ECUADOR HIGH-FREQUENCY PHONE SURVEY (HFPS) - PHASE 2 - VENEZUELANS IN ECUADOR

SAMPLING DESIGN, WEIGHTING, AND ESTIMATION *

Introduction

In 2021 the World Bank carried out the second phase of the High-frequency Phone Survey (HFPS) in Latin America and The Caribbean to assess the socioeconomic impacts of the COVID-19 pandemic on households. In Ecuador, in particular, four survey waves were conducted. One key objective was to characterize the population of Venezuelan migrants and their households, and the differences from the native population.

Sampling low-prevalence populations, like Venezuelan migrants in Ecuador, requires particular strategies to obtain unbiased precise estimates, so it is not necessary to select extremely large samples that would make the survey implementation unfeasible. Therefore, the strategy to identify and sample Venezuelan migrants is different from that used for the overall population.

This document describes the sample design and the weighting procedures for Venezuelan migrants in the four waves that made up the second phase of the HFPS in Ecuador[1]. The resulting estimates represent Venezuelan adult migrants, their households, children 0-4 years of age and children 5-17 in those households.

Wave 1 Sampling Strategy

The major difficulty in sampling rare populations, like Venezuelans in Ecuador, is counting on a good-quality sampling frame so the selected sample does not produce biased estimates. In order to build a frame of telephone numbers of Venezuelan migrants in Ecuador, the World Bank established an agreement with Movistar, the major mobile telephone provider company in Ecuador.

The company generated a list of all cell phone customers in Ecuador who registered regular incoming or outgoing calls from Venezuela, amounting to 65,622 cell phone numbers. A first-phase simple random sample of 11,200 phone numbers was selected from this frame and called to confirm if the owners were indeed Venezuelan and determine if they were willing to participate in the survey.

Of the 11,200 called persons, 4,984 answered the call, and 850 agreed to participate in the study and were confirmed as Venezuelan adults. Out of these, a second-phase sample of 620 was then selected and called, out of which 401 persons finally answered the survey questionnaire.

[1] The sampling design and weighting scheme corresponding to Ecuador´s overall population is described in the document *High-frequency Phone Survey (HFPS), Phase 2. Sampling Design, Weighting and Estimation.*

Wave 1 Weighting

The survey has four units of analysis: Venezuelan adult migrants (18 years of age and older), their households, children 0-4 years of age and children 5-17 in the household.

Sampling weights were computed for each unit and should be used according to the estimate of interest. The weighting process involves five steps:

1. Calculation of the inclusion probabilities of phone numbers.
2. Computation of design weights for households and individuals.
3. Nonresponse weighting adjustment.
4. Calibration of individual, household and child weights.
5. Weight trimming and recalibration.

Step 1: Inclusion probabilities of cell phone numbers

The first-phase inclusion probabilities of cell phone numbers are

$$\pi_{w1(1)i} = \frac{n_{w1(1)}}{N_M}$$

where

$\pi_{w1(1)i}$ is the wave 1 first-phase inclusion probability of the i-th cell phone number;

$n_{w1(1)}$ is the size of the wave 1 first-phase sample of cell phones; and

N_M is Movistar frame size.

The second-phase inclusion probabilities of cell phones conditional on being selected in the first phase and accepting to participate in the survey are

$$\pi_{w1(2)i|(1)i} = \frac{n_{w1(2)}}{n_{w1(1)}}$$

where

$\pi_{w1(2)i|(1)i}$ is the second-phase inclusion probability of the i-th cell phone number conditional on being selected in the first phase;

$n_{w1(2)}$ is the size of the second-phase sample of cell phones;

$n_{w1(1)}$ is the size of the first-phase sample of cell phones.

The final inclusion probabilities of cell phones are the product of their first-phase inclusion probabilities and the conditional second-phase inclusion probabilities.

$$\pi_{w1i} = \pi_{w1(1)i}\ \pi_{w1(2)i|(1)i}$$

Step 2: Design weights for households and individuals

The selection probabilities of households and individuals aged 18 years and older are based on the inclusion probabilities of the cell phones through which they can be reached.
The probability of an individual being selected equals the inclusion probability of his or her cell phone number. On the other hand, the selection probabilities of households with more than one cell phone should account for multiple chances of selection. Households with more than one cell phone are over-represented and their selection probabilities need to be adjusted to account for this increased chance of selection. The multiplicity-adjusted household selection probabilities in each frame are computed as

$$\pi_{w1j} = m_j \pi_{w1i}$$

where

π_{w1j} is selection probability of the j-th household, adjusted for multiplicity of cell phones in the household, and

m_j is the number of working cell phones in the j-th household.

Finally, individual and household design weights, w_{0k} and w_{0j} respectively, are the inverse of the corresponding selection probabilities

$$w_{0k} = \pi_{w1i}^{-1}$$

$$w_{0j} = \pi_{w1j}^{-1}$$

Step 3: Nonresponse adjustment

When a phone number is called, it is not always possible to carry out an interview. Nonresponse occurs because of a number of constraints. Most common are that nobody answers the call (no contact), or the respondent is unwilling to cooperate (refusal).

The weighting nonresponse adjustment was based on the inverse of the weighted response rate estimate, which is the ratio of the sum of the design weights of all units (respondents and nonrespondents) to the sum of the design weights of respondents.

$$a_k = \frac{\sum_{k \in c,R} w_{0k} + \sum_{k \in c,NR} w_{0k}}{\sum_{k \in c,R} w_{0k}}$$

$$a_j = \frac{\sum_{j \in c,R} w_{0j} \; + \; \sum_{j \in c,NR} w_{0j}}{\sum_{j \in c,R} w_{0j}}$$

where a_k is the nonresponse adjustment factor for responding individuals and a_j is the nonresponse adjustment factor that should be applied to responding households. *R* and *NR* indicate the responding and nonresponding units, respectively.

Step 4: Calibration of individual and household weights

Finally, the weights of the responding individuals, households and children were calibrated using a raking method to reflect the dictributions of Venezuelan migrants estimated by the 2019 Survey "Encuesta a Personas en Movilidad Humana y Comunidades Receptoras (EPEC)".

The variables involved in the calibration of weights for each unit of analysis were as follows:

- Households: region (Pichincha province, Guayas province, rest), household head´s educational attainment and household size;
- Individuals: region, sex, age, educational attainment;
- Children 0-4 years: region, sex;
- Children 5-7 years: region, sex, age (5-12, 13-17).

Step 5: Weight trimming and recalibration

Following calibration, the top weights were trimmed to reduce excess variation in the final weights introduced by multiplicity adjustments and calibration, thus mitigating the inflation of standard errors due to weighting.

Lastly, the trimmed weights were calibrated again to the same EPEC totals of Venezuelan migrants used in the first calibration.

Wave 2

For Wave 2, an additional 9,100 numbers were randomly selected from Movistar frame to expand the number of individuals willing to participate in the survey. From this new selection, 513 Venezuelan migrants were identified as willing to participate in the study.

From these migrants and the ones that had agreed to participate but were not called in Wave 1, a fresh supplement sample of 700 was randomly selected and called. 320 of these new sample migrants answered the questionnaire. Their weighting procedure was analog to that that described for Wave 1.

The 401 Wave 1 respondents were also called in Wave 2, and 260 responded. In this case, an additional weight adjustment was introduced to account for the attrition nonresponse between both waves.

This attrition nonresponse weighting adjustment of the panel cases involved the following steps.

1. Identify a set of variables collected in Wave 1 as potential predictors of response in Wave 2.
2. Examine any missing data patterns in those preselected variables.
3. Impute any item missing values in the potential predictors identified in Wave 1 using a sequential regression imputation procedure.
4. Fit a weighted response propensity logistic regression model with the preselected variables from Wave 1 as predictors and a dummy response indicator in Wave 2 as the outcome variable. Estimate a response propensity score for each respondent and nonrespondent unit in Wave 2. Stratify all sample units (respondents and nonrespondents) based on their response propensity scores to create equal-sized adjustment classes.
5. Adjust the weights of the respondent sample units using the inverse of the median propensity in each adjustment class.

The estimated response propensity $\hat{\phi}_i$ can be written as

$$\hat{\phi}(\boldsymbol{x}_i) = \frac{exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})}{1 + exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})}$$

where

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the vector of *p* predictors for unit i considered for the model; and

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ allude to the estimated logistic regression coefficients corresponding to the *p* predictor variables.

The predictors included in the model were sex, age and educational level of the respondent, geographical region, urban-rural area, household size, and overcrowding (whether the household had three or more people per bedroom).

The propensity model fitting and the subsequent nonresponse adjustment were made separately for individuals and households.

Finally, the weight calibration, trimming and recalibration were done on the entire sample, composed of the new supplement cases plus the panel cases.

Waves 3 and 4

Given the cumulative attrition and the decreasing sample size, in waves 3 and 4 it was decided to call all the previously selected phone numbers again regardless of their response status in the previous waves. Under this strategy, 445 complete interviews were obtained in Wave 3 and 356 in Wave 4.

Since all the ever-selected phone numbers were called, and not only those that had responded in the previous waves, no attrition adjustments were needed and the weighting procedures were equivalent to those applied in Wave 1.