Capítulo IV

Proceso metodológico desarrollado para el sector agropecuario

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 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 & Hidiroglou, 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).

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

donde s_h^2 es la cuasivarianza verdadera en el h-ésimo estrato, n_h 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.