ESTIMATING BILATERAL MIGRATION FLOWS AMONGST SOUTH AMERICAN COUNTRIES THROUGH INTEGRATING ADMINISTRATIVE DATA

Introduction

Migration is the most complex demographic event to estimate, considering that migrating might happen none, one or more times throughout the life of a person (Rowland, 2003; Newell, 1988). It is that difficult that not even the UN provides (probabilistic) estimates regarding migration flows as they do so for fertility and mortality (UN, 2019a). The UN only publishes deterministic values of migration (UN, 2019b), which, most of the time, are the residual after accounting for the natural increase (i.e. births - deaths) between pairs of censuses (UN, 2017:4).

The difficulty of providing migration flow estimates comes from the fact that the available data are usually incomplete and incomparable (Willekens *et. al.,* 2016:897-898; Nowok & Kupiszewska, 2005:15; Rodriguez, 2004:49-51; ECLAC, 1999:417-420). Therefore, we aim at integrating various types of data for estimating international migration flows. We use various types of data that vary due to the definition of what migration and a migrant are, measurement methods, population coverage, systematic bias and accuracy of the data collection mechanism.

These inconsistencies in data are sharper in regions such as South America, in which the data are much more sparse than in Europe or North America, and the data are potentially of lower quality, since there is no common regulation in South America to produce data. 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. In order to overcome the difficulties with South American data, this paper aims to develop a statistical model for estimating bilateral migration flows amongst South American countries through integrating the different types of administrative data.

We build on and advance the methods for measuring and estimating international migration flows developed in recent years. For example, flows in Europe were harmonised by using constrained optimisation (De Beer *et.al.,*2010) and hierarchical modelling (Raymer *et.al.,*2013; Wiśniowski *et.al.,*2013; Wiśniowski *et.al.,*2016; Wiśniowski,2017). Abel &Sander (2014) and Abel (2013) developed a log-linear model to estimate global flows. Bryant &Graham (2013) proposed a Bayesian approach to reconcile different data sources for New Zealand. Other methods have also been developed to forecasting migration (Bijak,2010; Bijak &Wiśniowski,2010; ; Disney *et.al.,2*015:29; Raymer & Wiśniowski, 2018).

Data

This study focuses on the ten biggest South American countries (Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay, and Venezuela) and the period extending from 1990 to 2018. Data are extracted from each of the migration offices of the previous South American countries. Migration offices produce annual data. Table 1 shows the available administrative data on bilateral migration flows reported by South American migration offices. For the years, in which there is no bilateral migration flow data, the total migration is available and those values are used for validation of the final estimates.

Table 1. Available administrative data on bilateral migration flows reported by South American migration offices.

Country	Years
Argentina	2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017
Bolivia	2012, 2016
Brazil	2010, 2011, 2013, 2014, 2015, 2016
Chile	2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017
Colombia	2012, 2013, 2014, 2015, 2016, 2017
Ecuador	1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2012, 2013, 2014, 2015, 2016, 2017
Paraguay	2015, 2016, 2017
Peru	2010, 2011, 2012, 2013, 2014, 2015, 2016
Uruguay	1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014

* There is no public data for Venezuela.

Moreover, administrative data contains three types of statistics: entries/departures, permanent residences and temporary residences. This implies some inconsistencies in the data. Figure 1 delineates the histograms and descriptive statistics of the intra-regional inflows based on the concepts of permanent and temporary residences. While the maximum of permanent dwellers is 1 619 290, the maximum of temporal residents is 1 832 514. This is sensible, since it is more likely to have more migrants who can cover the cost of living abroad for three months than migrants who can deal with the expenses of living overseas for three years.

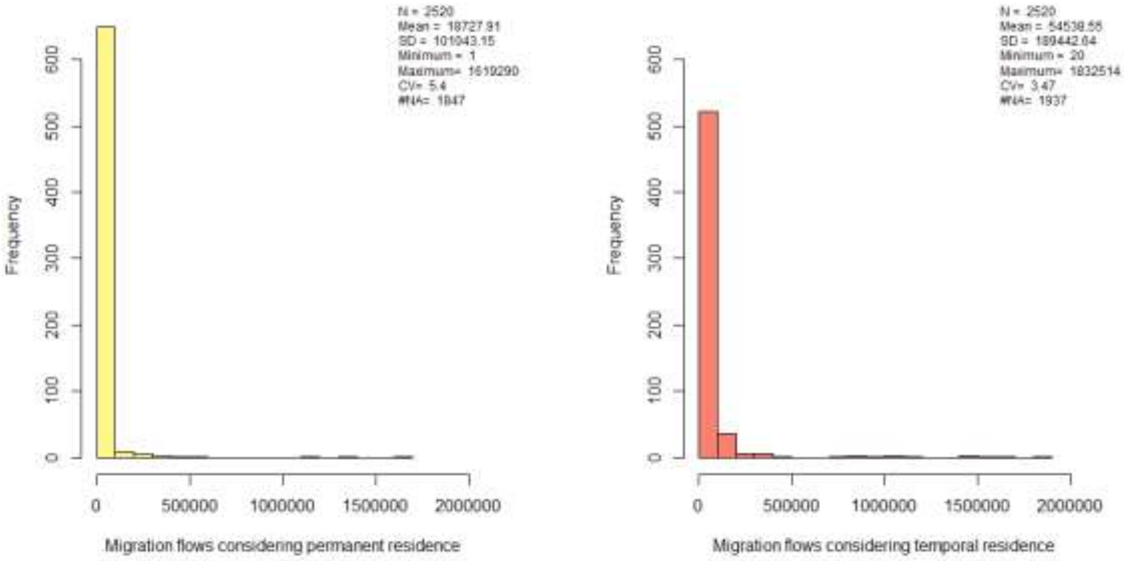

Figure 1. Histograms and descriptive statistics of the observed migration inflows based on permanent (left) and temporary (right) residences.

Administrative data also gives information about immigration and emigration. Figure 2 illustrates the migrants reported by sending and receiving countries (i.e. the number of emigrants and immigrants, respectively). It is clear that receiving countries usually report more migrants than sending countries (i.e. emigrants are usually undercounted).

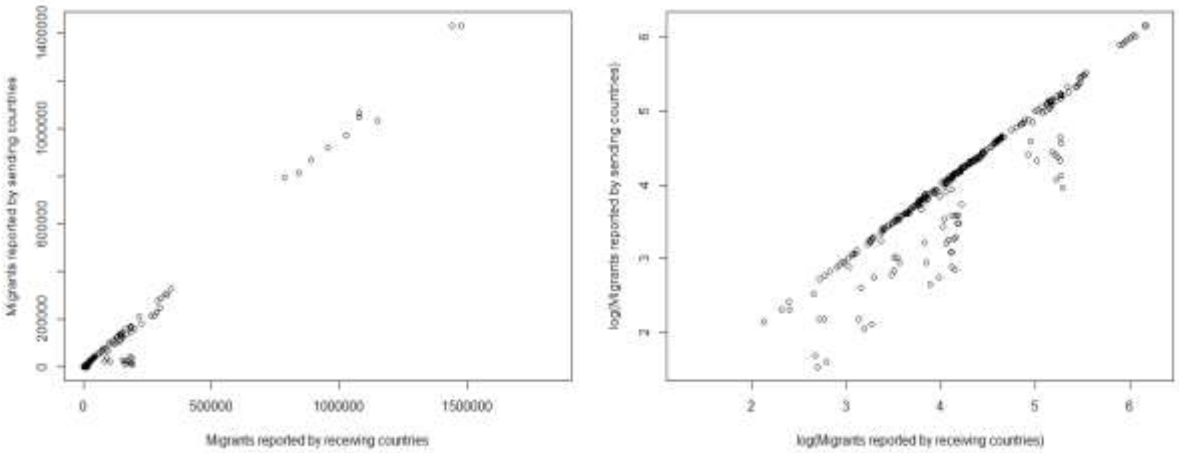

Figure 2. Migration flows (left) and logarithmic transformation (based 10) (right) of migrations flows reported by sending and receiving countries based on temporary residences.

Additionally, most of administrative data are generated by using citizenship. Following Nowok & Kupiszewska (2005:10), citizenship is the most common approach by which countries count migrants. However, some countries such as Colombia and Chile also identify migrants by their las place of residence. This implies inconsistencies in the data. This is evident in the flow from Bolivia to Colombia in 2017[1]. There were 31668 non-national migrants, whereas there were 7941 resident migrants.

It must be mentioned that administrative data are used instead of census data, given that besides the common census data limitations (Kupiszewska and Wisniowski, 2009:6; Week, 2007:267; Newell, 1988:87; CELADE, 1991:25-26), South American censuses do not fully comply with the UN recommendations (2017, 2008, 1998) relating to censuses (e.g. some South American census are taken more than 10 years apart). Further work will include alternative/new types of data (e.g. geotagged social media data).

Methods

This research is founded on the works of Raymer *et. al.* (2013) and Wiśniowski *et. al.* (2013) for reconciling differences between various measures of migration flows. Generalising the Raymer *et.al.*'s model *(2*013:803), 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*.

[1] The information for this flow is reported by the Colombian migration office (Migration Colombia, 2017).

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

Accepting that counts on migrants distribute $\mathrm{z}_{\mathrm{ijt}}^{\mathrm{k}} \sim Poisson(\mu_{ijt}^{k})$[2] and z_{ijt}^{k} have measurement errors, a component which represents the true (unobserved) migration flows y_{ijt} is defined. Considering that the component y_{ijt} is not directly observed, a measurement error model is required for correcting data inadequacies (Buonaccorsi, 2010:1, Gustafson, 2004:11).

$$\log(\mu_{ijt}^{k}) = \log(y_{ijt}) + \omega_{f(k)} + \boldsymbol{\theta}_{g(k)} + \boldsymbol{\delta}_{m(k)} + \boldsymbol{\lambda}_{n(k)} + \zeta_{r(k)} + \varepsilon_{ijt}^{k}$$

where $\log(y_{ijt})$ is the true flows, which captures the temporary residences[3], sorts migrants by their last place of residence, embodies the minimal duration of stay of 12 months, assume that neither immigrants nor emigrants are undercounted, has the accuracy of Ecuadorian data (which is the best one), and refers to migration performed in a year *t*.

Moreover, $\omega_{f(k)}$ captures whether the data refer to entries/departures or permanent residences, $\boldsymbol{\theta}_{g(k)}$ represents if migrants are classified by citizenship, $\boldsymbol{\delta}_{m(k)}$ embodies the minimum duration of residence, $\boldsymbol{\lambda}_{n(k)}$ includes undercounting of immigrants or emigrants, $\nu_{p(k)}$ signifies if the data come from tertiary sources, $\zeta_{r(k)}$ is related to level of coverage of the data, and $\varepsilon_{ijt}^{k} \sim N(0, \boldsymbol{\beta}_{q(k)})$ is the random errors, where $\boldsymbol{\beta}_{q(k)}$ captures the accuracy of the data. The posterior distribution of the $\log(y_{ijt})$ will represent the final synthetic data used as outputs for the model for estimates of migration flows over time with associated uncertainty.

Expected results

The resulting outcome is a set of synthetic annual estimates of migration flows with measures of uncertainty for South American countries from 1990 to 2018. Further contribution is the extension of Raymer *et. al.* (2013) and Wiśniowski *et. al.* (2016) models, which have been developed for other regions with more abundant data.

[2] μ_{ijt}^{k} is the expected number of migrants from country *i* to country *j* at time *t*, as reported by source *k*.

[3] Temporary residences are the closest definition to the long-term migrant concept of the UN (1998).