Small Area Estimation Methodology (SAE) applied on Bogota Multipurpose Survey (EMB)

José F. Zea (josezea@usantotomas.edu.co)
Santo Tomás University, Bogota, Colombia

Felipe Ortiz (andresortiz@usantotomas.edu.co)
Santo Tomás University, Bogota, Colombia

ABSTRACT

Small Area Estimation Methodology (SAE) is a widely used by statistical offices in several countries to reduce sampling errors with the help of auxiliary information. Different countries such as USA, Canada, England, Israel and European Community have within their statistical institutes offices dedicated to the application of SAE in several investigations. So far, the National Administrative Department of Statistics of Colombia (DANE), has not published official statistics that involve this methodology. The present work illustrates the advantages in the use and estimation of living conditions using SAE. Formally, the unemployment rate and the average income levels of municipalities of Cundinamarca are estimated. For this purpose, information of the Multipurpose Survey 2014 is used and is complemented with socio-demographic and economic related auxiliary information. A mixed Fay & Herriot (1979) model it is used in order to get the estimates.

We use R ecosystem to develop SAE methodology. R is used for data wrangling, model adjustment, parameter estimation and finally visualization with the aid of renowned packages such as tidyr, forcats, sae, ggplot2 among others.

We will show R implementation and some remarkable results. First, a good adjustment of the model to the data; second, a reduction in the sampling errors reported by the estimation in small areas compared to the direct estimates generated by the Bogota Multipurpose Survey (EMB); and finally acceptable estimates for municipalities that were not covered by the survey.

Keywords: *Small Area Estimation, Survey Sampling, Tidyvese, Household Survey, Colombia, Cundinamarca Municipalities.*

JEL classification: *O54, C63, C81, C83, C88*

INTRODUCTION

Obtaining reliable estimates for municipalities or smaller geographical areas has been an impossible task for household surveys with traditional sampling methodologies due to financial sources and operative costs. In the present work we seek to obtain precise estimates at municipality level of average income and unemployment rate for Cundinamarca to achieve more effective public policies.

Cundinamarca is one of 32 departments of Colombia with 116 municipalities including Colombian capital Bogota. In this study we make estimations of average household income and unemployment rates in 115 municipalities (excluding Bogota due to administrative issues and for obtaining a better adjustment in estimations).

Multipurpose Survey covers 32 municipalities, and Bogota which will not be taken into account in this study because it is an outlier municipality in terms of average income and has a very different economic dynamic from the other municipalities in Cundinamarca; due to its condition, as a capital of Colombia and the biggest city in terms of population, economic development, etc. More details of the survey are found in (DANE 2014).

In this study we have two main goals:

- To optimize the production of official statistics by improving the estimates of the average income and the unemployment rate to observe Cundinamarca municipalities in the 2014 Multipurpose Survey.
- To provide estimations of average income and unemployment rate for non-covered municipalities in the 2014 Multipurpose Survey.

In this work we use R ecosystem in different stages such as:

- Data wrangling of auxiliary information to municipality level and multipurpose sampling dataset (EMB).
- Direct estimates of average household income by municipality, unemployment rate and their respective variances by municipality.
- Estimates of average household income by municipality and unemployment rate by municipality using SAE in particular Fay-Herriot model.
- Visualization.

R code for reproducibility can be obtained from https://github.com/josezea/sae_emb

LITERATURE REVIEW

There are two main approaches in survey sampling theory: design-based survey sampling (Sarndal, Swensson, & Wretman 1992) and model-based survey sampling (Rao & Molina 2015).

1.1 Direct estimators for domains

In design-based approach the values of the interest variable y are 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 inclusion probabilities.

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

$$\hat{t}_{y\pi} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{n} w_k y_k$$

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 $1/\pi_k$.

A direct estimator (not unbiased) for the mean is:

$$\hat{\bar{Y}} = \frac{\sum_{k=1}^{n} w_k y_k}{\sum_{k=1}^{n} w_k}$$

The above estimator is a particular case of a ratio estimator:

$$\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}$$

where $\hat{t}_{y\pi} = \sum_{k=1}^{n} w_k y_k$ and $\hat{t}_{z\pi} = \sum_{k=1}^{n} w_k z_k \cdot y_k$ and z_k are variables of interest observed in the survey.

An example of a ratio is the unemployment rate, which is calculated with the previous expression. The numerator is the number of unemployed people in a reference period $\hat{t}_{y\pi}$, the denominator $\hat{t}_{z\pi}$ is the number of economically active people in the same period. Both the numerator and denominator are computed.

An example of a mean (computed as a ratio) is mean income. Numerator and denominator of a mean are random variables and a non-linear function of total. Therefore, the mean estimator is biased; the variance of mean is increased due to randomness of the numerator and denominator.

In many occasions the interest is the estimation of parameters for subpopulations such as age groups, municipalities, combination between age groups and sex categories, etc. Those groups are called domains. Estimation of parameters in domains can be done through direct estimators and SAE methodology. The direct estimation of total for domain d is get as follows:

$$\hat{t}_{y\pi,d} = \sum^{s} \frac{y_k z_{dk}}{\pi_k} = \sum^{s} w_k y_k z_{dk},$$

where $z_{dk} = 1$, if element *k* is in the domain d and $z_{dk} = 0$, in other cases.

An estimator for the mean can be computed as:

$$\hat{\bar{Y}}_d = \frac{\hat{t}_{y\pi,d}}{\hat{N}_d} = \frac{\sum^{s} \frac{y_{dk}}{\pi_k}}{\sum^{s} \frac{z_{dk}}{\pi_k}} = \frac{\sum^{s} w_k y_{dk}}{\sum^{s} w_k z_{dk}},$$

where $y_{dk} = y_k z_{dk}$.

FAY-HERRIOT ESTIMATOR

Small area estimation (SAE) refers to estimation in domains (population subgroups) which have a relatively small sample size. Some examples of a small area such as counties, city administrative divisions (eg. localities in Bogota). In this research municipalities of Cundinamarca are considered small area.

SAE is used to carry on the estimation of parameters for domains (small area) with a mixed approach between design-based and model-based estimation. SAE is based in the adjustment of mixed models which take into account within-domain variance. Auxiliary information for small area is added to decrease survey sampling error and therefore an increase in quality of the estimations. There are two main types of SAE estimators: individual-level models (Battese, Harter & Fuller 1988) and area-level models (Fay & Herriot 1979) (Rao & Molina 2015).

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

$$\hat{\bar{Y}}_d = \mathbf{z}_d^t \beta + u_d + e_d$$

where $u_d \sim N(0, \sigma_v^2)$, $e_d \sim N(0, \Sigma_e)$, with $\Sigma_e = diag(D_1, D_2, \dots D_D)$. In the above equation $\mathbf{z}_d^t \beta + u_d$ is the unknown average for d^{th} area and D_d is a known term, usually taken as variance of $\hat{\bar{Y}}_d$ (under the survey design $p(s)$).

The best linear unbiased predictor (BLUP) for θ_d, with β, σ_v^2 and D_d known is obtained as follows:

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

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

When β y σ_v^2 are estimated we obtained the empirical best linear unbiased predictor (EBLUP). The computation of EBLUP is carried on as:

$$\hat{\bar{Y}}_d^{FH} = \begin{cases} \mathbf{z}_d^t\hat{\beta} + \hat{\gamma}_d\left(\hat{\bar{Y}}_d - \mathbf{z}_d^t\hat{\beta}\right) & \text{Si } d \in A \\ \mathbf{z}_d^t\hat{\beta} & \text{Si } d \notin A \end{cases} \quad [1]$$

EBLUP can be seen as a weighted average of direct estimation $\hat{\bar{Y}}_d$ and indirect estimation $\mathbf{z}_d^t\hat{\beta}$. If $\hat{\gamma}_d$ is closed to 1, $\hat{\bar{Y}}_d^{FH}$ is similar to $\hat{\bar{Y}}_d$, on the other hand if $\hat{\gamma}_i$ is closed to 0, the estimator $\hat{\bar{Y}}_d^{FH}$ tends to $\mathbf{z}_d^t\hat{\beta}$.

Mean square error for Fay-Herriot estimator

The Mean Square Error (MSE) provides an approximation of mean square error of Fay-Herriot estimator which depends on estimation method of β and σ_u^2 (Prasad & Rao 1990). With the *moments method* (method developed by the same authors) and *restricted maximum likelihood - REML* the mean square error of estimations is obtained as:

$$MSE(\hat{\bar{Y}}_d^{FH}) = \begin{cases} g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2) & \text{Si } d \in A \\ \mathbf{z}_d(\mathbf{Z}\mathbf{V}^{-1}\mathbf{Z}^t)^{-1}\mathbf{z}_d^t + \hat{\sigma}_u^2 & \text{Si } d \notin A \end{cases}$$

where

$$g_{1d}(\hat{\sigma}_u^2) = \frac{\hat{\sigma}_u^2 D_d}{\hat{\sigma}_u^2 + D_d}, \qquad g_{2d}(\hat{\sigma}_u^2) = \frac{D_d^2}{(\hat{\sigma}_u^2 + D_d)^2}\mathbf{z}_d^t(\mathbf{Z}^t\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{z}_d$$

and

$$g_{3d}(\hat{\sigma}_u^2) = \left(\frac{2D_d^2}{m(\hat{\sigma}_u^2 + D_d)^3}\right)\left(\hat{\sigma}_u^4 + 2\,\hat{\sigma}_u^2\sum_{i=1}^{D} D_d/D + \sum_{i=1}^{D} D_d^2/D\right) \text{ with } \quad \mathbf{V} = diag(\sigma_u^2 + D_1, \ldots, \sigma_u^2 + D_D)$$

Basic concepts of Bogota Multipurpose Survey (EMB)

The concepts related to labor force and income that we will use in this section are based in the definitions used by the Colombian Official Statistics Agency (National Administrative National Department of Statistics - DANE) and may differ slightly from the concepts used in the International Labour Organization - ILO).

In EMB survey analyzed population corresponds to houses, households and people. In this study, all economically active household members were considered, household members were grouped according to the occupational situation:

a. Working Population: consist of 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 the previous week without compensation.
- Employees: worker, private company employee, public employee, domestic employee and hired worker.
- Independent worker: people who performed in the previous week any activity pays for one hour or more, or perform at work in the following activities: private enterprise worker/employee, government employee, domestic employee and day laborer.
- Unpaid Worker: people who performed the previous week some activity pays for one hour or more or perform at work in the following activities: unpaid assistant, unpaid family worker and unpaid worker.

b. Unemployed: it consists of people who are not working and not performing any paid activity or who have a job or business for which they receive income. This group excludes people who worked the previous week without compensation.

Household members can get revenues from rental income, financial aid, sales, pension, etc.

Employees get earnings mainly from salary, bonus, and payments in food, transportations, subsidies, indemnities and others. Independent workers get earnings related to any activity.

In order to compute unemployment rate it is necessary to identify employed and economically active people. In order to achieve that, DANE provides the next definitions:

- Population of working age (PWA): this group consists of people aged 12 and over in urban areas and over 10 years in rural areas.
- Economically Active Population (EAP): also called labor force and

consists of people who are in working age and are working, or are looking for a job.

- Economically Inactive Population (EIP): consists of people who are in working age who neither work nor look for a job.
- Employed (E): consists of economically active population who in the reference period were in one of the following situations: worked for at least one hour paid in cash in the reference week. Those who did not work the reference week, but had a job. Unpaid family workers who worked in the reference week for at least 1 hour
- Unemployed (U): consists of economically active population who in the reference period is not in any situation described for employed people.

DANE methodology for compute unemployed rate is estimated by the ratio the ratio $\frac{\text{Total unemployed}}{\text{Total EAP}}$.

METHODOLOGY

Data wrangling

The datasets corresponding to occupational condition was used in order to compute income for every household of Cundinamarca municipalities and unemployment for the surveyed people living in Cundinamarca.

Data wrangling was carried out using R tidyverse packages. For data importation we used haven (Wickham & Miller 2017) for datasets in .sav format (IBM SPSS), *readr* (Wickham, Hester & Francois 2017) for text files and *readxl* (Wickham & Bryan, 2017) to read xls and xlsx files. *dplyr* (Wickham et al. 2017) was used in different processes such as filtering, recoding, merging (left, and inner joins) and binding different datasets and to make necessary aggregations. The script "1.DataWrangling_EMB.r" in the *github* repository contains all the details of the processing information.

Datasets from different sources was explored to use as auxiliary information. Tidyverse packages was used in order to conform a unified database of auxiliary variables, some of the auxiliary variables considered are:

- Rate of beneficiaries of the selection system for beneficiary social programs (SISBEN)
- Average of "Prueba Saber 11" (Standardized test similar to the American - SAT) score in the municipality.
- Area (Km^2).
- 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 index.
- Municipality unsatisfied basic needs.
- 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-out rate.

Direct estimations

Once the EMB and auxiliary information is depurated, direct estimations are carried on. The survey design of EMB is probabilistic, each element of the population (households) has a non-zero probability to be selected. In particular, it is a clustered and stratified design.

In order to improve the accuracy of the estimations, it is considered as strata the social stratification in Bogota (local government mechanism to assign subsidies in public services to the poorest households).

In this design, clusters correspond to a set of houses located within the same block, this group of houses is called a size measure segment. In each selected segment all houses are surveyed.

The average of income is estimated as follows:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} w_k \, y_k}{\sum_{i=1}^{n} w_k}$$

where w_k is the survey weights (the inverse of inclusion probabilities $\frac{1}{\pi_k}$).

In R, We estimate the average using survey package (Lumley, 2017). The sampling design used in this exercise is complex (to increase cost efficiency), it involves stratification and cluster sampling. Unfortunately, population sizes, probabilities and weights associated to different survey sampling stages are not delivered by Colombian Statistical Office (DANE) to general public, instead they only provide the final expansion factor.

We use the fact that we can approximate inclusion probability π_k to mp_k (where p_k the selection probability associate to a with replacement

sampling design) and we estimate variance of the estimator through a with-replacement sampling design:

We estimate the average in every municipality with their respective standard errors and variance through survey package (Lumley, 2017) as follows:

```
aprox_design <- svydesign(ids =~ 1, weights =~ SURVEY_WEIGHTS,
                          data = HouseholdIncome )

df_est.income <- svyby(~INCOME, ~ID_MUNIC, aprox_design, svymean)
df_est.income$varmeansByMun <- df_est.income$se ^ 2

df_est.income$cvemeansByMun <- 100 * cv(df_est.income)
```

The selection probability of different stages are not available in DANE open data portal, instead of this we use final sampling weight (SURVEY_WEIGHTS) which is available in EMB dataset delivered by DANE. We do not incorporate *weights* parameter in order to develop a with replacement survey design.

The *svydesign* function allows to define survey sampling design, *svyby* function is used to get aggregated estimations (by municipalities for example).

Estimations of means, estimated variance of mean estimator, and *cve* is done for 31 observed municipalities in Multipurpose Survey.

Other variable of interest is the unemployment rate which is computed as the ratio of total of unemployed people (t_y), and total economically active population (t_z). Those totals are computed over all surveyed households.

$$\hat{R} = \frac{\sum^s w_k y_k}{\sum^s w_k z_k}$$

where w_k is the survey weights.

Ratio estimation is carried on for every municipality with their respective standard errors and variance through survey package (Lumley, 2017) as follows:

```
aprox_design_Unemployment <- svydesign(ids =~ 1,
                              weights =~ SURVEY_WEIGHTS, data = UNEMPLOYMENT
)

df_est.unemployment <- svyby(~Unem, by=~ID_MUNIC, denominator=~EAP,
                         design = aprox_design_Unemployment, svyratio)

names(df_est.unemployment)[c(1,2)] <- c(„ID_MUNIC”, „unempByMun”)
df_est.unemployment$varunempByMun <- df_est.unemployment$’se.DS/PEA’ ^ 2
df_est.unemployment$cve.unempByMun <- 100 * cv(df_est.unemployment)
}
```

In *svyby* we define numerator variable as total of Unemployed people, and denominator as Economically Active Population (EAP), `svyratio` indicates that we are estimating a ratio.

Fay-Herriot estimations

An unified dataset with direct estimation and auxiliary information for 31 municipalities is formed after joining respective dataframes:

```
library(dplyr)
df <- left_join(df_EMB, AuxInfo, by = „ID_MUNIC”)
```

For the estimated average of income a forward stepwise procedure is carried out to select auxiliary variables.

```
print(formula(income_step_model))
## IncomeMeansByMun ~ CONSUMO_ENERGIA_PER_HABIT + PUNTAJE_SABER +
##     AVALUOS_CATASTRALES_RURALES + NBI_2010
```

The selected variables for average income are:

- Energy use per capita in the municipality.
- Municipality unsatisfied basic needs.
- Average of "Prueba Saber 11".
- Average of cadastral appraisal in the municipality.

A Fay-Herriot model for estimated average income is adjusted with the aid of *sae* library as follows:

```
FH_income <- mseFH(IncomeMeansByMun ~ CONSUMO_ENERGIA_PER_HABIT +
                    PUNTAJE_SABER + AVALUOS_CATASTRALES_RURALES + NBI_2010),
                  vardir = VarIncomeMeansByMun, data = df_income_model)
```

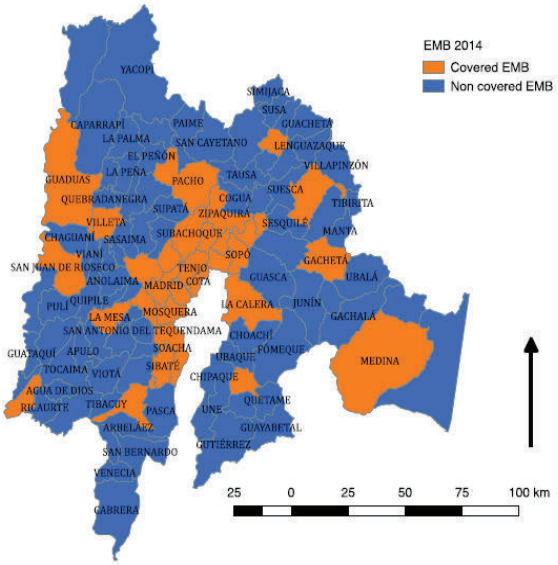
$\hat{\overline{Y}}_U$ is adjusted by energy consumption (CONSUMO_ENERGIA_PER_HABIT), score standardized national education test (PUNTAJE_SABER) and cadastral appraisals (AVALUOS_CATASTRALES_RURALES), the directed variance (VarIncomeMeansByMun) of every area is estimated by survey design using the variance estimation methodology previously described.

RESULTS

The covered municipalities (Covered EMB) in EMB survey are presented in figure 1:

Covered and non-covered municipalities in EMB 2014

Figure 1

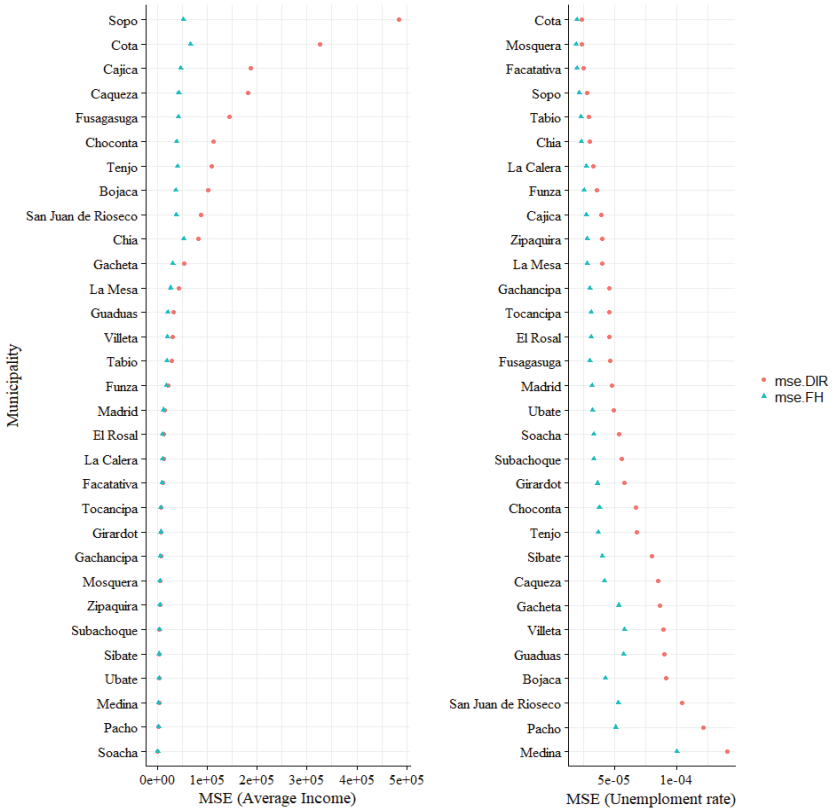


One of the main aspects of the Fay-Herriot estimator is that improves direct estimator in terms of accuracy. In figure 2 it can be noted that mean square error of Fay-Herriot estimator for average income is smaller than the mean square error of direct estimator.

The same occurs for the MSE of unemployment rates estimator, the Fay-Herriot estimator has a lower MSE than the direct estimator.

MSE of Direct and FH estimators in observed municipalities in EMB 2014

Figure 2

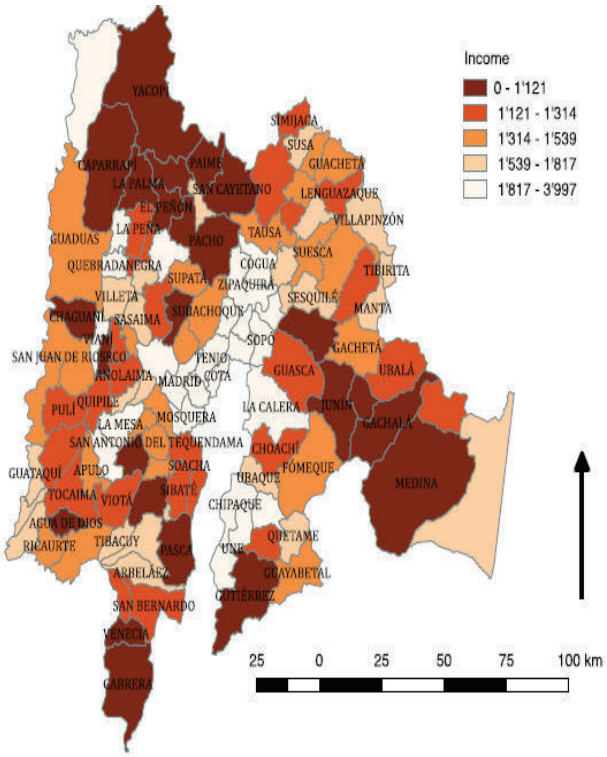


As mentioned earlier, one of the most attractive aspects of small area estimation methodology is obtaining estimates for domains where no surveys were conducted (see equation [1]). Observed domains ($d \in A$) are predicted as a linear combination of direct estimation and the model. In unobserved domain ($d \notin A$) the domain estimates are obtained only with model predictions.

In this case, predictions were made for 85 municipalities where no surveys were carried out. A map of average income (income in thousand Colombian pesos) by municipality is presented in figure 3.

Income in Cundinamarca municipalities

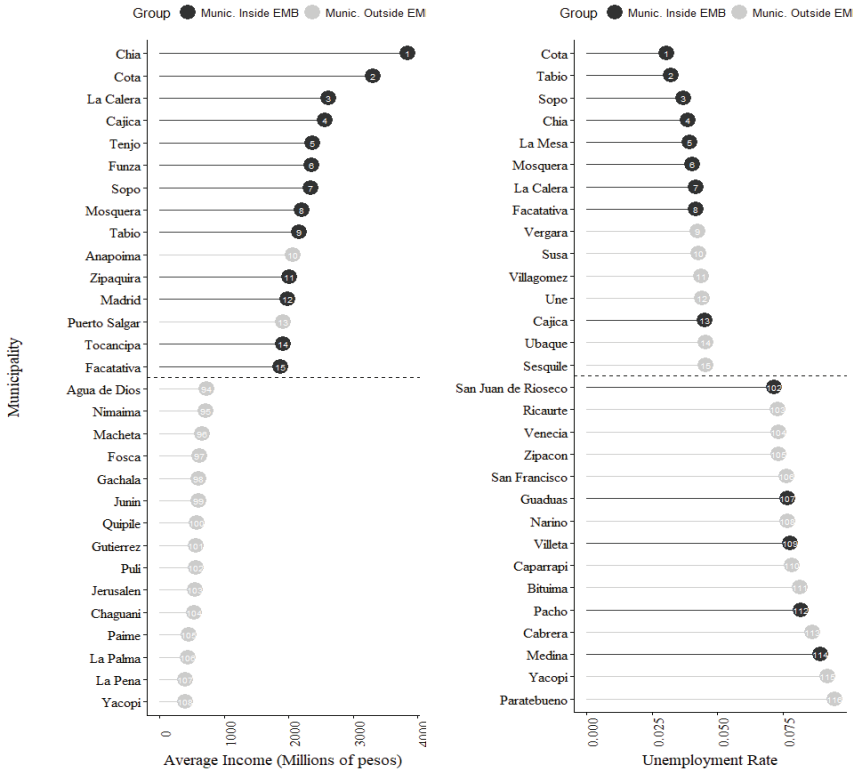
Figure 3



In figure 4 average income and unemployment are shown for both 15 top and bottom municipalities of Cundinamarca in terms of income (millions of Colombian pesos) and unemployment rate. Predictions for observed and not observed municipalities in EMB are computed using FH estimator (see equation [1]).

Top and bottom municipalities by average income and unemployment rate

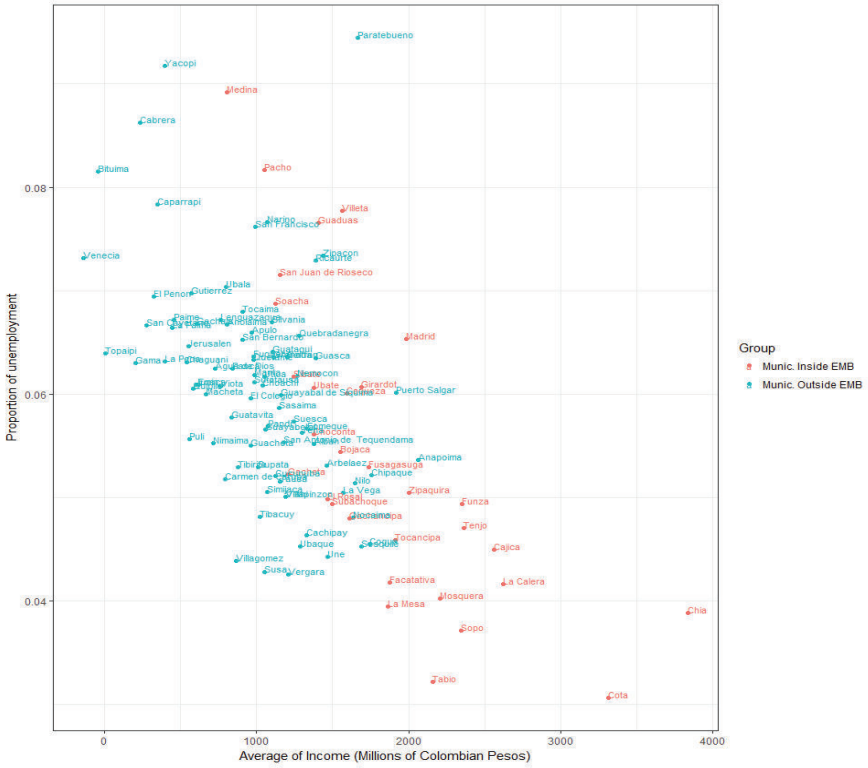
Figure 4



It is interesting to note that there is a very strong linear relationship between estimates for average income and unemployment rate generated by Fay-Herriot model as it is observed in figure 5.

Relationship between average of income and proportion of unemployment rate

Figure 5



CONCLUSIONS

In this article we developed SAE methodology in order to minimize sampling error of direct estimations. In addition, we obtained plausible predictions for municipalities not covered by the EMB 2014.

The R software was a very useful tool to carry out the information processing (with tidyverse R packages), calculations of the point estimates with the direct estimator and the Fay-Herriot estimator and their respective variances (sae package), as well as the visualization and diffusion of the results (document, presentation).

REFERENCES

1. **Battese, G. E., Harter, R. & Fuller, W A**, 1988, *"An error-components model for prediction of county crop areas using survey and satellite data"* Journal of the American Statistical Association, no. 401(83): 28–36.
2. **DANE**, 2014. "EMB2014." October. Taken from: https://formularios.dane.gov.co/Anda_4_1/index.php/catalog/189/.
3. **Lumley T.**, 2017, *"Survey: analysis of complex survey samples"*. R package version 3.32.
4. **Molina I** and **Marhuenda Y**, 2015, *"sae: An R Package for Small Area Estimation."*
5. **The R Journal**. 7(1), pp. 81-98. http://journal.r-project.org/archive/2015-1/molina-marhuenda.pdf
6. **Prasad, N.G.N.**, and **J.N.K. Rao,** 1990, *"The Estimation of Mean Squared Errors of Small Area Estimators."* Journal of the American Statistical Association, no. 85: 163–71.
7. **Rao, J.N.K.**, and **I. Molina,** 2015, *Small Area Estimation*. 2nd ed. Wiley.
8. **Sarndal, C. E., B. Swensson,** and **Jan Wretman**, 1992, *Model Assisted Survey Sampling*. 3rd ed. Springer.
9. **Wickham, Hadley** and **Jennifer Bryan**, 2017, Readxl: Read Excel Files. https://CRAN.R-project.org/package=readxl.
10. **Wickham, Hadley**, and **Evan Miller**, 2017, Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files. https://CRAN.R-project.org/package=haven.
11. **Wickham, Hadley, Romain Francois, Lionel Henry,** and **Kirill Muller**, 2017, Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.
12. **Wickham, Hadley, Jim Hester**, and **Romain Francois**, 2017, Readr: Read Rectangular Text Data. https://CRAN.R-project.org/package=readr.