Probabilistic method for combining internal migration data

Guillermo Vinué
Postdoctoral researcher

Wittgenstein Centre for Demography and Global Human Capital
(IIASA, VID/ÖAW, WU)
Vienna Institute of Demography/Austrian Academy of Sciences

SYSORM 2017
IEMath-GR, Spain

13-15 November 2017

Wittgenstein Centre
FOR DEMOGRAPHY AND
GLOBAL HUMAN CAPITAL

Vienna
Institute of
Demography

ÖAW
AUSTRIAN
ACADEMY OF
SCIENCES

Outline

Migration is a big issue for everybody

Research on migration (I)

- Researchers and policy makers require timely and consistent data to understand the causes and consequences of population movements.
- Data on migration is recorded by asking people their place of residence one or five years prior to the current census or survey.
- Migration data sources can vary in their measurement of accuracy, coverage and undercount of population and definitions of a migration event.
- The availability and reliability of current migration data are not completely optimal.

Research on migration (II)

- Until innovations in data collection methods reduce measurements errors and increase the comparability of data, probabilistic models should be used to produce valid migration data.
- The most effective strategy to produce high-quality data is to use a model that creates a synthetic database.
- A synthetic database results from the combination of both quantitative and qualitative data from different sources and contains estimates of harmonized "true" migration flows.
- We propose a Bayesian hierarchical model for combining migration data sources between nine USA-census divisions between 1980 and 2016 to provide synthetic estimates of true flows with uncertainty.

USA internal migration (I)

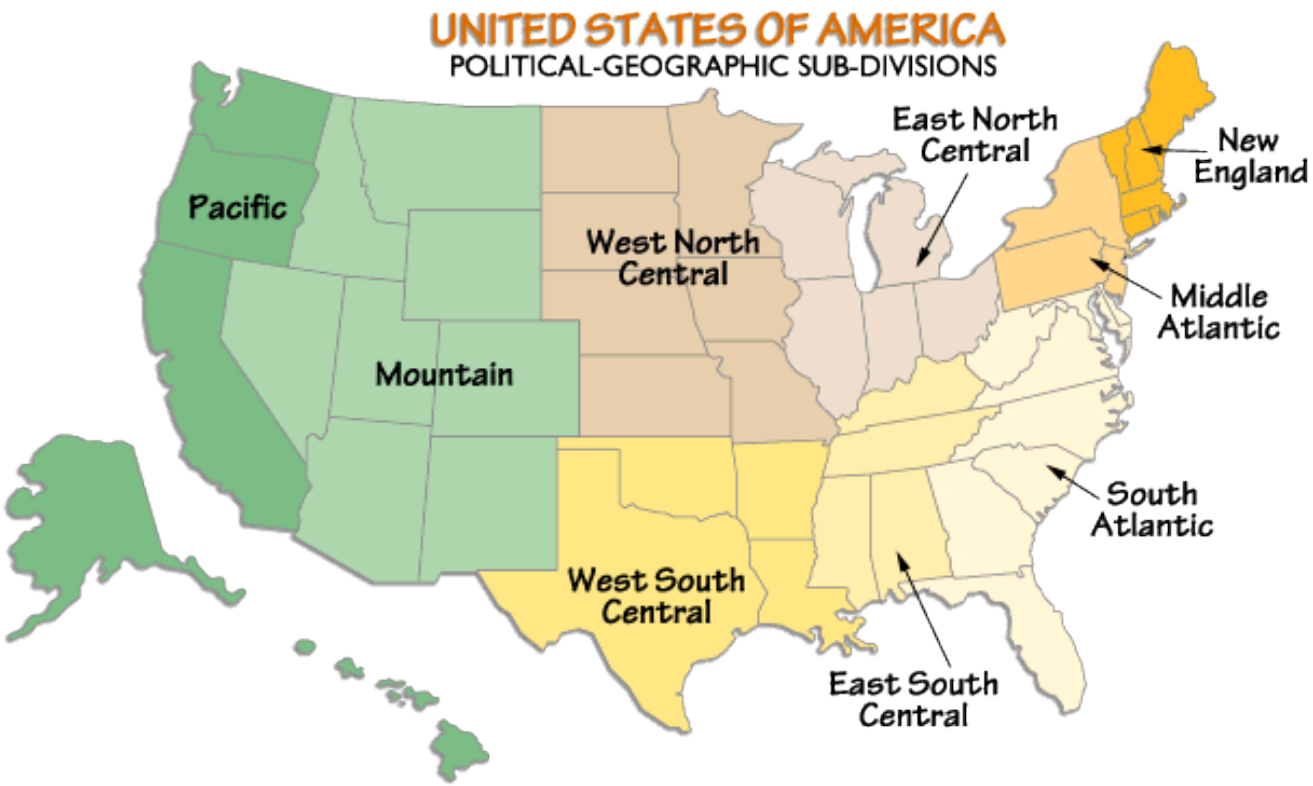
- We focus on USA data because there is plenty of official migration data that can be used to estimate overall population movements.

Data Source	Universe	Duration Migration	Years
Decennial Census	Age 5+	Residence five-year ago	1980, 1990, 2000
American Community Survey (ACS)	Age 1+	Residence one-year ago	2000:2015
Current Population Survey (CPS)	Age 1+	Residence one-year ago	1982:1984, 1986:1994, 1996:2016
Current Population Survey (CPS)	Age 5+	Residence five-year ago	1985, 1995, 2005, 2015
Internal Revenue Service (IRS)	Tax filers	Residence one-year ago	1991:2015

- CPS measures migration over both one-year and five-year periods. Census measures over five-year periods. ACS and IRS over one-year periods.

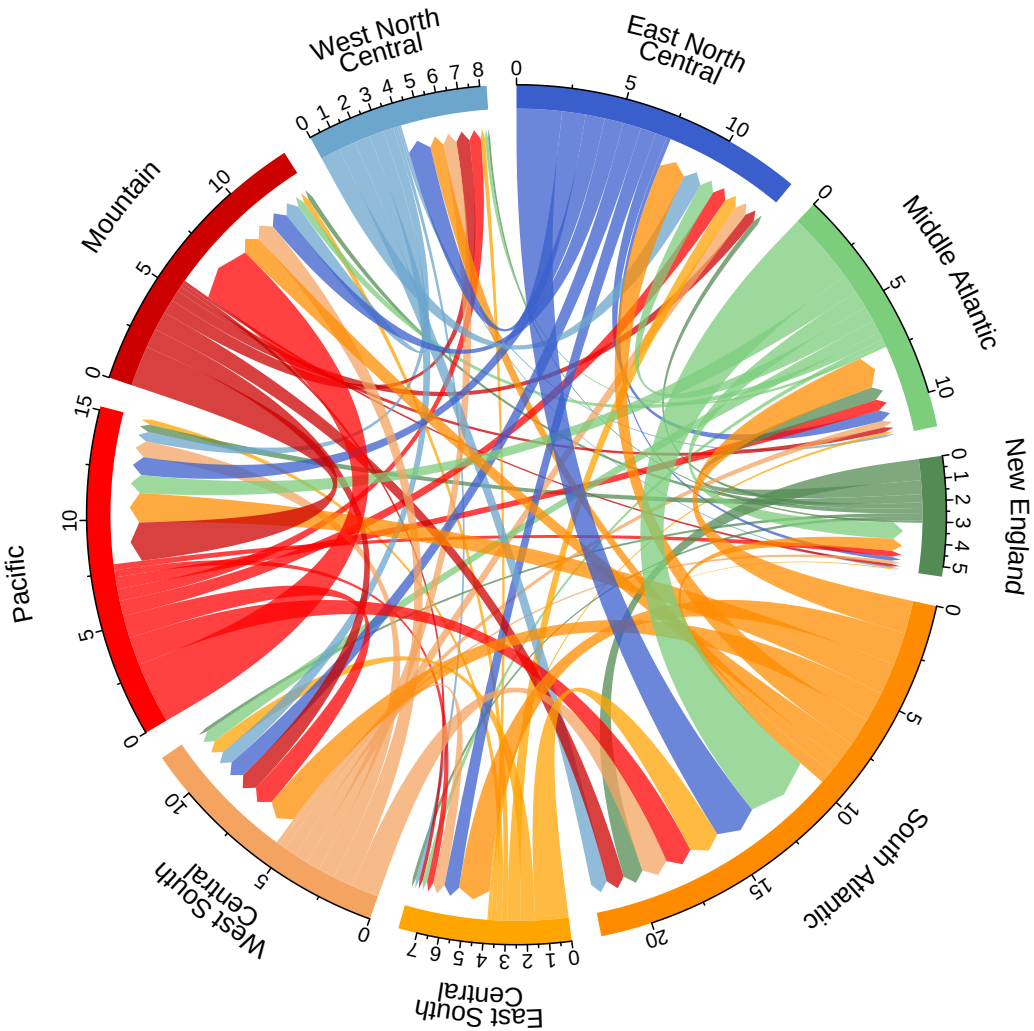
USA internal migration (II)

- Nine USA-census divisions:



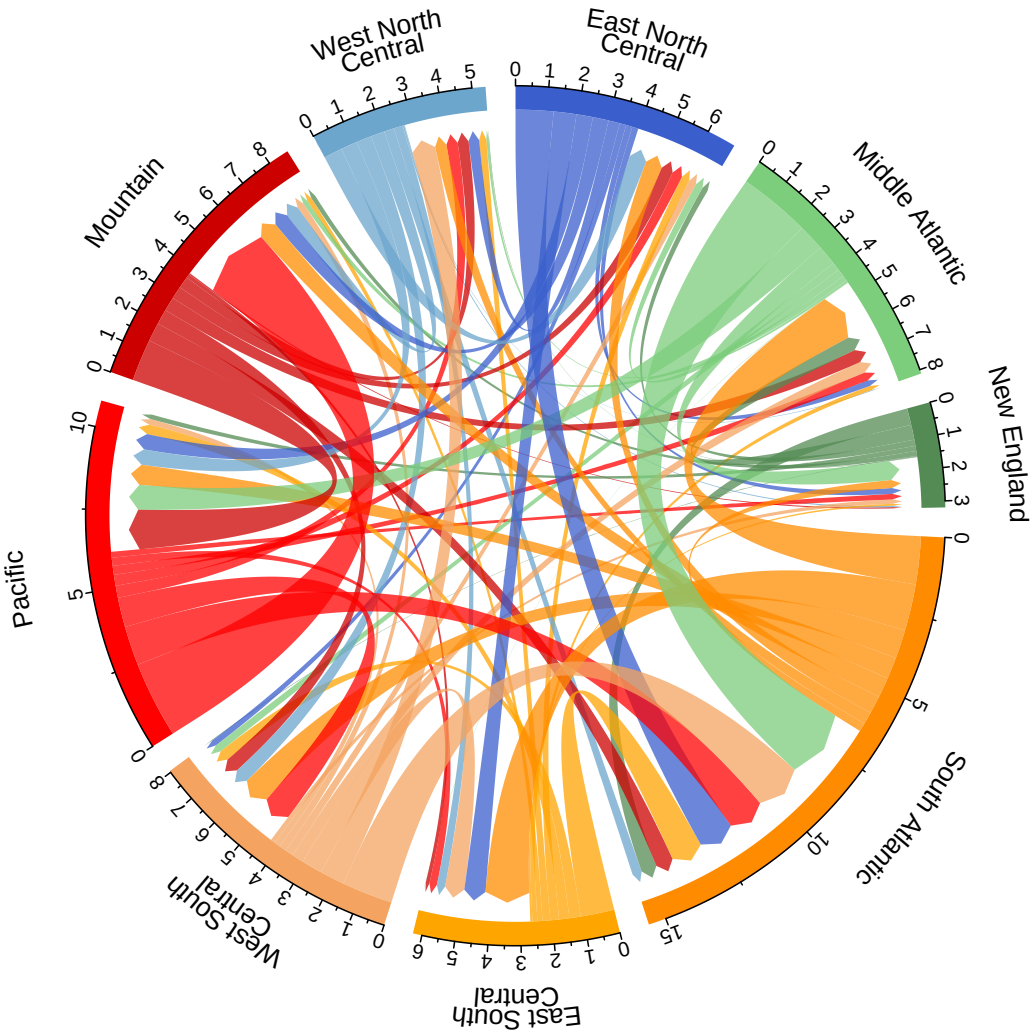
Data illustration

Origin-destination migration flows in 2015 by ACS One-Year



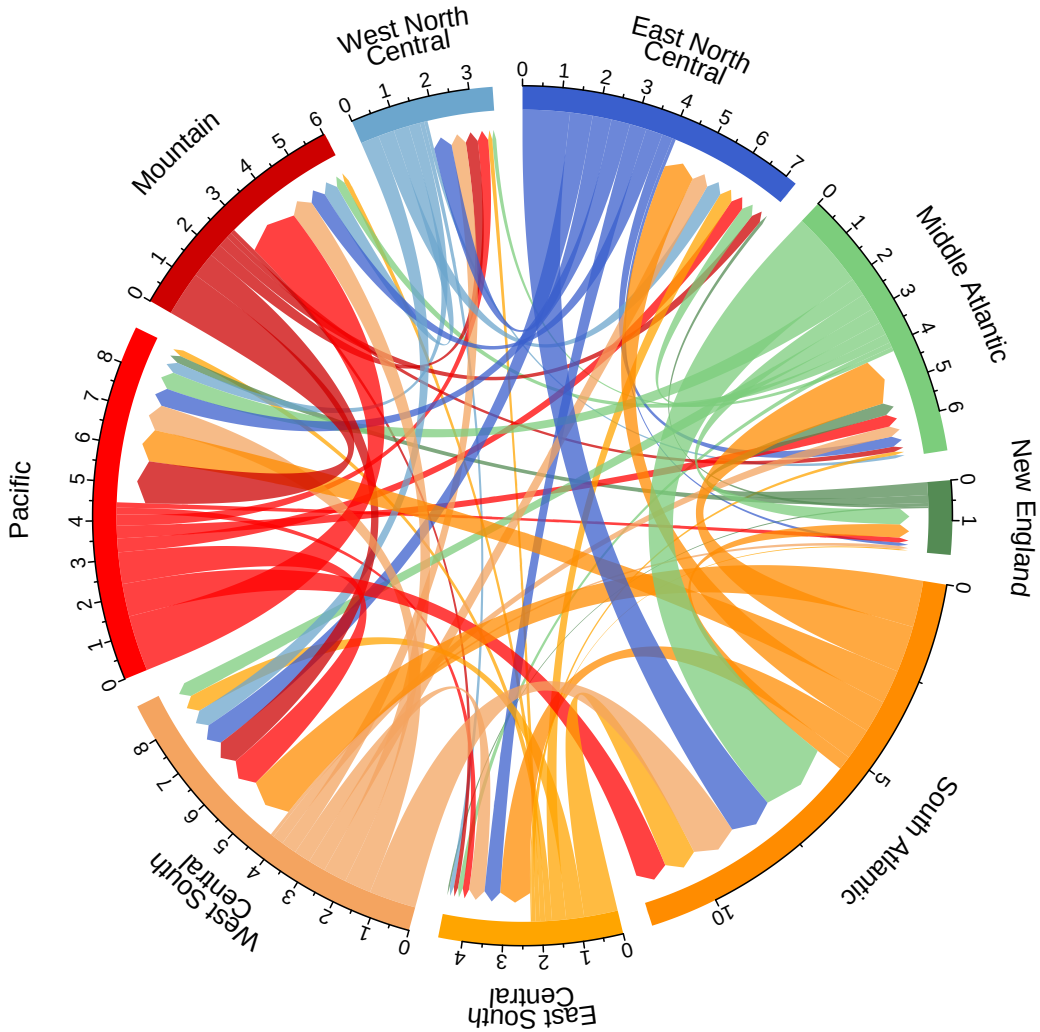
Data illustration

Origin-destination migration flows in 2015 by CPS One-Year



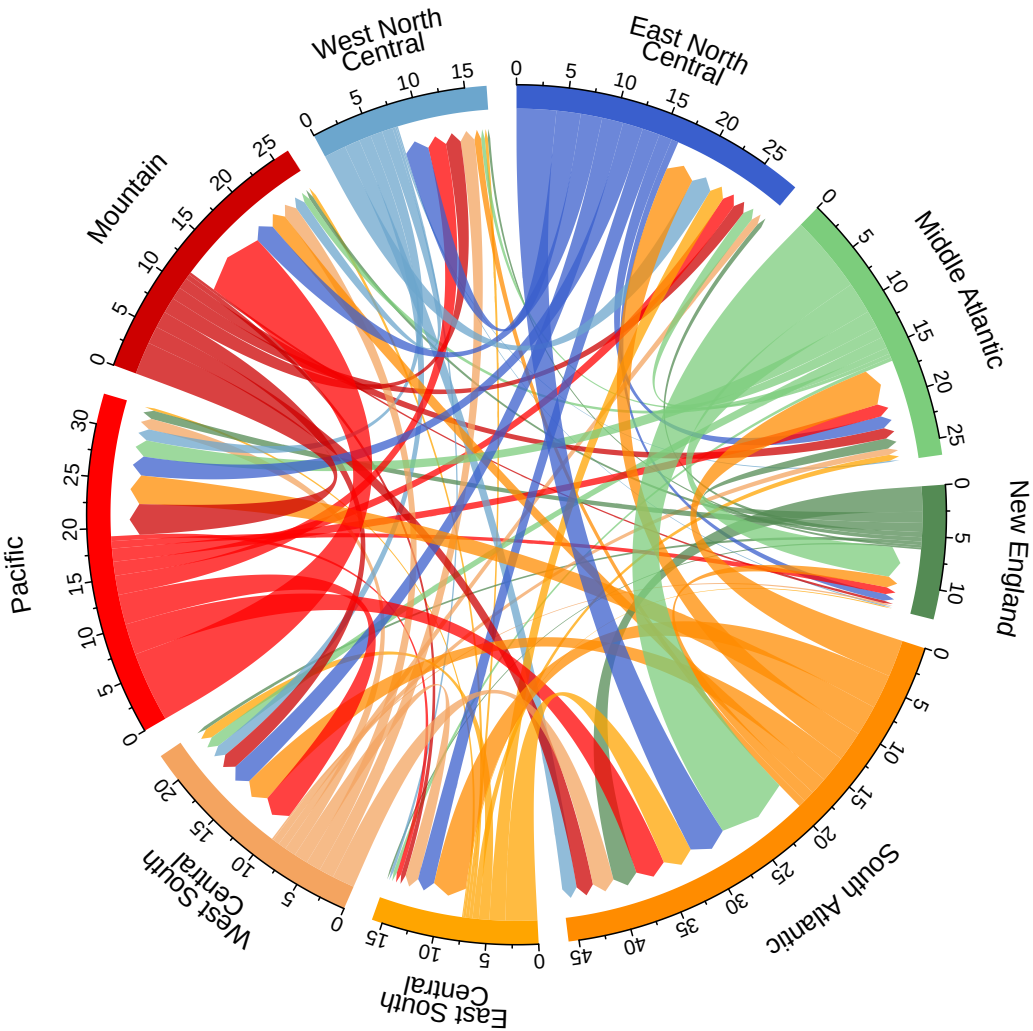
Data illustration

Origin-destination migration flows in 2015 by IRS One-Year

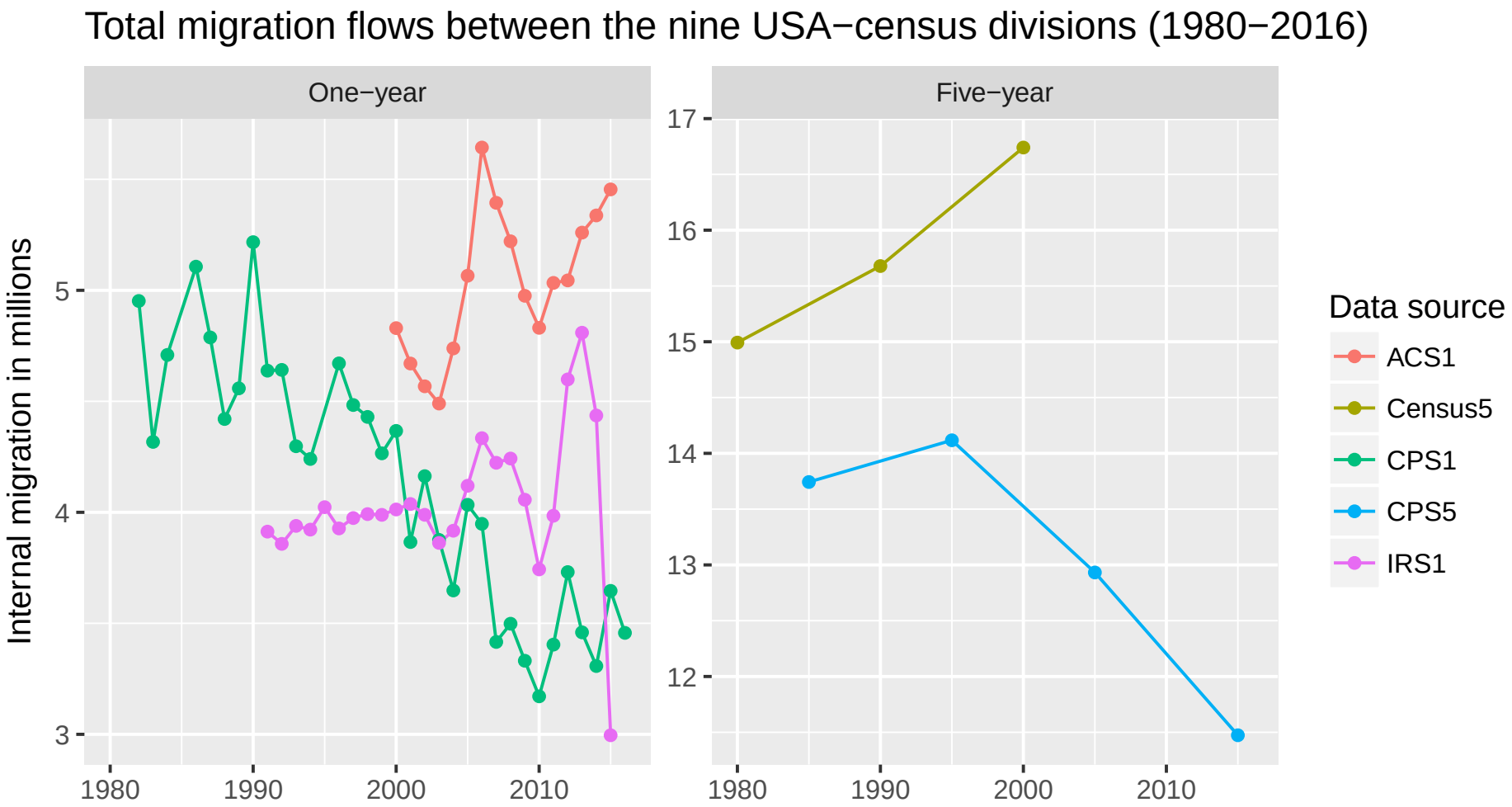


Data illustration

Origin-destination migration flows in 2015 by CPS Five-Year



Total migration flows



How can we deal with this data to do some nice statistics?

Research objective

- We observe flow counts z_{ijt}^k from origin i to destination j $(i = j = 1, \ldots, 9)$ during year t reported by data source k and duration interval status $d = 1, 5$.

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

- We aim to estimate true migration flows y_{ijt} using the observed z_{ijt}^k.

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

- We choose the Bayesian paradigm because of the flexibility to combine data from different sources and the possibility to incorporate prior knowledge.

Methodology (I)

- Measurement Model:

$$\log z_{ijt}^k = \log y_{ijt} + \log \lambda^k + \log \gamma^k + \log \delta^k + \epsilon^k$$

where $\epsilon^k \sim N(0, \tau_\epsilon^k)$

- Data Generating Model:

$$\log y_{ijt} = \mu + \alpha_{ij} + \beta_{ij} \log y_{ij(t-1)} + \eta_{ijt}$$

where $\eta_{ijt} \sim N(0, \tau_\eta)$

- Model Priors:

$$\mu \sim N(0, 10^{-6}) \qquad \alpha_{ij} \sim N(0, \tau_\alpha) \qquad \beta_{ij} \sim N(\mu_\beta, \tau_\beta)$$

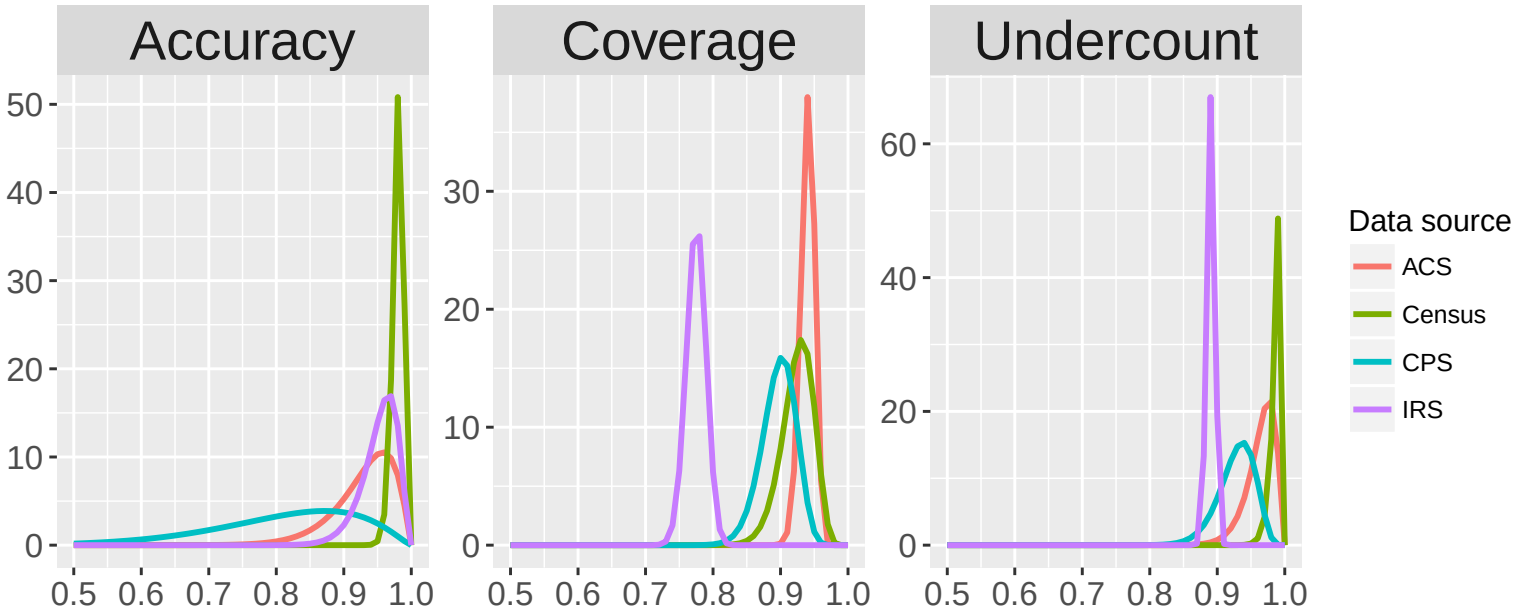
$$\tau_\alpha \sim U(0, 100) \qquad \mu_\beta \sim N(0, 10^{-3}) \qquad \tau_\beta \sim U(0, 100)$$

$$\tau_\eta \sim U(0, 100)$$

- $N(\mu, \tau)$ is a normal distribution with mean μ and precision (inverse variance) τ.

Methodology (II)

- γ^k (coverage): Percentage of people to be interviewed ($\approx$ the population of interest).
- λ^k (undercount): Percentage of people who response to the survey (response rates).
- δ^k (duration): Difference of the duration of movements (one-year, five-year).
- ϵ^k (accuracy): How well the estimate reflects the true value (margin of error).
- For accuracy, coverage and undercount, we derived informative priors from the literature surrounding the data sources.

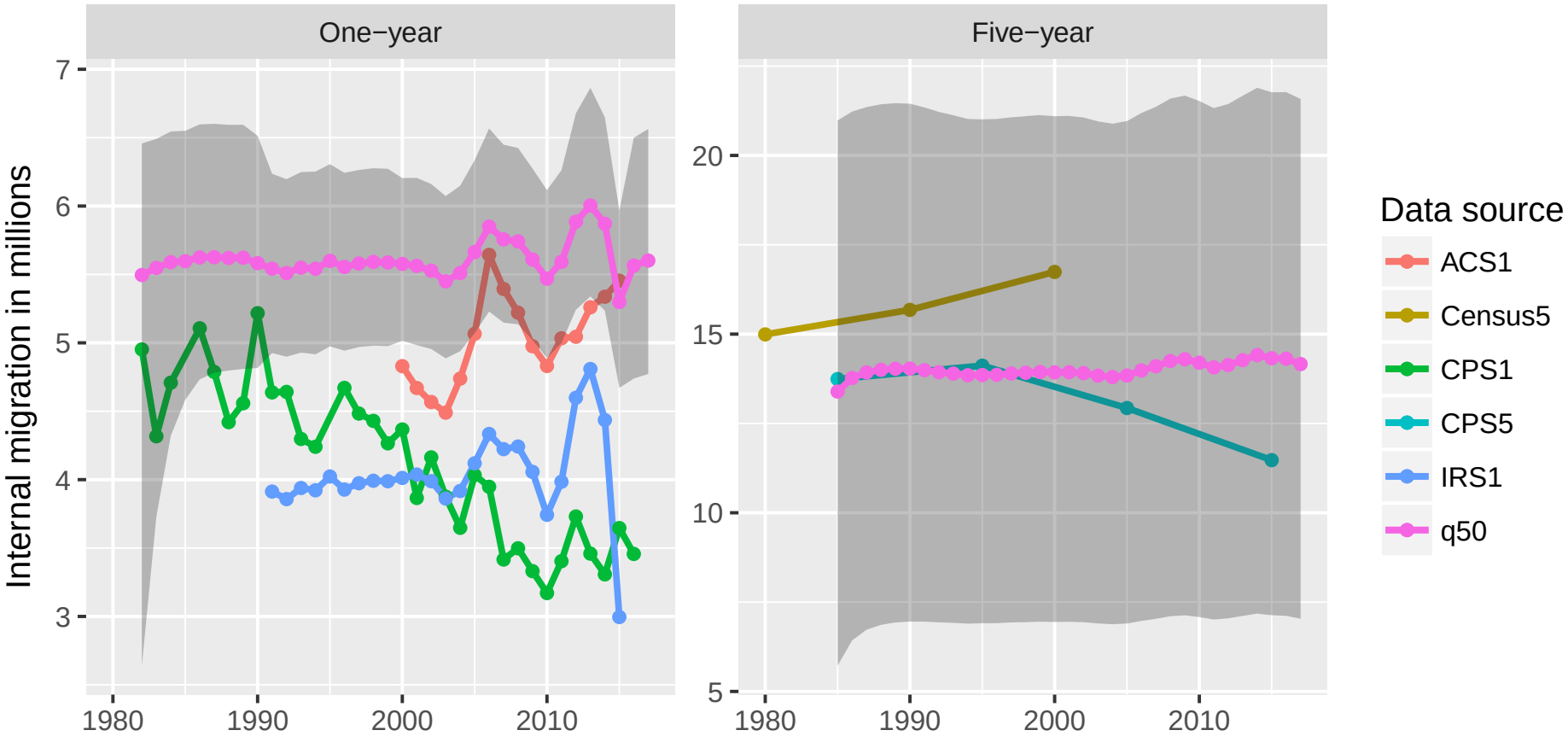


- The prior for duration comes from the equations of expectation and variance of the log-normal distribution.

Results (I)

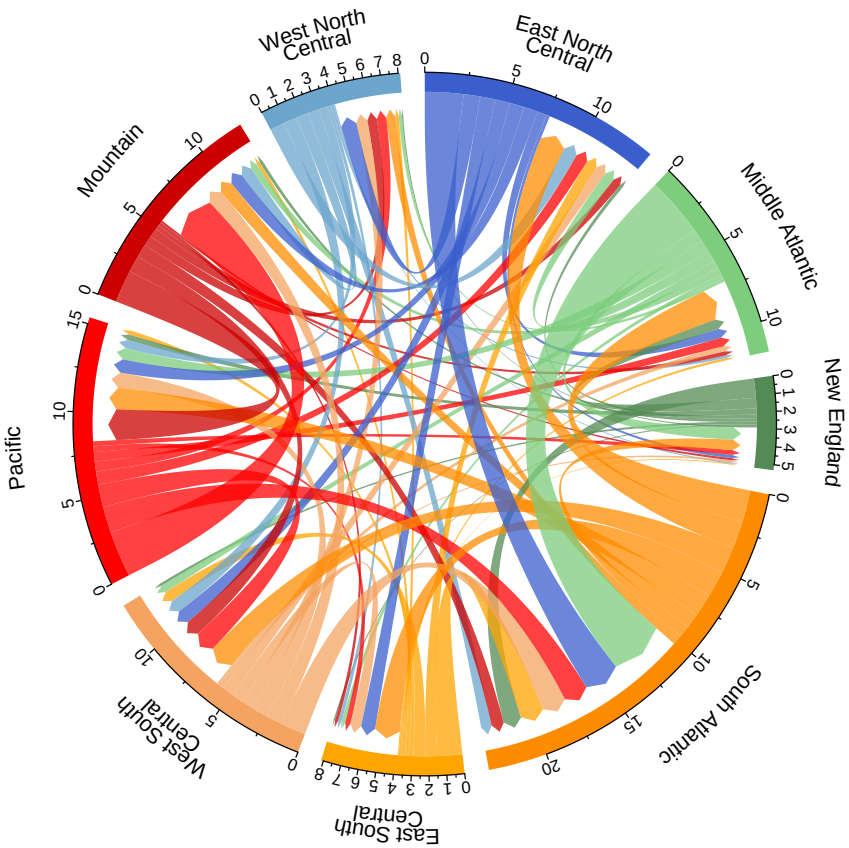
- The model is written in **JAGS** and run with the R package **rjags**.

Total migration flows between the nine USA−census divisions (1980−2017).
Medians and credible intervals.

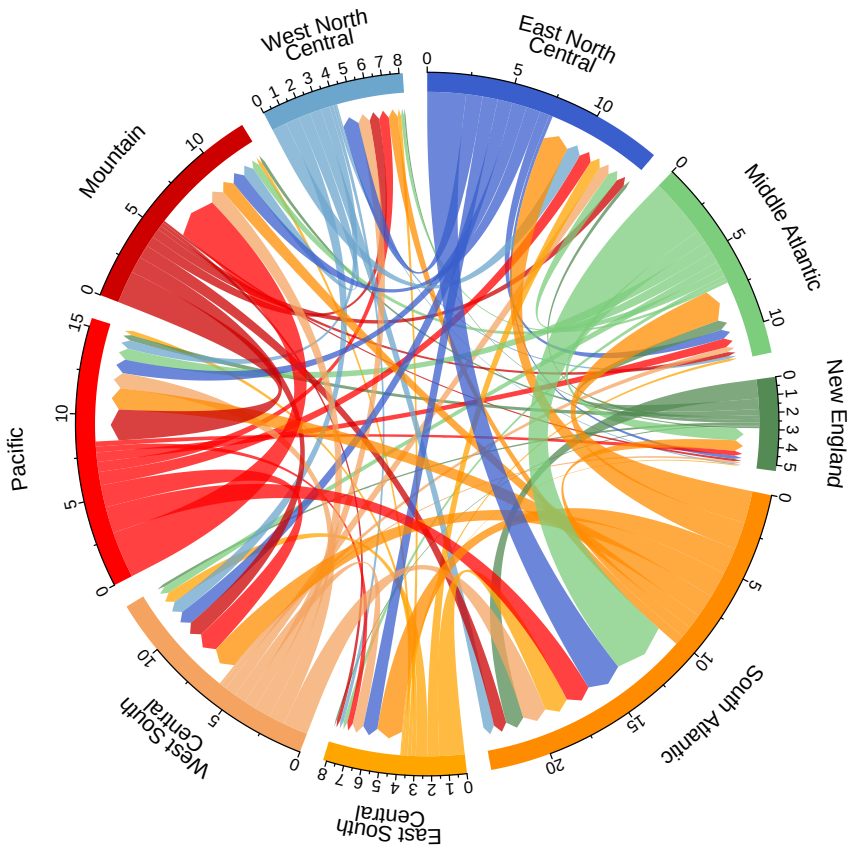


Results (II)

2016 True Flow Estimate, One-Year



2017 True Flow Forecast, One-Year



Conclusions and future work

- Our method estimates synthetic bilateral migration flows that borrow strength over multiple data sources.
- The resulting estimates allow for a better understanding of people's movement patterns beyond the confines of a single source.
- Migration is forecasted in future periods using past data.
- Use data from Brazil. We aim to create a flexible modelling framework that can be applied to estimate internal migration in any country.
- We will expand the model to incorporate geo-located Twitter data.

References

- Nowok, B. and Willekens, F. A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place* 17, 521-533, 2011.

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

- Willekens, F. Evidence-based monitoring of international migration flows in Europe. *Eurostat conference "Towards more agile social statistics"*, 1-42, 2016.

In collaboration with:

- Guy Abel (Asian Demographic Research Institute, Shanghai University, China & Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Austria)
- Dilek Yildiz (Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Austria)
- Arkadiusz Wiśniowski (Cathie Marsh Institute for Social Research, School of Social Sciences, University of Manchester, United Kingdom)

Funding award:

- STE-Projekts 0059 (Jubiläumsfonds der Stadt Wien)
  "Combining Traditional and Emerging Big Data Sources to Model Population Movement Patterns (BIGMIG)"

THANKS FOR THE ATTENTION

www.uv.es/vivigui

Guillermo.Vinue.Visus@oeaw.ac.at

Guillermo.Vinue@uv.es