Producing Small Area Estimates for Labor Market Indicators in Latin America
A Bayesian Perspective

Statistics Division
UN-ECLAC

Andrés Gutiérrez, PhD.
Regional Adviser on Social Statistics
andres.gutierrez@un.org
The need for disaggregations in labour force surveys

An SDG perspective
8 DECENT WORK AND ECONOMIC GROWTH

Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all.
Target 8.3. Promote policies to support job creation and growing enterprises.

- Promote development-oriented policies that support productive activities, decent job creation, entrepreneurship, creativity and innovation, and encourage the formalization and growth of micro-, small- and medium-sized enterprises, including through access to financial services.

  - Indicator 8.3.1: Proportion of informal employment in total employment, by sector and sex.
SDG 8: Decent work

Target 8.5. Full employment and decent work with equal pay.

- By 2030, achieve full and productive employment and decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value.
  
  - Indicator 8.5.1: Average hourly earnings of employees, by sex, age, occupation and persons with disabilities
  - Indicator 8.5.2: Unemployment rate, by sex, age and persons with disabilities.
SDG 8: Decent work

Target 8.6. Promote youth employment, education and training.

• By 2020, substantially reduce the proportion of youth not in employment, education or training.
  • Indicator 8.6.1: Proportion of youth (aged 15–24 years) not in education, employment or training.
Target 8.7. End modern slavery, trafficking and child labor.

- Take immediate and effective measures to eradicate forced labour, end modern slavery and human trafficking and secure the prohibition and elimination of the worst forms of child labour, including recruitment and use of child soldiers, and by 2025 end child labour in all its forms.

  - Indicator 8.7.1: Proportion and number of children aged 5–17 years engaged in child labour, by sex and age.
LEAVE NO ONE BEHIND
Fundamental principle of data disaggregation

Sustainable Development Goal indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migration status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics.

**General Assembly resolution - 68/261**
Fundamental principles of official statistics

The public’s essential confidence in the integrity of official statistical systems and the credibility it gives to statistics depend to a large extent on respect for the fundamental values and principles that underpin any society that seeks to understand itself and respect the rights of its members and which, in this context, the professional independence and accountability of statistical agencies are crucial.

**General Assembly resolution - 68/261**
17
PARTNERSHIPS FOR THE GOALS

Strengthen the means of implementation and revitalize the global partnership for sustainable development.
SDG 17: Partnership for the goals

Target 17.1. Enhance availability of reliable data

- By 2020, enhance capacity-building support to developing countries, including for least developed countries and small island developing States, to increase significantly the availability of high-quality, timely and reliable data disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location and other characteristics relevant in national contexts.
Quick SAE intro
What is it about?

Surveys depend on large sample size and a proper sampling strategy (sampling design and estimator); they also rely on a robust inferential system that provides precise and exact estimation in planned domains.

When the sample size of the survey is not enough for some subgroups of interest, it is necessary to resort to external auxiliary information (censuses, administrative records, satellite imagery) so that together (surveys and external data) a precise inferential system can be built.

UN-ECLAC uses SAE models to integrate data from different sources.
Coefficients of variation in different disaggregations. Source: NSO - Chile
What is a small area?

Estimates of the parameters of interest might be necessary for geographical disaggregation, allowing visualization on a map, or for sociodemographic subgroups.

- Geographical units could include states, provinces, departments, or municipalities.
- Specific subgroups might involve combinations such as age \times sex \times ethnicity \times immigration status.

In general, if the subgroups are not part of the survey design domains, their sample size is not planned in advance and will therefore be random, which increases the uncertainty of the direct estimate.
Parsimonious solution

The general idea is to support estimation within existing relationships in other areas by fitting statistical models that establish a connection between subgroups of interest through complementary information present in censuses, administrative records, or remote sensing data sources.

Confronted with the challenge of producing accurate estimates in small domains, models emerge as an alternative for estimating indicators of interest, even in domains where the sample size is small or nonexistent.
SAE models for labor market indicators:
A non-comprehensive review
Molina, Saei and Lombardía (2007)
Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model

They proposed a methodology for estimating unemployment or employment characteristics in small areas, based on the assumption that the sample totals of unemployed and employed individuals follow a multinomial logit model with random area effects.

They considered the multinomial vector that counts the number of sampled unemployed \(y_1\), employed \(y_2\), and inactive \(y_3\) individuals within each area-sexage group.

\[
\log \left( \frac{p_{dij}}{p_{dij3}} \right) = \mathbf{x}_{di} \mathbf{\beta}_j + u_d \quad j = 1, 2; \ i = 1, ..., 6; \ d = 1, ..., 406
\]
She defined a model-based approach to producing small area estimates of counts for different categories of the Australian labour force based on a multinomial logit mixed model with category specific random effects.

Within each small area there are two correlated random effects, one associated with the employed category and the other associated with the unemployed category.

\[
\log \left( \frac{p_{dij}}{p_{dij3}} \right) = x_{di} \beta_j + u_{dj}
\]

\[
u_d \sim Normal \left( 0, W \right)
\]
López-Vizcaíno, Lombardía, Morales (2013)
Multinomial-based small area estimation of labour force indicators

They proposed small area estimators of labour force characteristics (totals of employed, unemployed and inactive people and unemployment rates) that were derived from a multinomial logit mixed model with independent random effects on the categories of the response vector.

\[
\log \left( \frac{p_{dij}}{p_{dij3}} \right) = x_{dij} \beta_j + u_{dj}
\]

\[
u_d \sim \text{Normal} \left( 0, V = \text{diag}(\phi_1, \phi_2) \right)\]
They proposed small area estimators of labour force derived from a multinomial logit mixed model that included correlated time and area random effects

\[ \log \left( \frac{p_{dij}}{p_{dij}} \right) = x_{dij} \beta_j + u_{dj} \]

It is also assumed that \( u_1 \) and \( u_2 \) are independent, with:

\[ u_1 \sim Normal \left( 0, V_1 = \text{diag}(\phi_1, \phi_2) \right) \]

\[ u_2 \sim Normal \left( 0, V_2 \sim AR(1) \right) \]
A non-informative Bayesian approach
Under revision
ECLAC in collaboration with Franco

We considered a sampling and a linking model in an area-level set up, where the random effects are correlated

$$\log \left( \frac{p_{dij}}{p_{dij}} \right) = x_{di} \beta_j + u_{dj}$$

$$u_d \sim Normal(0, W)$$

In the sampling model, we considered the design-effects in the following way:

$$\tilde{n}_{dij} \sim \frac{p_{dij}(1-p_{dij})}{\nu_{dij}}$$

$$\tilde{y}_{dij} = \tilde{n}_{dij} \ p_{dij} \sim Multinomial(\sum \tilde{y}_{dij}, p = (p_{d1}, p_{d2}, p_{d3}))$$
Cholesky decomposition and prior distributions

Under the Cholesky decomposition, we assume that $L$ is the Cholesky factor of the correlation matrix $\rho$. This way:

$$W = \text{diag}(\sigma_1, \sigma_2) \times \rho \times \text{diag}(\sigma_1, \sigma_2)$$

$$\rho = L \times L'$$

We considered a traditional noninformative approach

$$\beta_j \sim \text{Normal} \ (0, A_j)$$

$$\sigma_j \sim \text{Inverse-gamma} \ (0.00001, 0.000001)$$
parameters {
  matrix[P-1, K] beta;
  vector<lower=0>[P-1] sigma_u;
  cholesky_factor_corr[P-1] L_u;
  matrix[P-1, D] z_u;
}

transformed parameters {
  simplex[P] theta[D];
  matrix[P-1, D] u; // random effect matrix
  u = diag_pre_multiply(sigma_u, L_u) * z_u;
}
model {
    L_u ~ lkj_corr_cholesky(1);
    to_vector(z_u) ~ normal(0, 10000);
    sigma_u ~ inv_gamma(0.0001, 0.0001);

    beta[p-1, k] ~ normal(0, 10000);
    target += multinomial_lpmf(hat_y[d, ] | theta[d, ]);
}
Benchmarking

Once the chains have reached convergence, we use the direct estimates at the national and regional levels to benchmark the small area estimates. This process is carried out for each iteration of the MCMC process.

Further research should be conducted regarding the benchmarking process and how it can take into account the model-based variance.
Implementation of the model in LAC region
ECLAC’s Household Survey Data Bank (BADEHOG)

This is a repository of household surveys from 18 Latin American countries maintained by the Statistics Division.

• In the case of Chile, the 2017 National Social and Economic Survey (CASEN survey) corresponds to a representative sample at the national, regional, urban, and rural levels.
• For Colombia, the Great Integrated Household Survey of 2018, which is representative of the national, urban, rural, and regional levels, along with departments and their capitals, was used.
• In the case of Peru, the 2017 National Household Survey (ENAHO), which is representative of the national, urban, rural, and departmental levels, was considered. Table 1 shows a comprehensive summary of the household surveys used in this system.
<table>
<thead>
<tr>
<th>Country</th>
<th>Survey</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>Permanent Household Survey (EPH)</td>
<td>2019</td>
</tr>
<tr>
<td>BOL</td>
<td>National Household Survey</td>
<td>2020</td>
</tr>
<tr>
<td>BRA</td>
<td>National Survey by Continuous Household Sample</td>
<td>2020</td>
</tr>
<tr>
<td>CHL</td>
<td>National Socioeconomic Characterization Survey (CASEN)</td>
<td>2020</td>
</tr>
<tr>
<td>COL</td>
<td>Large Integrated Household Survey</td>
<td>2020</td>
</tr>
<tr>
<td>CRI</td>
<td>National Household Survey (ENAHO)</td>
<td>2020</td>
</tr>
<tr>
<td>DOM</td>
<td>National Continuous Labour Force Survey (ENCFT)</td>
<td>2020</td>
</tr>
<tr>
<td>ECU</td>
<td>National Survey on Employment, Unemployment and Underemployment (ENEMDU)</td>
<td>2020</td>
</tr>
<tr>
<td>GTM</td>
<td>National Survey on Living Conditions</td>
<td>2014</td>
</tr>
<tr>
<td>HND</td>
<td>Multipurpose Household Survey</td>
<td>2019</td>
</tr>
<tr>
<td>MEX</td>
<td>National Household Income and Expenditure Survey (ENIGH)</td>
<td>2020</td>
</tr>
<tr>
<td>NIC</td>
<td>National Household Survey on Living Standard Measurement</td>
<td>2014</td>
</tr>
<tr>
<td>PAN</td>
<td>Multipurpose Survey</td>
<td>2019</td>
</tr>
<tr>
<td>PER</td>
<td>National Household Survey - Living Conditions and Poverty</td>
<td>2020</td>
</tr>
<tr>
<td>PRY</td>
<td>Continuous Permanent Household Survey (EPHC)</td>
<td>2020</td>
</tr>
<tr>
<td>SLV</td>
<td>Multipurpose Household Survey</td>
<td>2020</td>
</tr>
<tr>
<td>URY</td>
<td>Continuous Household Survey</td>
<td>2020</td>
</tr>
</tbody>
</table>
ECLAC’s census data bank (CELADES’)

This is a repository of LAC censuses maintained by the ECLAC Population Division (CELADE), which has pursued its ongoing activity of disseminating the census information derived from its data bank.

Most Latin American countries have shared their databases to CELADE, allowing the availability of microdata from the previous and current rounds of censuses.

We used the software Redatam for statistical processing specialized in microdata of population and housing censuses developed by CELADE.
<table>
<thead>
<tr>
<th>Country</th>
<th>Census</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>National Census of Population, Households, and Housing</td>
<td>2010</td>
</tr>
<tr>
<td>BOL</td>
<td>Population and Housing Census</td>
<td>2012</td>
</tr>
<tr>
<td>BRA</td>
<td>Demographic Census</td>
<td>2010</td>
</tr>
<tr>
<td>CHL</td>
<td>Population and Housing Census</td>
<td>2017</td>
</tr>
<tr>
<td>COL</td>
<td>National Census of Population and Housing (CNPV)</td>
<td>2018</td>
</tr>
<tr>
<td>CRI</td>
<td>X National Census of Population and Housing</td>
<td>2011</td>
</tr>
<tr>
<td>DOM</td>
<td>IX National Census of Population and Housing</td>
<td>2010</td>
</tr>
<tr>
<td>ECU</td>
<td>VII Population and Housing Census</td>
<td>2011</td>
</tr>
<tr>
<td>GTM</td>
<td>XII National Census of Population and VII Housing Census</td>
<td>2018</td>
</tr>
<tr>
<td>HND</td>
<td>XVII Population and Housing Census</td>
<td>2013</td>
</tr>
<tr>
<td>MEX</td>
<td>Population and Housing Census</td>
<td>2020</td>
</tr>
<tr>
<td>NIC</td>
<td>Population and Housing Census of Nicaragua</td>
<td>2005</td>
</tr>
<tr>
<td>PAN</td>
<td>Census 2010</td>
<td>2010</td>
</tr>
<tr>
<td>PER</td>
<td>XII Census of Population, VII Housing Census, and III Indigenous Communities Census</td>
<td>2017</td>
</tr>
<tr>
<td>PRY</td>
<td>II National Indigenous Census of Population and Housing</td>
<td>2002</td>
</tr>
<tr>
<td>SLV</td>
<td>VI Population Census and V Housing Census</td>
<td>2007</td>
</tr>
<tr>
<td>URY</td>
<td>Population Census</td>
<td>2011</td>
</tr>
</tbody>
</table>
Satellite Imagery

We access this information through Google Earth Engine, which provides facilities to analyze and obtain this data through the Javascript and Python programming languages, and recently since 2021 in R with the rgee package.

- Among the main advantages of information based on remote sensing is the ease of access to data with deep geographic coverage that is impossible to obtain by traditional means such as surveys or administrative records.

- Data panels can be built at a low marginal cost of variables as diverse as night lights, rainfall, wind speed, floods, topography, forest cover, types of crops, urban development, kind of road services, among many other variables that can be proxies for different economic aspects.
Some preliminary maps
Use with caution
Teamwork

Rolando Ocampo
Director of the Statistics Division

Xavier Mancero
Chief of the Social Statistics Unit

Stalyn Guerrero
Consultant and R expert

Carolina Franco
Principal Statistician at NORC
Thank you for your attendance!