Small Area Estimation Methodology in Bogotá Multipurpose Survey

Content – November 2017

1. Objetivos
2. Análisis exploratorio de la información
3. Componentes del proyecto
4. Resultados
Small Area Estimation Methodology in Bogotá Multipurpose Survey

José F. Zea
josezea@usantotomas.edu.co

Felipe Ortiz
andresortiz@usantotomas.edu.co

Universidad Santo Tomás,
Bogotá - Colombia

November, 2017
uRos 2017
Bucharest, Romania
Small Area Estimation Methodology in Bogotá Multipurpose Survey

Content – November 2017

• Motivation
• Scope
• Outcomes
• Bibliography
Cundinamarca location

Latitude: 4.5981, Longitude: 74.0758
Cundinamarca

Cundinamarca has 22,623 km² (8,735 sq mi) and a population of 2,680,041

✓ One of 32 departments of Colombia
✓ 115 municipalities (31 observed in multipurpose survey)
Small Area Estimation

✓ Small area estimation refers to domain estimation with an effective small simple size.

✓ Administrative registers increase available information to area level, with new information sampling error decreased and leads to a improvement of the quality of the estimations.

- Area-level Models (Fay – Harriot)
- Individual - levels models (Batseее – Harter – Fuller)

✓ It is based in mixed models adjustment models which take into account within domain variability.

Universidad Santo Tomás
Bogotá - Colombia

uRos, 2017
Bucharest, Romania,
6-7 November
Income per household

☑ Working Population: It is made up by people who are working, performing some paid activity or who have a job or business for which they receive income. This group also includes people who worked last week without compensation.

• Employees: They can get earnings mainly from salary, bonus, payment in kind (food, transportations, etc), subsidies, indemnities and others. Independent workers get earnings related with some activity.

• Independent worker
• Unpaid Worker
• Unemployed
They can get revenues from rental income, financial aid, sales, pension etc.

☑ Information is aggregated to household level.
Occupational conditions

Unemployed rate is estimate the ratio \((\text{Total unemployed})/(\text{Total EAP})\)

Population of working age

- Economically active population
- Economically inactive population

Employed
Unemployed
Fay-Herriot Model

Fay-Herriot (FH) model links estimated average of interested variable in the area $d$ (for $d = 1, \ldots, D$) $\hat{Y}_d$ with auxiliary information vector $z_d$:

$$\hat{Y}_d = z_d^t \beta + u_d + e_i$$  \hspace{0.5cm} (1)

$u_i \sim N(0, \sigma^2_v)$,

$e_i \sim N(0, \Sigma_e)$, with $\Sigma_e = \text{diag}(D_1, D_2, \ldots, D_m)$.

In above equation $z_d^t \beta + u_d$ is the unknown average for $d$-th area

The best linear unbiased predictor (BLUP) for $\theta_d$, with $\beta$, $\sigma^2_v$ and $D_d$ known is get as follows:

$$\hat{Y}^{BLUP}_d = \begin{cases} z_d^t \beta + \gamma_d \left( \hat{Y}_d - z_d^t \beta \right) & \text{Si } d \in A \\ z_d^t \beta & \text{if } d \notin A \end{cases}$$

where $\gamma_d = \frac{\sigma^2_v}{\sigma^2_v + D_d}$ and $A$ denotes the selected areas in the sample.
Direct Estimations

✓ In this approach $y$ is considered as fixed and the sample is considered as a random variable.

✓ In this approach each sample has a probability $p(s)$ to be selected. In general, $p(s)$ is not straightforward to compute, in order to estimate parameters such as total, mean, ratio and proportions is enough to have only inclusion probabilities. The inclusion probability is computed as:

$$\pi_k = \sum_s p(s)$$

✓ An unbiased direct estimator (under the survey design $p(s)$) for the population total is the Horvitz-Thompson (HT) estimator:

$$\hat{y}_π = \sum_s \frac{y_k}{\pi_k} = \sum_s w_k y_k$$

• HT estimator is widely used to produce official statistics. HT estimator can be written in terms of expansion factor or survey weight, $w_k$ which is the inverse of inclusion probabilities $\frac{1}{\pi_k}$.
Direct Estimations (2)

✓ A direct estimator (not unbiased) for the mean in domain d is given by:

\[ \hat{Y}_d = \frac{\sum_s w_k y_k z_{dk}}{\sum_s w_k z_{dk}} \]

The average of income by municipality is estimated.

✓ A direct estimator (not unbiased) for the ratio in domain d is given by:

\[ \hat{R}_d = \frac{\hat{\gamma}_{\text{y}}_{\text{r},d}}{\hat{\gamma}_{\text{x}}_{\text{r},d}} = \frac{\sum_s w_k y_k z_{dk}}{\sum_s w_k x_k z_{dk}} \]

An example of a ratio is the unemployment rate, which is calculated with the previous expression.

✓ We produce survey estimation with survey package.
Variance estimation of mean and ratio estimators

In order to estimate the variance of this estimator
a. We select 1000 samples (with replacement and probability weights equal to the inverse of survey weights).
b. We estimate the mean for every sample.
c. An approximate design variance is obtained computing variance over 1000 means.

```r
estvar.mean <- function(x, weights, nsim){
  est_mean <- numeric(nsim)
  for(i in 1:nsim){
    set.seed(i)
    pi_k <- 1/weights
    sel <- sample(length(x), replace = T, prob = pi_k)
    est_mean[i] <- sum(x[sel] * weights[sel]) / sum(weights[sel])
  }
  est_var <- var(est_mean)
  est_var
}

estvar.ratio <- function(y, z, weights, nsim){
  est_ratio <- numeric(nsim)
  for(i in 1:nsim){
    set.seed(i)
    pi_k <- 1/weights
    sel <- sample(length(y), replace = T, prob = pi_k)
    est_ratio[i] <- sum(y[sel] * weights[sel]) / sum(z[sel] * weights[sel])
  }
  est_var <- var(est_ratio)
  est_var
}
```

uRos, 2017
Bucharest, Romania,
6-7 November
Auxiliar Information

- Rate of beneficiaries of the Selection System for Beneficiary Social Programs (SISBEN)
- Average of "Pruebas Saber" (Standardized Test similar to the American - SAT) score in the municipality.
- Area (Km²).
- Affiliates Health Contributory and Subsidiary Regime
- Average of Cadastral Appraisal in the municipality
- Rural and Urban Cadastral Appraisals.
- Coverage of Primary and Secondary Education.
- Energy use per capita in the municipality
- Oil royalty payments dependence of the municipality.
- Vaccination Rate
- Municipality Budget Execution (2000 - 2012)
- Poverty Incidence by municipality
- Multidimensional Poverty Index by index
- Municipality unsatisfied basic needs (NBI).
- Homicide Rate by municipalities.
- Average of Cadastral Appraisal in the municipality (rural an urban)
- Population projections (to relativize some measures)
- Sexual Assault Rate
- School drop-off rate
Data Manipulation

Data wrangling was carried on using R tidyverse packages.

Data importation:
✓ haven (Wickham and Miller 2017) for datasets in .sav format (IBM SPSS)
✓ readr (Wickham, Hester, and Francois 2017) for text files
✓ readxl (Wickham and Bryan 2017) to read xls and xlsx files.

Data manipulation

✓ dplyr (Wickham et al. 2017) was used to different process such as filtering, recoding, merging (left, and inner joins) and binding different datasets and to make necessary aggregations.
✓ forcats: define order of categories
✓ tidyr: reshape datasets
BLUP and EBLUP (Fay – Herriot Model)

✓ The best linear unbiased predictor (BLUP) for $\theta_d$, with $\beta$, $\sigma_\nu^2$ and $D_d$ known is get as follows:

$$\hat{Y}_{d}^{BLUP} = \begin{cases} z_d^t \hat{\beta} + \gamma_d \left( \hat{Y}_d - z_d^t \hat{\beta} \right) & \text{Si } d \in A \\ z_d^t \hat{\beta} & \text{Si } d \notin A \end{cases}$$

(2)

where $\gamma_d = \frac{\sigma_\nu^2}{\sigma_\nu^2 + D_d}$ and $A$ denotes the selected areas in the sample.

✓ When $\beta$ y $\sigma_\nu^2$ are estimated we obtained the empirical best linear unbiased predictor (EBLUP). The computation of EBLUP is carried on as:

$$\hat{Y}_{d}^{FH} = \begin{cases} z_d^t \hat{\beta} + \gamma_d \left( \hat{Y}_d - z_d^t \hat{\beta} \right) & \text{Si } d \in A \\ z_d^t \hat{\beta} & \text{Si } d \notin A \end{cases}$$

(3)

EBLUP can be seen as an weighted average of direct estimation $\hat{Y}_d$ and indirect estimation $z_d^t \hat{\beta}$.

• $\hat{Y}_d$ is close to 1, $\hat{Y}_d$ is similar to $\hat{Y}_d$, in the other hand if $\hat{Y}_i$ is closed to 0, the estimator $\hat{Y}_d$ tends to $z_d^t \hat{\beta}$
MSE (Fay – Herriott Model)

(Prasad and Rao 1990) provides an approximation of mean square error of Fay-Herriott estimator which depend of estimation method of $\beta$ and $\sigma^2_u$. With moments method (method developed by the same authors) and restricted maximum likelihood - REML the mean square error of estimations is obtained as:

$$\text{MSE}(\tilde{Y}_i) = \begin{cases} 
  g_{1i}(\hat{\sigma}^2_u) + g_{2i}(\hat{\sigma}^2_u) + 2g_{3i}(\hat{\sigma}^2_u) & \text{Si } i \in A \\
  z_i^t(ZV^{-1}Z^t)^{-1}z_i + \hat{\sigma}^2_u & \text{Si } i \notin A 
\end{cases}$$

where

$$g_{1i}(\hat{\sigma}^2_u) = \frac{\hat{\sigma}^2_u D_i}{\hat{\sigma}^2_u + D_i}, \quad g_{2i}(\hat{\sigma}^2_u) = \frac{D_i^2}{(\hat{\sigma}^2_u + D_i)^2} z_i^t(ZV^{-1}Z)^{-1}z_i$$

and

$$g_{3i}(\hat{\sigma}^2_u) = \left(\frac{2D_i^2}{m(\hat{\sigma}^2_u + D_i)^3}\right) \left(\hat{\sigma}^4_u + 2\hat{\sigma}^2_u \sum_{i=1}^m D_i/m + \sum_{i=1}^m D_i^2/m\right)$$

with $V = \text{diag}(\sigma^2_u + D_1, \ldots, \sigma^2_u + D_m)$.
Small Area Estimation

✓ We select variables with high (absolute value) correlation between direct estimations and auxiliary variables.
✓ Stepwise
✓ To produce small area estimations and their respective MSE we used sae library

threshold <- 0.55
selected_var <- as.character(df_cor$variables[abs(df_cor$V1) > threshold])

formula_income <- as.formula(paste("IncomeMeansByMun ~ ", paste(selected_var, collapse="+")))
income_model <- lm(formula_income, data = df)
income_step_model <- step(income_model)

FH_income <- mseFH(eval(formula(income_)), VarIncomeMeansByMun, data = df)
Top and Bottom Municipalities by Average Income and Unemployment Rate

Visualizations are produced with ggplot2.
MSE of Direct and FH estimators in observed municipalities in EMB

University Santo Tomás
Bogotá - Colombia

uRos, 2017
Bucharest, Romania,
6-7 November
Average income and unemployment rate in Cundinamarca municipalities

Universidad Santo Tomás
Bogotá - Colombia

uRos, 2017
Bucharest, Romania,
6-7 November
Accuracy of estimates

Universidad Santo Tomás
Bogotá - Colombia

uRos, 2017
Bucharest, Romania,
6-7 November
Direct estimates vs FH estimator in observed municipalities
Conclusions and future work

✓ SAE methodology lets minimize survey error of direct estimates.

✓ It is possible to get prediction of different parameters (average of income and unemployment rate) for uncovered municipalities of EMB 2014.

✓ SAE methodology allows to optimize the production of statistics.

✓ R ecosystem makes easy the complete process: data exportation, data manipulation, direct estimations, sae estimation visualization.
References


Thank you

José F. Zea

josezea@usantotomas.edu.co

Andrés Felipe Ortiz

andresortiz@usantotomas.edu.co