCUESTA = t2.ENCUESTA **AND** t1.COD_VEREDA =
t2.COD_VEREDA);

QUIT;

/*Unión de base de datos Cultivos_Encabezado y predominancia étnica*/

PROC SQL;

CREATE TABLE WORK.Datos_Cultivos_Final **AS**

SELECT t1.P_DEPTO, t1.P_MUNIC, t1.COD_VEREDA, t1.ENCUESTA,
t1.PRED_ETNICA, t1.P_S6P46, t1.P_S6P59_UNIF_Mean, t1.AREA_
SEMBRADA_Sum, t1.AREA_COSECHADA_Sum

FROM WORK.Combinacion_Cultivos_Encabezado t1

WHERE t1.AREA_SEMBRADA_Sum >= /* Valor de referencia */;

QUIT

/*Generación de estratos*/

DATA WORK.Datos_Cultivos_Final_Estratos; **SET** WORK.

Datos_Cultivos_Final;

IF AREA_SEMBRADA_Sum <= /*límite superior estrato*/ **THEN**

ar_sem="Est_1";

ELSE IF AREA_SEMBRADA_Sum <= /*límite superior estrato*/ **THEN**

ar_sem="Est_2";

ELSE IF AREA_SEMBRADA_Sum <= /*límite superior estrato*/ **THEN**

ar_sem="Est_3";

```

ELSE IF AREA_SEMBRADA_Sum <= /*límite superior estrato*/ THEN
ar_sem="Est_4";
ELSE IF AREA_SEMBRADA_Sum <= /*límite superior estrato*/ THEN
ar_sem="Est_5";
ELSE IF AREA_SEMBRADA_Sum <= /*límite superior estrato*/ THEN
ar_sem="Est_6";
ELSE ar_sem="Est_7";
RUN;

/*Selección de una proporción de la población*/
PROC SORT DATA= WORK.DATOS_CULTIVOS_FINAL_ESTRATOS;
BY ar_sem;
RUN;
/*Contar el número de datos por estrato (Frecuencia por estrato)*/
PROC FREQ DATA= WORK.DATOS_CULTIVOS_FINAL_ESTRATOS
NOPRINT;
TABLE ar_sem /MISSING OUT=Estratos_area_sembrada (DROP= percent
RENAME=(count=_nsize_));
RUN;
/*Cálculo del 10% del número de datos por estrato*/
DATA ESTRATOS_AREAS;
SET ESTRATOS_AREA_SEMBRADA;
_nsize_ = ceil(_nsize_*0.10);
RUN;
/*Seleccionar aleatoriamente n datos por estrato*/
PROC SURVEYSELECT DATA= WORK.
DATOS_CULTIVOS_FINAL_ESTRATOS
    SAMPSIZE= ESTRATOS_AREAS OUT=SAMPLE_ESTRATOS;
    STRATA ar_sem;
RUN;
/*Cálculo de Media, Desviación estándar y Varianza por estrato*/
PROC TABULATE
DATA= WORK.SAMPLE_PLATANO;
VAR AREA_COSECHADA_Sum;
CLASS ar_sem / ORDER=UNFORMATTED MISSING;
TABLE /* Dimensión de la fila */
ar_sem,/* Dimensión de la columna */
AREA_COSECHADA_Sum*( Mean StdDev Var) ;
RUN;

```



Terminó de diseñarse en abril de 2021,
Bogotá, D. C., Colombia

Una de las principales limitantes en la realización de estudios socioeconómicos por encuestas en el sector agropecuario es la ausencia de marcos muestrales actualizados y suficiente información con respaldo estadístico. El primer reto por superar es la definición de un diseño muestral que permita elegir una muestra representativa.

En este libro se presentan conceptos generales sobre métodos de muestreo, una propuesta metodológica para el uso de los datos del Tercer Censo Nacional Agropecuario (CNA), así como un ejemplo de aplicación para definir el marco muestral y la muestra, los cuales emplean los registros de las unidades productoras agropecuarias (UPA).

A su vez, está dirigida a personas y entidades que, en sus dinámicas productivas, de mercado y política pública requieren realizar estudios por encuestas y recolectar y analizar datos para la toma de decisiones, convirtiéndose así en una guía práctica para la correcta aplicación de metodologías de muestreo en el sector agropecuario del país.



BIBLIOTECA AGROPECUARIA DE COLOMBIA

CORREO: **bac@agrosavia.co**

TELÉFONO: **(57 1) 422 73 00** EXT. 1257 o 1274

SKYPE: **[biblioteca.agropecuaria](https://www.skype.com/name/identifiers/alias/biblioteca.agropecuaria)**

www.agrosavia.co

Distribución gratuita
Prohibida su venta



**El campo
es de todos**

Minagricultura