Integrating census and residence permit data to estimate annual bilateral migration flows for South America, 1986-2020.

Background

International migration flows are a difficult component of population change to measure. The difficulty comes from the fact that many countries do not produce consistent flow data, and the available data are usually incomplete and incomparable (Willekens et. al., 2016, pp. 897-898). Even the two main data sources, censuses and residence permit data (RPD), are subject to data-source-specific systematic biases and random variation. The limitations of these sources can be overcome by integrating census and RPD to exploit their strengths and compensate for their weaknesses (Bryant & Zhang, 2018).

The strengths of census data lie mainly in their comparability, which comes from the fact that censuses are collected following commonly shared guidelines (UN, 2017, 2008, 1998). Nonetheless, (i) censuses do not measure migration directly, i.e. by counting events; (ii) some of them only provide information on residency 5 years before the census date and, (iii) many censuses are collected approximately every 10 years. These three features entail a need for (i) correcting systematic biases caused by the differences in the census approaches (i.e. *de jure* or *de facto*), omission of infant migrants, migrant deaths and migration of the native-born population, and dissimilarities in census data quality; (ii) translating from five- to one-year transitions to avoid undercounting migrations (events); and, (iii) imputing missing data in the intercensal periods, for which census data are not available.

In the case of RPD, the strengths of these data are associated with their availability, access and production frequency. Nevertheless, using RPD implies (i) dealing with country-specific legislation that may cause considerable cross-national differences in the definition as to who constitutes a migrant, and therefore, who is being counted in national administrative procedures; and also (ii) handling the dissimilarities of national data collection systems (Texidó & Gurrieri, 2013, pp. 75-78). These two characteristics lead to (i) systematic biases due to country-specific timing requirements to obtain a residence permit, undercount of migrants who do not need a residence permit due to bilateral or multilateral free movement agreements; and, overcount of migrants who are part of regularisation processes; as well as (ii) differences in cross-national data quality.

Given the strengths and limitations of the two mentioned data sources, this paper aims at integrating census data and RPD to estimate annual bilateral migration flows through developing a three-level Bayesian hierarchical model. The resulting estimates are the "true" (unobserved) international migration flows amongst countries, where a migrant is tacitly defined as a person whose country of residence differs at the start and end of a year.

We illustrate the use of our model with data reported by the ten biggest South American countries (i.e. Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay and Venezuela), for which the data are much more sparse than in Europe or North America. Thus, there is a need to use cutting-edge methods to deal with the lack of high-quality information on the number of international migrants within, out and into the region.

Data

We draw data from two data sources. The first data source is census microdata of the ten biggest South American countries. The second data source is RPD reported by the Continuous Reporting System on International Migration in the Americas (SICREMI, by its Spanish acronym). Data cover the period from 1986 to 2020.

We consider a migration system composed of 18 origins and 18 destinations. Origins and destinations correspond to Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay, Venezuela, the USA, Canada, Spain and the rest of the world grouped into continents (i.e. America, Africa, Asia, Europe and Oceania)[1]. Each origin *i* and destination *j* pair conforms a migration corridor, where $i \neq j$. Intraregional corridors are composed of only South American country pairs, whereas interregional corridors comprehend corridors where either the origin or the destination is a non-South American country. In total, there are 250 migration corridors (i.e. 90 intraregional corridors + 160 interregional corridors). This implies a dataset of 8.500 entries (i.e. 250 corridors x 34 years).

Research methods

This research is founded on the integrated model that Raymer *et. al.* (2013) proposes for reconciling differences between various measures of migration flows. Generalising Raymer *et.al.*'s model *(2*013, p. 803), migration flows can be conveniently expressed in contingency tables, where rows are origin *i*, and columns indicate destination *j*. There is a contingency table per source *k* and period in time *t*.

$$z_{ijt}^{k} = \begin{pmatrix} 0 & z_{12} & z_{13} & \cdots & z_{1n} \\ z_{21} & 0 & z_{23} & \cdots & z_{2n} \\ z_{31} & z_{32} & 0 & \cdots & z_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & 0 \end{pmatrix}$$

Raymer *et.al.*'s model *(2*013) comprehends two levels: a data model for correcting biases of the data and a migration model for imputing missing data. Our proposed Bayesian hierarchical model involves three levels (see Figure 1). The first level (i) translates from five- to one-year intervals in the case of census data, and (ii) separates permanent and temporary residence permits in the RPD.

The second level of our Bayesian hierarchical model is similar to the data model in Raymer *et.al.* (2013). This second level corrects census and RPD by their systematic biases and accounts for each data source random variation. This level or sub-model corrects census data by (i) standardising values to the most common census approach in South America (i.e. *de facto* perspective), (ii) removing biases due to the omission of infant migrants, migrant deaths and flows of the native-born, and (iii) quantifying differences in census data quality. In RPD terms, the data model standardises the data by accounting for (i) country-specific minimum timing requirements to acquire a residence permit; (ii) data quality; (iii) migrants who do not need a residence permit because of the implementation of

[1] Specific corridors were defined for the USA, Canada and Spain, based on the fact that 18.5% of five-year migrants registered in the 1990, 2000 and 2010 South American censuses moved from these countries to South America. Additionally, the USA, Canada and Spain concentrated the majority of South American migrant stocks of their respective continents in 2019 (UN, 2019). Spain had 55.8% of the total South American migrant stocks of Europe. Likewise, the USA and Canada had 80.5% and 6.4% of the total South American migrant stocks of America, respectively.

bilateral or multilateral free movement agreements; and (iv) unauthorised migrants who are part of national regularisation processes.

The third level of our Bayesian hierarchical model is analogous to the migration model in Raymer *et.al.* (2013), which imputes missing data. As opposed to Raymer *et.al.*'s model, which uses auxiliary variables correlated with migration flows, the third level is defined as an additive mixed gravity sub-model. This sub-model overcomes the dependence of the imputation process on incomplete and unreliable auxiliary data, given that gravity models use widely available data on the population size of origins and destinations, and a measure of the distance between them to predict migration volumes. The additive mixed part of the third level of our hierarchical model enables (i) modelling the non-linear relationship between flows and time, and (ii) capturing the non-uniformly-spaced trends of flows.

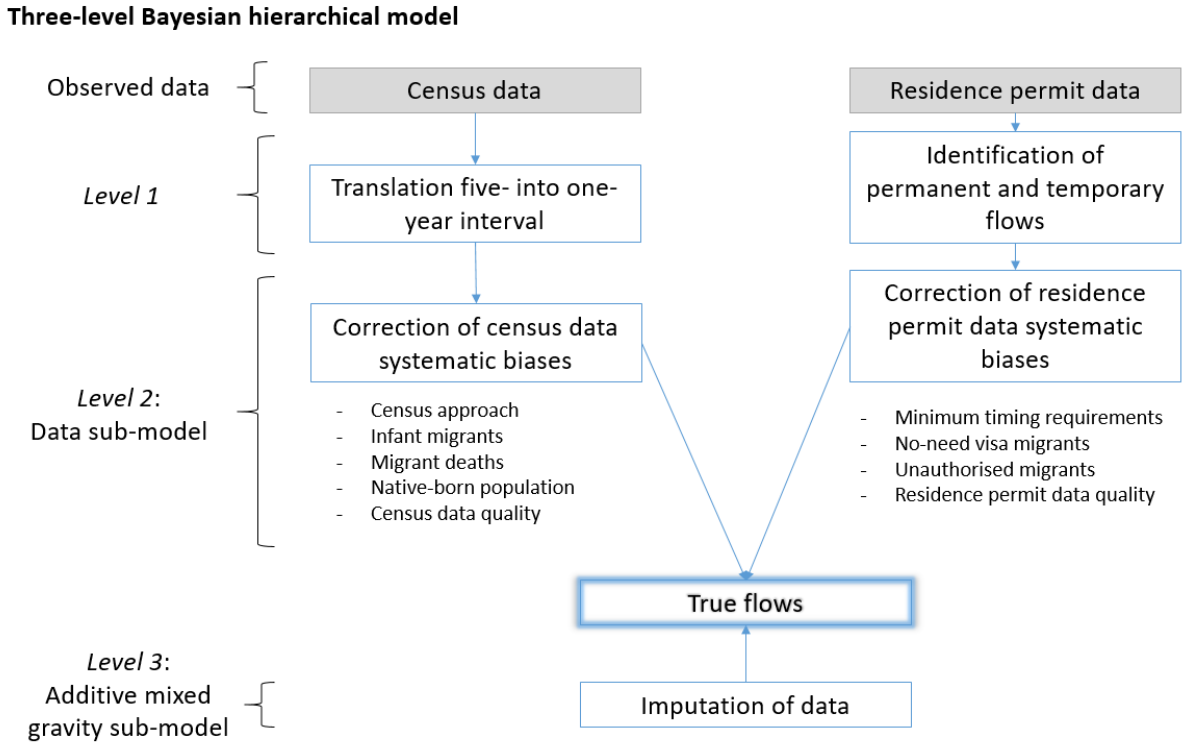

Figure 1. Diagram of our proposed three-level Bayesian hierarchical model to estimate annual migration flows through integrating census and residence permit data (RPD). Grey rectangles represent observed data, while the blue rectangle corresponds to the estimated values.

Expected findings

The output of this study is a set of synthetic estimates of bilateral migration flows with measures of uncertainty for South America from 1986 to 2020. Since there is not a set of gold standard estimates for South America that can be used, we compare our resulting estimates against two sets of existing migration flows. The first set corresponds to the estimates of Abel & Cohen (2019), who calculate them by using the Demographic Account Pseudo Bayesian closed (DAPBC) method. Secondly, we compare our true flows to the number of five-year migrants reported by the Latin American and Caribbean Demographic Center (CELADE, by its Spanish acronym) of the Economic Commission for Latin America and the Caribbean (ECLAC) of the United Nations (2020). For the comparative analysis, we compute Pearson correlation coefficients between sets of flows.

References

Abel, G. J., & Cohen, J. E. (2019). Bilateral international migration flow estimates for 200 countries. *Scientific data*, *6*(1), 1-13.

Bryant, J., & Zhang, J. L. (2018). *Bayesian demographic estimation and forecasting*. CRC Press.

ECLAC (2020), 'Investigación de la Migración Internacional en Latinoamérica (IMILA). Retrieved from: https://celade.cepal.org/bdcelade/imila/

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, *108*(503), 801-819.

Texidó, E., & Gurrieri, J. (2013). Panorama migratorio de América del Sur. *International Organization for Migration.*

United Nations (UN), Department of Economic and Social Affairs, Population Division (1998), Recommendations of statistics of international migration, revision 1, *Statistical Papers, M*(58). Retrieved from: https://unstats.un.org/unsd/publication/seriesm/seriesm58rev1e.pdf

United Nations (UN), Department of Economic and Social Affairs, Population Division (2008), Principles and recommendations for population and housing censuses, Revision 2, *Statistical Papers*, M(67). Retrieved from: https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/PrinciplesandRecommendations/Population−and−Housing−Censuses/SeriesM67Rev2−E.pdf

United Nations (UN), Department of Economic and Social Affairs, Population Division (2017), Principles and recommendations for population and housing censuses, Revision 3, *Statistical Papers* M(67). Retrieved from: https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/PrinciplesandRecommendations/Population−and−Housing−Censuses/SeriesM67rev3−E.pdf

United Nations (UN), Department of Economic and Social Affairs, Population Division (2019). *International migrant stock 2019*. Retrieved from: https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp

Willekens, F., Massey, D., Raymer, J. & Beauchemin, C. (2016), International migration under the microscope, Science, 352(6288), 897–899.