Bridging Data Gaps: Social Media and Traditional Sources in Assessing Migration in Latin America

Introduction

While web social media data offers a real-time alternative for assessing migrant stocks and their sociodemographic characteristics, it has inherent limitations. Several studies have evaluated the biases of social media data, particularly Facebook, by comparing it with high-quality traditional data sources such as censuses and administrative records. These studies have highlighted the potential and the limitations of using digital data to capture migration dynamics. For instance, Zagheni et al. (2017) analyzed Facebook's advertising data to estimate migrant stocks across the United States by comparing figures on monthly expat users to the American Community Survey data. Similarly, Spyratos et al. (2019) used Facebook to estimate international migration from several large diasporas comparing its performance against UNDESA/OECD figures, and Palotti et al. (2020) estimated the international migration of Venezuelans in several destinations providing insights into the discrepancies between digital and official estimates from register data. All these studies have relied on more traditional data sources that enable the assessment of biases and limits of alternative web social media data.

However, in Latin America, with no population register data and where statistical data sources often have their own limitations in measuring migrant populations, establishing a reliable 'gold standard' for comparison is particularly challenging. For example, Gutiérrez et al. (2020) highlight several challenges in using household surveys to measure and characterize migrant populations in the region: (i) insufficient sample sizes for accurate estimates across multiple levels of disaggregation; (ii) outdated sampling frames that fail to capture significant changes in migration flows between censuses; and (iii) underestimation of migrant stocks, either due to the exclusion of collective dwellings in survey designs or migrants concealing their status during interviews. In addition, during the 2020 census wave, several Latin American countries incorporated self-report strategies using telephone and online methods (Argentina 2022, Costa Rica 2022, Chile 2024, Ecuador 2022, Mexico 2020, and Uruguay 2023), making census collection more flexible but also raising concerns about coverage and the digital divide, especially in rural and poor urban communities or among people on the move. And while, there was greater integration of census data with administrative records (e.g., civil registries, social security databases) to complement and enhance traditional census data when in-person data collection faced limitations, other aspects seem to have followed a backward path regarding quality measurement of migrant populations. For example, Brazil (2023) has included exhaustive data on international migration in the extended questionnaire applied only to a sample of the total population, while recently released data for Argentina (2022) show a high share of non-response for country of birth among the foreign-born population. On top of this, preliminary findings point to an increase in census omission and overuse of imputation of population in non-interviewed housing (Del Popolo, 2024). In this context, the lack of consistent, high-quality traditional data across Latin America makes it difficult to establish a reliable gold standard for measuring migration. As a result, web social media data emerges as a relevant alternative, potentially facing similar limitations but offering a timely and flexible supplement to existing sources that can enable exercises of nowcasting migrant populations.

To further this discussion, this paper compares figures on expat users of Meta web social media (Facebook, Instagram, and Messenger) with census and household survey data from a selection of Latin American countries. Using the data from household surveys and population censuses, we assess the correlation between the number of absolute migrants, disaggregated by place of birth and sex, and analogous data on daily user data from Meta platforms, disaggregated by previous residence. We focus on Latin American countries with both annual labor force surveys that include questions to identify migrant populations and recent census data. At present, these criteria are met by Argentina (2022), Ecuador (2022), and Mexico (2020).

Recently, Varona et al. (2024) and Montiel (2024) conducted similar assessments of Facebook data against the Mexican census and household surveys for several Latin American countries, respectively. Nevertheless, this is the first simultaneous comparison for the same period where three types of data sources are available, which could provide further evidence to understand how much the direction and magnitude of bias of web social media data varies when we change the standard of comparison. In addition, our analysis here is extended to the whole population, which allows us to understand whether the discrepancy between the different data sources is migrant-specific.

This comparative approach contributes to a broader understanding of the validity and limitations of digital data in migration research for Latin America. It also contributes to the discussion on the general measurement of migrant populations based on novel data from household surveys and the 2020 Census round, while problematizing the idea of a gold standard for traditional statistics, at least in the assessment of migrant populations.

Data and methods

For this analysis, we rely on three types of data sources. First, the most recent wave of published census data, which is available for a limited number of countries—Argentina, Ecuador, and Mexico[1]. For Argentina, we use published tabulations, for Ecuador, published microdata, and in the case of Mexico, we use a 10% sample of the total census because the full sample is only accessible through the INEGI computing center. All are *de jure* censuses conducted over several months and weeks. Second, we incorporate household or labor force survey microdata for the same countries, selecting the trimester that overlaps with the census. From both, census and household survey data we use the information on the total population and the population by place of birth, broken down by sex. Third, we utilize data extracted from the Facebook API on the daily number of Facebook, Messenger, and Instagram users (here on referred to as Facebook data), disaggregated by place of previous residence, gender, and country of current residence. These extracts cover 13 Latin American and Caribbean countries of origin for various destinations in the region, though our analysis focuses on Argentina, Ecuador, and Mexico. To estimate a unique number of users by country of previous residence, aligned with the census enumeration period (Table 1), we calculate a median value from multiple weekly extracts in each destination, using bootstrap estimation to generate confidence intervals around the median.

Table 1. Period of reference for used data sources

Country	Census period	Household Survey period	Facebook period
Argentina	Mar 16-May 18, 2022	Second quarter of 2022 of Permanent Household Survey (*Encuesta Permanente de Hogares, EPH*).	Mar 16-May 18, 2022
Ecuador	Oct-Dec, 2022	Fourth quarter of 2022 of National Survey of Employment, Underemployment and Unemployment (*Encuesta Nacional de Empleo, Subempleo y Desempleo, ENEMDU*)	Oct-Dec, 2022
Mexico	Mar 2 to 27, 2020	First quarter of 2020 of National Survey of Occupation and Employment (*Encuesta Nacional de Ocupación y Empleo, ENOE*)	Jan-Mar, 2020

[1] The Dominican Republic has very recently published census data for 2022 including information on migrant population and it also possible to access this data through the *Encuesta Nacional de Hograes de Propósitos Múltiples*. Panama also has published aggregated data on population by place of birth but the *Encuesta de Mercado Laboral Telefónica* for 2022 did not include information to identify international migrants by national origin (Montiel, 2024). Though, Paraguay and Uruguay have not released microdata, yet we plan to include it the analysis too for the final version together with the Dominican Republic.

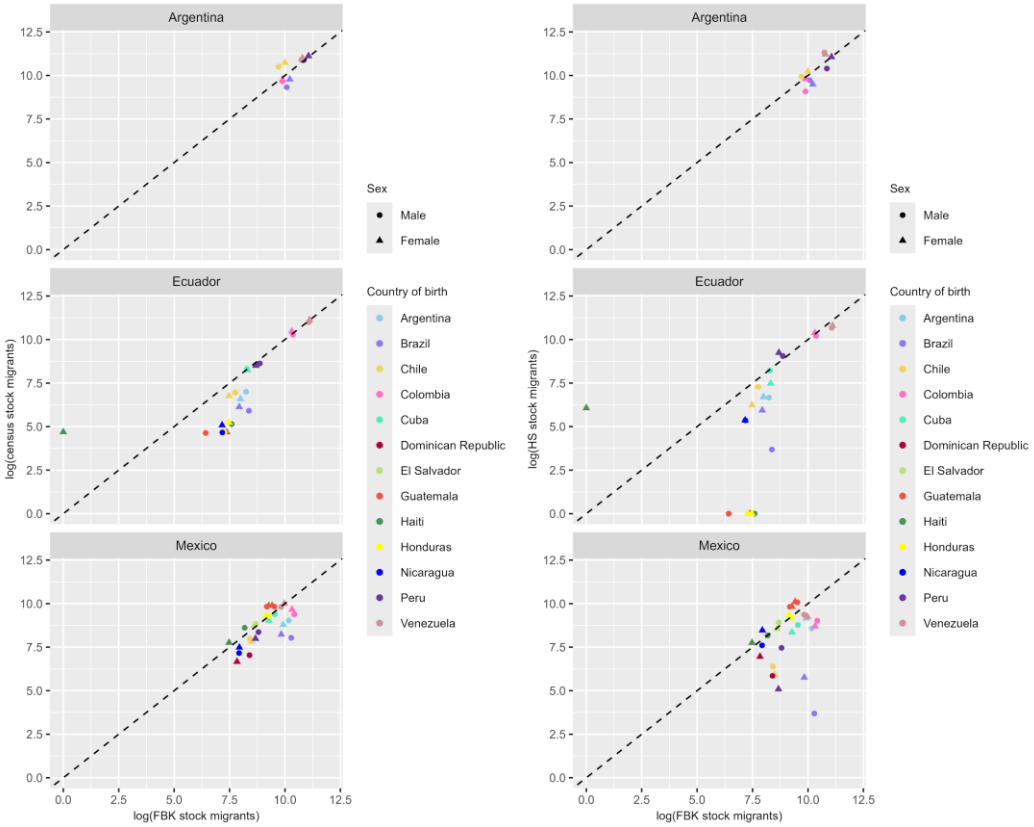
As for the definition of migration used, we limit it to absolute migrants, i.e. people born abroad. This is because not all countries have at the moment published information on population by place of residence and sex on a fixed previous date (five years earlier) that would allow us to work with the census definition of recent migrants which, as Varona et al. (2024) showed, has higher levels of fit with the Facebook data. The analysis focuses on the population between 25 and 65 years of age. This selection is because the information published for Argentina with age disaggregation is not the same as the Facebook data extracts we have made.

Regarding the empirical strategy, we aim to conduct a multivariate analysis following the approach of Zagheni et al. (2017), Montiel (2024), and Varona et al. (2024). First, we plan to predict the total number of migrant populations by sex and origin as reported by the census using figures on the median of Facebook DAU by origin and sex (controls). Second, to replicate this using instead the figures provided by household surveys. In both cases, the country of destination/enumeration can be included as a fixed effect. Third, we plan to compare the results for both exercises using the Chow test. Finally, we plan to replicate this empirical strategy for the whole population and to compare the results analytically with those obtained for migrant populations to identify differences and similarities in the magnitude and direction of the biases.

Preliminary results

Figure 1 shows a general alignment between the population figures of interest obtained from Facebook data, surveys, and censuses, with most of the values closer to the identity line. Generally, Facebook data exhibits a better fit with census figures than with survey data, a pattern particularly evident in Ecuador and Mexico.

Figure 1. Scatter plot for Facebook data against census (left) and household survey data (right) by origin and sex. Argentina, Ecuador, and Mexico, circa 2020

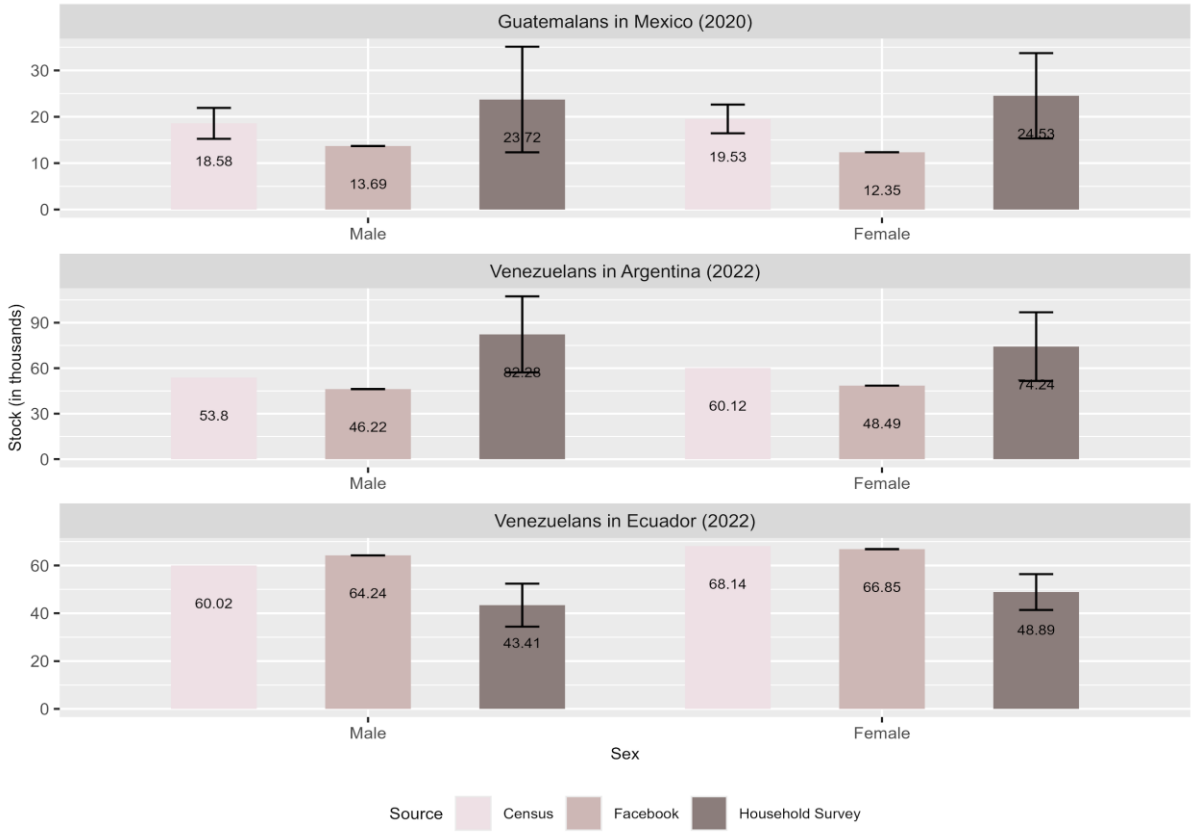


Source: own elaboration based on Facebook API extracts on people previously living in Guatemala (MEX) or Venezuela (ARG and ECU); people born in Guatemala (MEX) and Venezuela (ARG and ECU) reported by national census data (Argentina 2022, Ecuador 2022, Mexico 2020) and ENOE 2020 (Mexico), ENEMDU 2022 (Ecuador), and EPH 2022 (Argentina).

The degree of alignment varies depending on the community of origin; notably, there is a very strong congruence for the Venezuelan community across all three destination countries, regardless of whether Facebook data is compared against census or survey figures. Additionally, differences between sexes are minimal, except in the case of Haitians in Ecuador.

A closer look at the magnitudes from different data sources for selected dyads reveals remarkable similarities between the population estimates from census data and Facebook, while household surveys show less alignment (Figure 2). However, as observed before this pattern varies by origin. Also, in this case, no difference by sex is found but for Venezuelans in Ecuador.

Figure 2. Number of people/users born in/previously living in selected origins according to different data sources. Argentina, Ecuador, and Mexico, circa 2020



Note: These origins were selected as they represent the largest foreign-born group in Mexico, Ecuador, and Argentina[2] respectively. Source: own elaboration based on Facebook API extracts on people previously living in Guatemala (MEX) or Venezuela (ARG and ECU); people born in Guatemala (MEX) and Venezuela (ARG and ECU) reported by national census data (Argentina 2022, Ecuador 2022, Mexico 2020) and ENOE 2020 (Mexico), ENEMDU 2022 (Ecuador), and EPH 2022 (Argentina).

Further work will focus on multivariate analysis, which may provide more insight into the overall pattern by gender and origin. We will also include Dominican Republic (2022), Paraguay (2022), and Uruguay (2023) in the analysis as census and household survey data become available.

[2] The largest origin in Argentina corresponds to Paraguay and Bolivia but these together with Uruguay are some of the origins not reported by Facebook API.

References

Del Popolo, F. (2024). Breve panorama de los censos de población y vivienda 2020 en América Latina y el Caribe y principales desafíos de cara a la ronda. Presentation, Aug 27 2024. Available at 2030https://www.cepal.org/sites/default/files/presentations/panorama_censos_2020_desafios_ronda-2030_cepal-celade_28ag.pdf

Gutiérrez, A., Mancero, X., Fuentes, A., López, F., & Molina, F. (2020). Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional. Serie Estudios Estadísticos, CEPAL, no. 101.

Montiel, C. (2024). Facebook versus encuestas de hogares: oportunidades y límites de los datos de Facebook para el estudio de la migración internacional en América Latina. Tesis de Maestría en Demografía y Estudios de Población.

Palotti J, Adler N, Morales-Guzman A, Villaveces J, Sekara V, Garcia Herranz M, Al-Asad M, Weber I. (2020). Monitoring of the Venezuelan exodus through Facebook's advertising platform. PLoS One. 2020 Feb 21;15(2): e0229175.

Spyratos S, Vespe M, Natale F, Weber I, Zagheni E, Rango M. (2019). Quantifying international human mobility patterns using Facebook Network data. PLoS One. 2019 Oct 24;14(10): e0224134.

Varona, T., Masferrer, C., Prieto Rosas, V., & Pedemonte, M. (2024). Which definition of migration better fits Facebook 'expats'? A response using Mexican census data. *Demographic Research,* 50(39), 1171-1184.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4), 721-734.