IPS23, New Data Sources Meet Household Surveys: Scaling-up Innovation for Official Statistics: Colombia’s experience
Integrated social security register (PILA), business register and population register

Context

- The enhancement of Administrative Records (AARR) as one of DANE’s objectives for the 2018-2022 period.
- Despite that DANE has been increasing the use of AARR since before the COVID-19 pandemic, this was an opportunity for innovation.
- The change in the household survey data collection operative highlighted the need to use external sources to check the validity of the information and complementing it.

Advantages

- Increase the granularity of the information. The use of AARR allowed us to reach sample representativeness for new subgroups. For example, we managed to present sex composition by economic sector.
- The imputation process, using machine learning techniques, of informality condition in GEIH (Great Integrated Household Survey) could be made using the linkage with PILA data.

Challenges

- IT capacity
- Updated information
- Linkage: The personal identification number (cédula) is not mandatory for respondents in the GEIH. Thus, it is troublesome to match GEIH with PILA. To increase the matching rate GEIH-PILA, we used record linkage methods where no perfect integration key existed, using variables such as names, surnames and birth dates with reference tables.
- Integration of multiple data sources: It is necessary to harmonize the information since not all administrative records have statistical purposes
- Statistical Unit: PILA excludes especial regimes of social security such as district school's teachers as well as informal employees without social security payments. This could result in a bias as there are differences in gender participation between included and excluded groups.
Cronología de eventos
Gran Encuesta Integrada de Hogares (GEIH) 2020

January – February 2020

In-person GEIH
• In the months of January and February the GEIH was implemented in person.

March 6th 2020

COVID-19
• First positive case of COVID-19 in Colombia

March 26th 2020

Isolation measures
• National mandatory preventive isolation begins

March 25th 2020

Urban telephone GEIH operation starts
• Telephone collection begins in the 23 cities and A.M.
• Reduced form: only salary and net earnings questions are included
• Rural area and urban rest remained face-to-face

August 11th 2020

In-person GEIH
• Complete collection is resumed for in person GEIH
• Complete form (including all income modules)
• August to December is in person

Income information recovery operation
• An income information recovery operation was implemented for the period March-July 2020

August 31st 2020

7th of November 7 – 31st of December, 2020

Operativo de recuperación de información de ingresos
• Se implementó un operativo de recuperación de información de ingresos del periodo marzo-julio 2020
Agenda

1. Machine Learning imputation – informality rate

2. Substitution – coldeck imputation - poverty and inequality rates
Imputation process
ML
Cold deck

GEIH + AARR
Integration of PILA and the household survey

Microdata imputation - ML

First approach we are currently working on:

**DANE's measures:**
- In March and April, DANE measured the unemployment and employment rates, but due to the changes in data collection the informality rate was not calculated

**Goal:**
- Estimate a dummy informality variable to impute the microdata of the household survey to recover the informality series

**Sources:**
- Social security register - PILA
- Household survey - GEIH

**Methodology**

**Input**
- Link PILA-GEIH
- PILA status, and GEIH covariates

**Modelling**
- Machine learning
- Classification model – Random Forest

**Estimate**
- Random Forest model
- Impute GEIH microdata

**Results and Validation**
- Symmetric confusion matrix
- Precision, Recall F1 score greater than 0.8
Coldeek imputation – Poverty and inequality rates
Integration of PILA + government allowances AARR + household survey

DANE’s measures:
During the March – July period, the data collection faced two main challenges:
- Change of face to face to a phone operative.
- The validity re-call operative that was done in December due the reduction of the questionnaire

Goal:
- Use administrative records to improve the GEIH, and reduce possible coverage (non-response) and accuracy biases

Sources:
- Social security register – pensions PILA
- Government allowances i.e FA; JA; CM
- Household survey - GEIH

Methodology

**Input**
- Link pensions register – GEIH
- Link governance allowances registers – GEIH

**Harmonization in statistical unit and time variable**
- The definition in some registers about household is different between the survey and de AARR
- The reference period in the GEIH and AARR is different

**Input –Output Quality Assessment**
- Evaluate the different coverages, the non-response at the different income sources
- We check the consistency between the answer report in the survey and the observed value in the register

**Results and Validation**
- We used the 2019 survey as a reference period to validate the results in a period without pandemic status, and check if the integration produced biases from the official results.
Income Sources

The added income sources were:

- First Activity Monetary Income
- Second Activity Monetary Income
- In Kind Income
- Inactive and Unemployed Monetary Income
- Interests and Dividends Per Investment (IOF1)
- Pensions (IOF2)
- Alimony and Household Assistance (IOF3H)
- Institutional Aid (IOF3I)
- Effective Rent (IOF6)
The statistical commission of ECLAC suggests using Administrative Records (AARR) to solve possible coverage biases, given the variations that may occur due to changes in collection operations. In this regard, DANE uses AARR as a source of contrast and imputation to minimize possible coverage biases in the income recovery operation, due to the time lag in the collection of information.

Although the GEIH does not ask for mandatory identification document type and number fields since 2018, the completeness and consistency (evaluated with the REBP) of these fields allow linking the survey with different data sets.

The integration of surveys with AARR has the potential to improve the precision and accuracy in the measurement of some sources of income subject to possible: coverage biases, non-response biases and measurement biases.

For the fiscal year 2020, 2019 was used as a counterfactual to assess the relevance of the use of AARR. So for 2019 and 2020, a coverage analysis was carried out, which consisted of 1) contrasting the observed value in the survey and the population value of AARR; and 2) an analysis of reporting biases, based on the comparison between the observed income in the common population between the survey and the AARR.
• For the construction of the First Activity Monetary Income (IMPA) the salaries and net earnings reported in GEIH were used. For the March-July 2020 period, in which the form was shortened and the operation shifted to a CAPI collection, the other incomes that are comprised by the IMPA could be estimated thanks to the income information recovery operation. This source allows to build the information of ISA, IMDI, IOF1 (interest and dividends), IE (income in kind), IOF6 (effective rent).

• For the sources of income IOF2 (pensions) and IOF3I (institutional aid), administrative records are used to enrich and complement the information reported in the survey:
  
  • Pensions (IOF2): the amount of the pension table reported in the PILA is used.
  • Institutional aid (IOF3I): the monthly amount of the total charges observed in the "portability zone" of each individual during the 2020 term is used.

• Considering the structural shock on income caused by the COVID-19 situation for 2020, the partitions of the sample for the imputation of income are the following: January-February, March-July and August-December.
Quality Management of the Administrative Records

Input Protocol

1. Preparation of the administrative records
   - Cleaning of variables in terms of present invalid characters.
   - Elimination of leading zeros in identification documents.
   - Variable improvements.
   - Verification of formats (i.e., dates), detection of default values.

2. Homologation, standardization and codification
   - The standardization of the received administrative record is carried out to adapt it to a standard structure and variable coding, according to the established reference tables.

3. Consistency Validation
   - Checking for missing and consistency of the identification variables (use of reference tables).
   - Duplication detection and protocols per record.
   - Consistency checks in each variable, this is determined against their attributes.

4. Statistical Unit Definition
5. GEIH-AARR Substitution and Integration
- Matching GEIH information with RRAA
- GEIH-RRAA match consistency tests
- Information substitution

6. Design of scenarios and imputation windows
- Jan - Jun / Jul - Dec
- Jan - Feb / Mar - Jul / Aug-Dec

7. Hot deck – process ind. By scenario
- Construction of the ICV.
- Imputation of the ICV for households with information gaps.
- Generation of sextiles.
- Stratum imputation
- Calculation of shortages and conformation of income sources.
- Generation of variables for quantile regressions. (They are different in each source).
- Imputation of years of education, time worked and time in that job. The imputation of education is made for each source of income
- Estimation of quantile regressions and identification of outliers.
- Imputation.
- Output consistency validation
Module M. Other Income Sources

Government aid questions flow

During the PAST TWELVE MONTHS, did you receive money from other households, people, or institutions; money for interest, dividends, profits or severance pay?

YES

C. Aid in money from institutions in the country or from outside the country?

YES

C1. Was the aid money from government entities?

YES?

¿Más Familias en Acción? How much?

¿Jóvenes en Acción? How much?

¿Colombia Mayor? How much?

OTHER YES? How much? When?
## Extreme Income Poverty by Scenarios

Main geographical domains and 23 cities, and metropolitan areas

<table>
<thead>
<tr>
<th>Source</th>
<th>Administrative Record</th>
<th>Aid Type</th>
<th>GEIH enhanced question</th>
<th>2019</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Prosperity Department (DPS)</td>
<td>Mas Familias en Acción (FA)</td>
<td>Ordinary transfer</td>
<td>P1661S1A1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additional transfer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jóvenes en Acción</td>
<td>Ordinary transfer</td>
<td>P1661S2A1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additional transfer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Colombia Mayor (CM)</td>
<td>Ordinary transfer</td>
<td>P1661S3A1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additional transfer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compensación de IVA-CM transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compensación de IVA-FA Transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ingreso Solidario (IS) Transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bogotá Bogotá Solidaria Transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bucaramanga Bono vital Bucaramanga Transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medellín Medellín Transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Health Ministry Entidad Pensionados Transfer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

**Fuente:** DANE, match GEIH – RRAA Ayudas y PILA, 2019 - 2020.
### Más Familias en Acción (FA) – GEIH Match

**Deterministic Match Summary – Households and/or Holders**

<table>
<thead>
<tr>
<th>Year</th>
<th>Observations</th>
<th>Holders</th>
<th>Beneficiaries</th>
<th>Household Match GEIH-AARR</th>
<th>No match - but Household reports the aid program in GEIH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GEIH 2019</strong></td>
<td>756,063 obs</td>
<td>2,531,366 holders</td>
<td>33,854 obs (2,505,189)</td>
<td>1,450,638</td>
<td>2,528,571</td>
</tr>
<tr>
<td><strong>GEIH 2020</strong></td>
<td>747,822 obs</td>
<td>2,422,077 holders</td>
<td>30,053 obs (2,263,073)</td>
<td>1,435,833</td>
<td>2,422,077</td>
</tr>
</tbody>
</table>

**Fuente:** DANE. GEIH 2020. - DPS. RRAA FA 2020. Match Determinístico
Further questions:
• Juan Sebastian Ordoñez
  (jsordonezh@dane.gov.co)