MUESTREO EN POBLACIONES FINITAS 2016

Fundamentos y Métodos

Trabajo Práctico 1

PARTE B

En el dataframe “Marco.Agropecuario.RData” adjunto a la Parte B del TP se presenta un marco muestral de 16050 unidades agropecuarias con la siguiente información censal:

UNIDAD: Identificador único de la empresa agropecuaria (1:16050)
REGION: Región a la que pertenece el establecimiento más importante de la empresa
(1=”Norte”,2=”Centro”,3=”Sur”),
TIPOEMP: Tipo de Empresa
(1=”Familiar”, 2=”SA”, 3=”Otros”),
SUPERFICIE: Hectáreas propias dedicadas a la siembra durante Julio2013 – Julio2014
(se incluyen los establecimientos con 50 ha y más de superficie)
CANTEMP: Cantidad de empleados en el establecimiento a noviembre del 2014,
(se incluyen a las empresas con más de 4 empleados)
SOJA: Hectáreas dedicadas a la siembra de Soja en el periodo Julio2013 – Julio2014
TRIGO: Superficie sembrada de Trigo en el periodo Julio2013 – Julio2014,
BOVINOS: Total de cabezas Bovinas a noviembre del 2014,

El objetivo es buscar una estratificación y un tamaño de muestra óptimo, sujetos a restricciones de precisión sobre algunas variables y en dominios de estimación para el futuro diseño de una encuesta anual agropecuaria.

La propuesta inicial evalúa emplear como características que estratifican al marco muestral a CANTEMP y SUPERFICIE. La primera categorizándola en 5 (cinco) categorías, de la siguiente manera: 1=[min(CANTEMP),20], 2=(20,30], 3=(30,40], 4=(40,50] y 5=(50,max(CANTEM)). En cuanto a la variable SUPERFICIE se acordó dividirla también en 5 estratos: 1=(min(SUPERFICIE),500], 2=(500,1000], 3=(1000,1500], 4=(1500,2000] y 5=(2000,max(SUPERFICIE)].

De acuerdo a las categorizaciones propuestas a las empresas se la estratifica inicialmente en 5*5=25 estratos de diseño o “microestratos”.

Por otro lado para el estudio se piensa emplear a las variables SOJA,TRIGO y BOVINOS como variables objetivo y a la variable REGION como la que define los dominios de estimación ya que son de interés para la encuesta. Se considera que la estratificación deberá permitir estimaciones de dichas variables con coeficientes de variación (CV) que no superen a los definidos por la siguiente tabla

CV deseados para las estimaciones en cada Dominio

REGION	CV(SOJA)	CV(TRIGO)	CV(BOVINOS)
NORTE	0.02	0.01	0.02
CENTRO	0.01	0.01	0.02
SUR	0.05	0.03	0.05

Estos CV son diferenciales por variable y dominio y respetan tanto la intensidad del cultivo en la regiones como la cría de ganado, buscando más precisión en aquellas donde la presencia es más fuerte de acuerdo a la información con la que se cuenta.

Los responsables sospechan que el diseño que involucraría a los 25 microestratos por dominio de estimación (Regiones) atendiendo las restricciones de precisión podría ser

ineficiente ya que llevaría a muchos estratos a tener muy pocas unidades, que los tamaños de muestra sean muy pequeños, o que los cambios de estrato de las unidades a la fecha de la encuesta genere una pérdida en la precisión importante dado que la información no es actual.

Se considera que tener N_h<=10 por estrato equivaldría a pensar en aplicar un censo en él, ya que por principio es poca la cantidad de unidades involucradas para que el muestreo tenga sentido, pero al mismo tiempo censarlos incrementaría los costos; por otro lado tener tamaños de muestra n_h<=4, llevaría a tener muy malas estimaciones de la varianza dentro del estrato de cada dominio que impactarían en los CV finales estimados en las variables objetivo.

Es por esto que se desea evaluar la alternativa de reagrupar los estratos originales definiendo una nueva estratificación a través del algoritmo Genético empleando SamplingStrata.

El estudio es de carácter exploratorio y no necesariamente definitivo pero permitiría comparar ambas alternativas, una empleando el algoritmo de Bethel (para H=25) y la otra surgida del algoritmo Genético sujetas a restricciones de CV en los dominios de estimación. Al mismo tiempo los responsables dispondrían de los primeros tamaños muestrales para ir dimensionando costos y la magnitud del futuro operativo antes de tomar una decisión final.

Se pide:

1) Resolver la optimización por el algoritmo de Bethel empleando los 25 microestratos y las restricciones en CV a nivel Región o dominio detalladas.

2) Si el algoritmo convergió, por dominio analizar y comprobar si existen estratos con N_h<=10 y/o n_h <=4 y/o los que el algoritmo propone como "censo" por dominio de estimación. Se sugiere presentar en una tabla con los resultados del óptimo por dominio y estrato, con sus tamaños poblacionales, muestrales. Completar la tabla con las probabilidades π_k y agregar un resumen con los tamaños muestrales por dominio y los CV alcanzados vs. los propuestos.

3) Resolver la optimización por el algoritmo de Genético buscando reagrupar con un criterio óptimo a los 25 microestratos por dominio, imponiendo la condición que la nueva estratificación no supere los 10 estratos, una restricción tal que n_h>=5 en cada estrato y bajo los mismos CV a nivel Región empleados para el algoritmo de Bethel.
   Teniendo en cuenta que el algoritmo puede llevar a mínimos locales, se propone estudiar y presentar no menos de 3 alternativas modificando algunos parámetros que por default propone el algoritmo (pops, iter, mut_chance y elitism_rate) antes de quedarse con aquella que responde a las precisiones deseadas y al menor tamaño de muestra final alcanzado por samplingStrata[1].

4) Presentar para cada solución o estratificación un resumen con:

   a) Los parámetros u opciones empleadas en el algoritmo
   b) El total de estratos en cada dominio
   c) Tamaños muestrales alcanzados y por dominio

5) Proponer a su criterio una estratificación de las estudiadas como "definitiva"

[1] Retener los dataframes y elementos necesarios para el análisis comparativo y no perderlos entre ensayo y ensayo. En particular el que se origina de *updateStrata()* o en su defecto su equivalente "strata_aggregation.txt" en la carpeta de trabajo renombrándolo para cada ejecución, ya que lo reemplaza en cada ejecución. Una alternativa es ir guardando cada ensayo con su Workspace empleando *save.image()* o el comando que le es habitual para estas situaciones.

acompañándola con una síntesis con los motivos que llevaron a elegirla, más los gráficos de convergencia (plotdom1.pdf-plotdom3.pdf). Comparar los tamaños de muestra total por dominio con los propuestos por Bethel.

6) Presentar la tabla de conversión de los estratos originales en los definitivos en cada dominio señalando en la microestratificación cómo quedaron agrupados.

7) Actualizar el marco muestral con los nuevos estratos.

8) Evaluar el comportamiento de los CV y las estimaciones de cada variable objetivo en cada dominio a través de una simulación con 1000 muestras seleccionadas del marco muestral empleando el comando *evalSolution()* del Package. Interpretar y brindar una explicación de los resultados que exporta: cv.pdf, differences.pdf o differences.csv.

9) Seleccionar una muestra según la solución definitiva empleando el comando *selectSample()* del Package.

10) Presentar una tabla resumen con las probabilidades π_k y las π_{kl} del diseño final adoptado.