

Probabilistic method for combining internal migration data

G. Abel^{*2,1}, G. Vinué^{†1}, D. Yildiz^{‡1}, A. Wisniewski^{§3}, L. Fiorio^{¶4}, and J. Cai^{||5}

¹Wittgenstein Centre for Demography and Global Human Capital (International Institute for Applied Systems Analysis, Vienna Institute of Demography/Austrian Academy of Sciences, Vienna University of Economics and Business), Austria

²Asian Demographic Research Institute, Shanghai University, China

³Cathie Marsh Institute for Social Research, School of Social Sciences, University of Manchester, United Kingdom

⁴University of Washington, Seattle, USA

⁵Chinese University of Hong Kong, Hong Kong, China

Abstract

In order to fully understand the causes and consequences of population movements, researchers and policy makers require timely and consistent data. Migration data are commonly obtained from censuses, registers or surveys. Each of these data sources can vary in their measurement of accuracy, coverage of population, undercount and definitions of a migration event. This paper proposes a Bayesian probabilistic methodology to harmonize migration data from different sources. In particular, we build a hierarchical model for combining migration data sources in the USA between 1980 and 2016. The model allows for estimates of true migration flows that explicitly compensates for the inadequacies in each data source and provides one-step ahead forecasts of bilateral migration patterns.

Keywords— Migration forecasting, Combining data, Bayesian, USA, Internal migration

1 Introduction

In order to fully understand the causes and consequences of population movements, and how they evolve over time, researchers and policy makers require timely and consistent data [3]. Data are traditionally obtained from censuses,

*Corresponding Author: guy.abel@oeaw.ac.at

†guillermo.vinue.visus@oeaw.ac.at

‡dilek.yildiz@oeaw.ac.at

§a.wisniewski@manchester.ac.uk

¶fiorio@uw.edu

||caijixuan@link.cuhk.edu.hk

surveys or administrative records. These sources provide estimates of movements with different qualities according to their data collection methods and sample sizes. Moreover, the definition of “migration event” changes both across sources and by time. In order to provide more coherent migration data, statistical methods have previously been utilized to overcome data issues, in particular for international migration [9, 11]. In this paper we propose a Bayesian probabilistic methodology to combine and harmonize internal migration data between nine USA-census divisions between 1980 and 2016. Further, we extend our method to provide one-step ahead forecasts of future migration patterns.

2 Data: USA internal migration

Table 1 presents the available data sources to estimate USA internal migration. Decennial censuses provide information on the place of residence five-years ago. When ACS replaced the long-form questionnaire in decennial census in 2010, the duration of stay criteria in the migration related questions changed to one-year from five-years. Hence, the migration estimates are not directly comparable to those from decennial census ¹ [4].

Source	Years available	Universe	Definition
Census	1940, 1960-2000	Age 5+	State/country 5-years ago
ACS	2000-2015	Age 1+	State/country 1-year ago
CPS	1982-2016 (-1985)	Age 1+	State/country 1-year ago
CPS	1985, 1995, 2005, 2015	Age 5+	State/country 5-years ago
IRS-Tax Records	1991-2015	Tax filers	State/county 1-year ago
SIPP	1993-2001	Individuals	Individual’s residence history

Table 1: Available USA data sources.

The sample size of ACS has changed over time, reached full implementation in 2005, whereas the information on group quarters are added in 2006. The CPS data source has been collecting information on place of residence either one-year or five-years ago since 1982. The CPS migration estimates are also not directly comparable to the ACS estimates since the former only collects information from the civilian non-institutionalized population. The CPS is a voluntary survey while the ACS is mandatory. The IRS provides state to state and county to county one-year migration tables for tax payers and for years between 1991 and 2015. There have been a series of changes in the migration data produced by the IRS, most notably in the population coverage, which varies year on year. The last data source is SIPP, which is a continuous series of national panels. Each panel features a nationally representative sample interviewed over a multi-year period lasting approximately four years. It collects information about the residence histories of the individuals ². In this study we focus on ACS, Census,

¹<http://www.census.gov/topics/population/migration/guidance/state-to-state-migration-flows.html>

²http://www.census.gov/programs-surveys/sipp/about/sipp-content-information.html#par_textimage_9

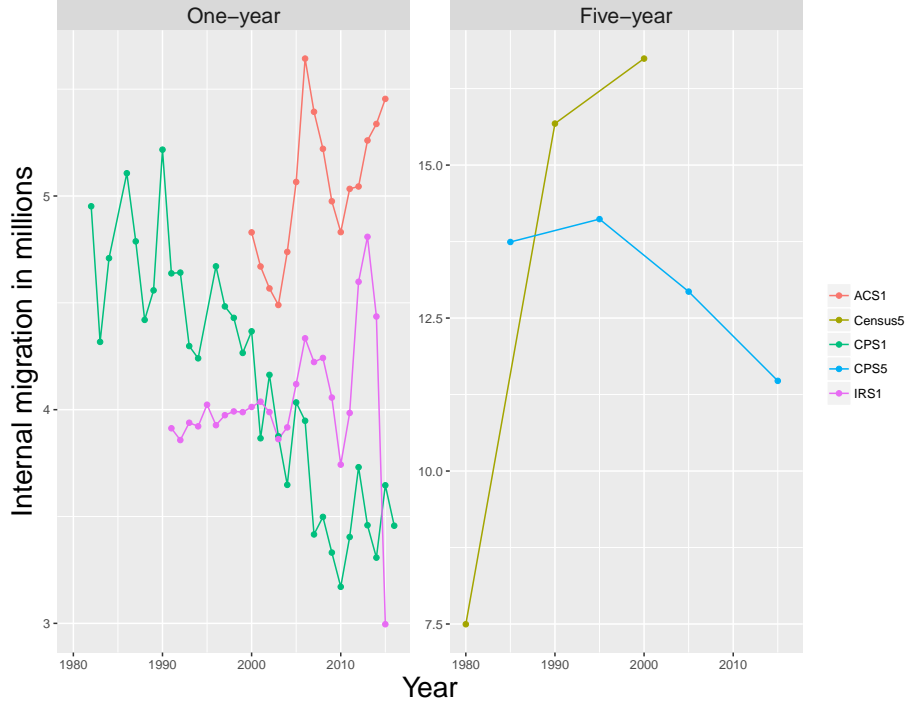


Figure 1: Time series plot of total migration flows between the nine USA-census divisions (1980-2016).

CPS and IRS. A time series plot of total migration flows between the nine USA-census divisions is shown in Fig. 1, during the period of interest (1980-2016). In Fig. 2 we plot the origin-destination migration flows between each of these data sources in 2015 to illustrate the bilateral patterns in this period.

3 Methodology

Migration data of interest can be conveniently represented in a two-way contingency table of origin-destination movements, with cells representing the count of moves over a specified period between the nine USA-census divisions. We observe flow counts z_{ijt}^k from region i to region j during year t reported by data source k and interval status (1-year or 5-years). These flows can be represented by a matrix:

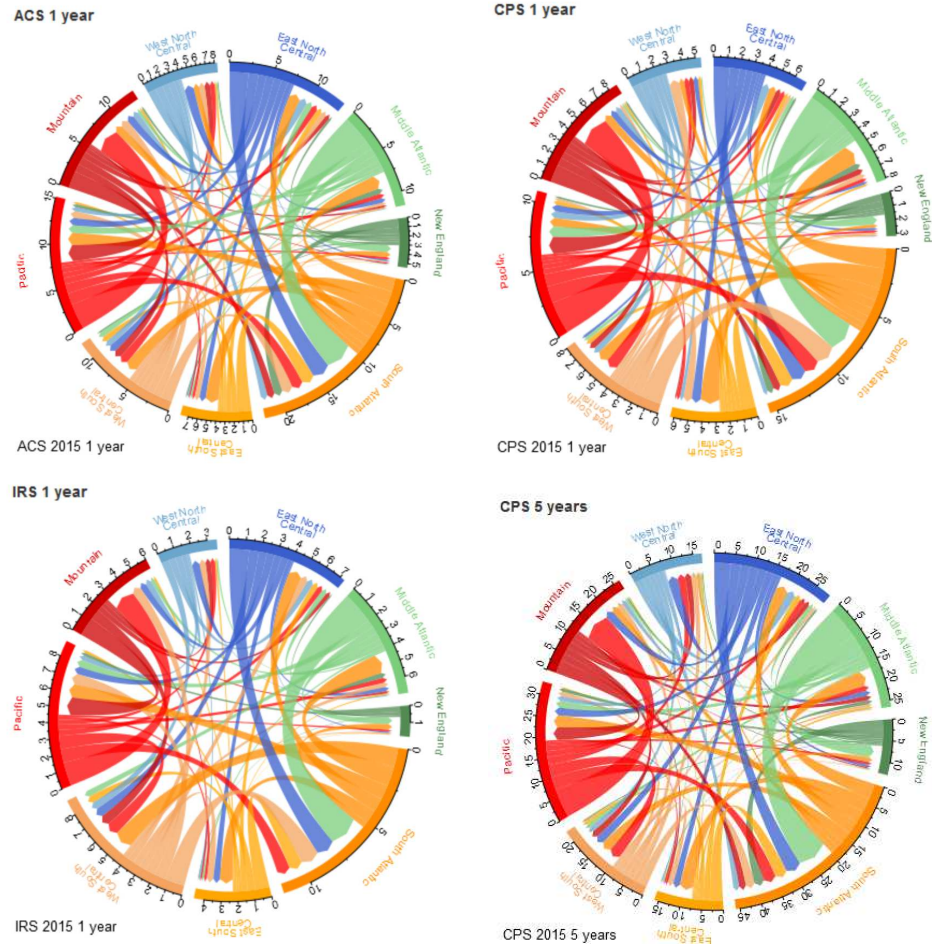


Figure 2: Origin-destination migration flows between each of these data sources in 2015.

$$z_{ijt}^k = \begin{pmatrix} 0 & z_{12t}^k & z_{13t}^k & \dots & z_{19t}^k \\ z_{21t}^k & 0 & z_{23t}^k & \dots & z_{29t}^k \\ z_{31t}^k & z_{32t}^k & 0 & \dots & z_{39t}^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{91t}^k & z_{92tp}^k & z_{93t}^k & \dots & 0 \end{pmatrix}$$

Our objective is to estimate true migration flows for a table for each year between 1980 and 2016 with the diagonals of the tables (i.e. the within-region flows) excluded:

$$y_{ijt} = \begin{pmatrix} 0 & y_{12t} & y_{13t} & \dots & y_{19t} \\ y_{21t} & 0 & y_{23t} & \dots & y_{29t} \\ y_{31t} & y_{32t} & 0 & \dots & y_{39t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{91t} & y_{92tp} & y_{93t} & \dots & 0 \end{pmatrix}$$

Following the general statistical perspective of [11] and [9], we can merge multiple tables of reported flows z_{ijt}^k in a general framework to combine data sources and adjust for data source limitations. This can be expressed as follows:

$$z_{ijt}^k = y_{ijt} \times c^k \times u^k \times d^k \times \tau_{ijt}^k \quad (1)$$

The y_{ijt} component represents the true migration flow from origin i to destination j . The posterior distribution of the y_{ijt} will represent the final synthetic estimates of bilateral migration flows. They will be estimates of migration flows over time with associated uncertainty. The additional parameters in Eq. 1 form a measurement model to link the data properties of the observed data (z) to the true unknown flows. The c^k and u^k parameters reflect the coverage and undercount, respectively, from data source k . These parameters lie between zero and one. The d^k term represents the difference of the duration of stay criteria for movements measured in data source k .

Through analysis of the literature related to each data source³⁴⁵⁶ (see also [10, Table 2], [2, Table A.7], [5, Figure 5], [6, Figure 16-2], [1, Figure 1]), we elicited the following prior distributions for coverage and undercount:

$$c^{ACS} \sim \text{beta}(483.8, 30.8), u^{ACS} \sim \text{beta}(71.5, 2.7)$$

³<https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/>

⁴https://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html

⁵<http://www.nber.org/data/current-population-survey-data.html>

⁶<https://www.irs.gov/statistics/soi-tax-stats-migration-data>

$$c^{Census} \sim \text{beta}(114.9, 9.4), u^{Census} \sim \text{beta}(156.1, 1.7)$$

$$c^{CPS} \sim \text{beta}(127.8, 14.7), u^{CPS} \sim \text{beta}(83.1, 6.5)$$

$$c^{IRS} \sim \text{beta}(638.9, 185.6), u^{IRS} \sim \text{beta}(2501.3, 307.1)$$

The prior distributions for each c^k and u^k are plotted in Figs. 3. The prior parameters for the duration have been derived from the equations of expectation and variance of the log-normal distribution.

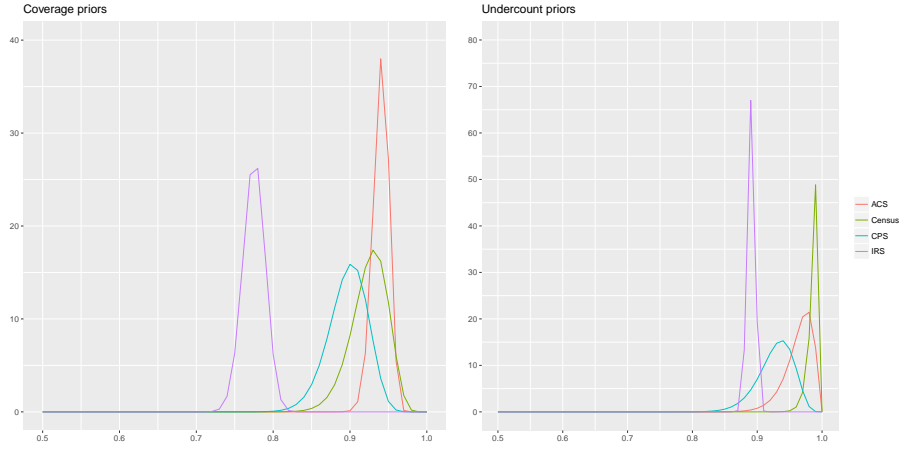


Figure 3: Coverage priors (left) and undercount priors (right) for each data source.

In order to generate true migration flow estimates for past and future periods, we use a set of random effects for both the mean level of each migration corridor and the lagged to capture auto-correlation:

$$\log(y_{ijt}) = \mu + \alpha_{ij}^k + \beta_{ij}^k \times y_{ij(t-1)}^k + \epsilon_{ij}^k \quad (2)$$

where non-informative priors are used for μ , α_{ij}^k and β_{ij}^k . The parameter ϵ_{ij}^k is assumed to be normally distributed with mean zero and precision τ^k , a non-informative prior representing the accuracy of data source k ⁷.

Using Bayesian estimation methods, each item of Eq. 1 can be splitted into components with possible underlying sub-model components. These can be estimated relatively easily, in comparison with other estimation approaches, using Markov chain Monte Carlo (MCMC) methods. The model is run with the JAGS (Just Another Gibbs Sampler) software [7] though R using the the `rjags` package [8].

⁷We are currently developing an informative τ^k for each data source k .

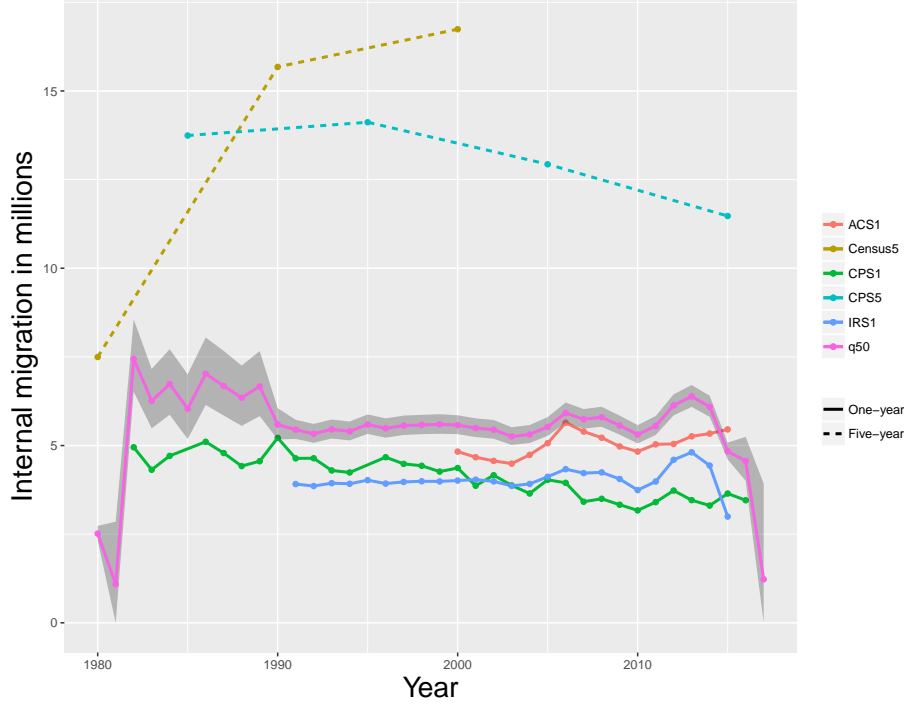


Figure 4: Time series plot of total migration flows between the nine USA-census divisions together with the median one-year flows predicted by the model (1980-2017).

4 Results

A time series plot of total migration flows between the nine USA-census divisions, together with the median one-year flows predicted by our model, is shown in Fig. 4, during the period of interest (1980-2017). The estimates obtained are a bit higher than the observed values because we are controlling for undercount. In addition, we are getting some unexpected behaviour at the beginning and at the end of the time series. We will look closely at this issue.

In addition, Fig. 5 represents two circular plots of median one-year flows for 2016 and 2017 (our forecast). Related to aforementioned issue, the flows between regions for 2017 are smaller than expected.

5 Conclusions

The methodology presented in this paper estimates synthetic bilateral migration flows that borrow strength over multiple data sources. The resulting estimates

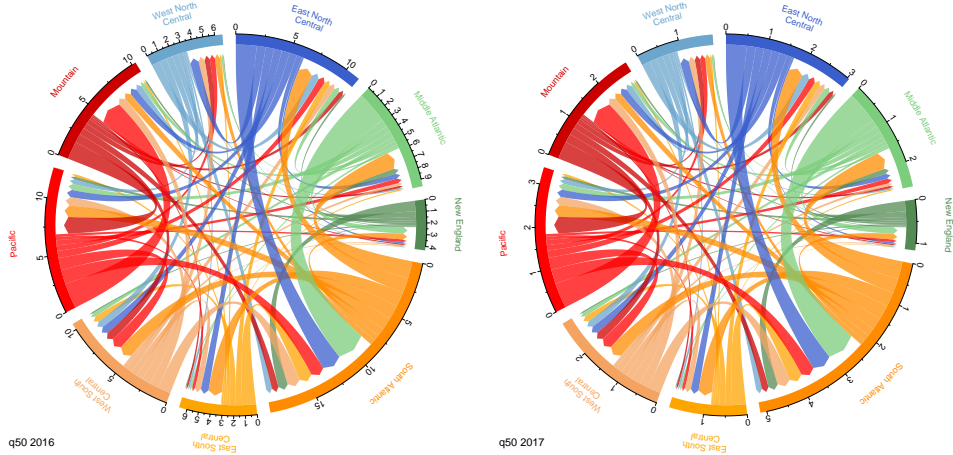


Figure 5: Circular plots of median one-year flows for 2016 (left) and 2017 (right, our forecast).

allow for a better understanding of people’s movement patterns beyond the confines of a single source. The model also utilizes past data to forecast migration in future periods.

6 Future work

Our future work, that we hope to complete in the upcoming months, include: (1) derive and incorporate informative prior distributions for data source accuracy measures. (2) expand the model to incorporate geo-located Twitter data for one-year and three-months migration flows. We have already collected the data and are currently setting up our informative priors for undercount, coverage and accuracy. (3) expand the model to estimate true flows for five-years and three-months intervals (alongside our current one-year estimates).

References

- [1] R.E. Brown and M.J. Mazur. IRS’s comprehensive approach to compliance measurement. Technical report, Internal Revenue Service, 2003.
- [2] J.L. Czajka and A. Beyler. Declining response rates in federal surveys: Trends and implications. Technical report, Mathematica Policy Research, 2016.
- [3] J. De Beer, J. Raymer, R. Van der Erf, and L. Van Wissen. Overcoming the problems of inconsistent migration data: A new method applied to flows in Europe. *European Journal of Population*, 26:459–481, 2010.

- [4] D. Ihrke, W. Koerber, and A. Fields. Comparison of migration data: 2013 American Community Survey and 2013 annual social and economic supplement of the current population survey. <http://www.census.gov/library/working-papers/2015/demo/SEHSD-WP2015-21.html>, 2015. Social, Economic, and Housing Statistics Division (SEHSD) Working Paper Number 2015-21. U.S. Census Bureau.
- [5] A.E. Johnson, S.L. Botman, and P. Basiotis. *Statistics of income: Turning administrative systems into information systems*, chapter Nonresponse in Federal Demographic Surveys: 1981-1991, pages 183–194. Internal Revenue Service. Statistics of Income Division, American Statistical Association, 1994.
- [6] United States. Bureau of Labor Statistics and United States. Bureau of the Census. *Current Population Survey: Design and methodology*. Technical paper. U.S. Department of Labor, Bureau of Labor Statistics, 2006.
- [7] M. Plummer. JAGS. A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, pages 1–8, 2003.
- [8] M. Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2016. R package version 4-6.
- [9] J. Raymer, A. Wisniowski, J.J. Forster, P.W.F. Smith, and J. Bijak. Integrated modeling of European migration. *Journal of the American Statistical Association*, 108(503):801–819, 2013.
- [10] J.D. Williams. *The 2010 Decennial Census: Background and Issues*. DIANE Publishing, 2011.
- [11] A. Wisniowski, J.J. Forster, P.W.F. Smith, J. Bijak, and J. Raymer. Integrated modelling of age and sex patterns of European migration. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 179(4):10071024, 2016.