UCLouvain

Understanding migration within countries: A supervised Machine Learning perspective

Jean-Francois Maystadt, Silvia Peracchi, Ella Sargsyan, Liangzhi You

IRES/LIDAM, Université catholique de Louvain; University of Lancaster; IFPRI

September 18, 2025

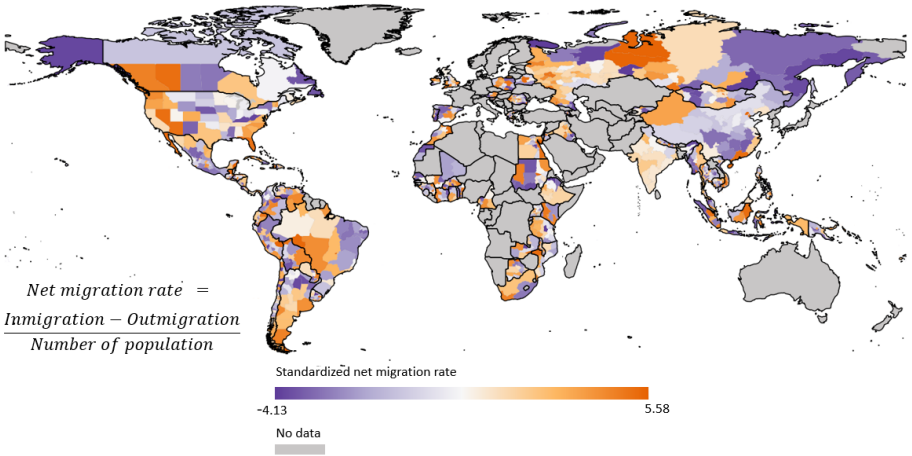
The majority of migrants remain within their own borders. While 281 million people have migrated in another country in 2020, the number of internal migrants is estimated to stand at around 763 million (IOM, 2021; UN DESA, 2016).

What tools and variables better predict internal migration globally?

- Contributions
  - Scarce systematic, harmonized bilateral data on internal migration: we provide data and explore within-country variations at a **global level**, both pooled and by socio-economic group
  - Compare standard **PPML versus ML approaches** in predicting internal migration globally
  - Quantify the relative **importance of different predictors** of migration.
  - Policy relevance by exploiting **heterogeneity by region, by gender (by skill, by urban)**

Net Migration Rates Within Countries

UCLouvain



Source: The figure is compiled by the authors using data from IPUMS International.

Data

Combining 3 main sources of data:

- IPUMSI census data – 1st subnational geographic level of residence 5 years (or 1 year) prior to survey,
  - gender-specific
  - age-specific
  - education-specific
  - urbanity-specific
- PRIO-GRID – grid-level determinants of migration
- UCDP – event-level conflict data

Resulting dataset

- unbalanced bilateral panel data of 1st admin divisions
- 66 countries covering 1992-2014 period worldwide

Full list of variables

- **Conflict**
  - Conflict events, fatalities, excluded groups
- **Land use**
  - Shares of land types: pasture, forest, agriculture, urban, etc.
- **Economic**
  - Crop price indices, nightlights, human capital (education shares, shares of employed/active, shares of occupied in skilled agriculture).
- **Demographic**
  - Population size, density, age structure, foreign inflows
- **Food security**
  - Malnutrition, infant mortality
- **Climate**
  - Temperature and precipitation in levels and anomalies, droughts
- **Geographic**
  - Bilateral distance and contiguity, Land area, border distance, travel times to major cities.

Gravity

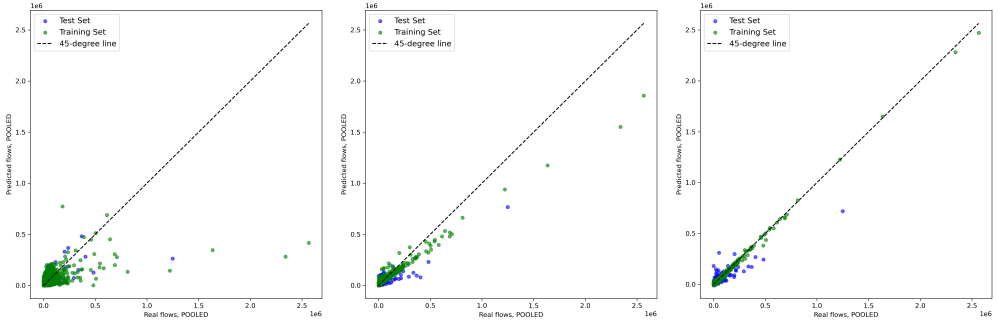
- The **PPML estimator** is the most widely used specification in *gravity model* frameworks.
-

$$Migration_{ij|t,c} = \alpha_t \alpha_c (Econ_{i|j,t,c})^{\beta_1} (Dem_{i|j,t,c})^{\beta_2} (Food_{i|j,t,c})^{\beta_3} (Geo_{i|j,t,c})^{\beta_4} (Env_{i|j,t,c})^{\beta_5} (Pol_{i|j,t,c})^{\beta_6} (Land_{i|j,t,c})^{\beta_7} (I_{ij|t,c})^{\beta_8} \epsilon_{ij|t,c}.$$

- **ML estimators** :
  Random Forest and Gradient Boosting Trees are *ensemble* methods that build multiple decision trees during training.
  Potential advantages:
  - Designed for prediction
  - Better handling of missing values
  - More flexible handling of non-linear relationship between the regressors and the outcome.
  - Research has progress to offer some interpretability tools from ML models.

Model performance

Figure: Predicted vs actual values



(a) PPML (b) Random Forests (c) XGBoost

Notes: Scatter plots representing the performance of the three main models. The x-axis represents the actual migration values. The y-axis represents the predicted ones. A 45-degree dashed line represents perfect prediction. Green dots correspond to in-sample predictions, blue dots represent predictions out of sample.

Results visualisation

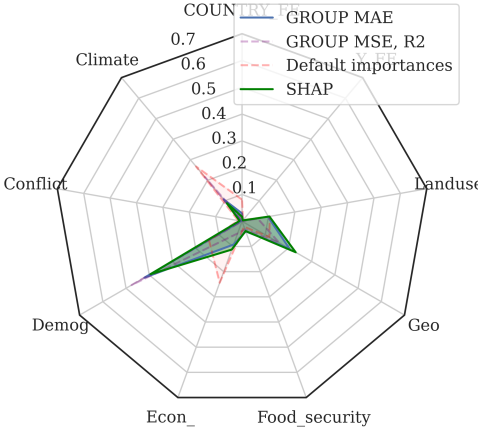
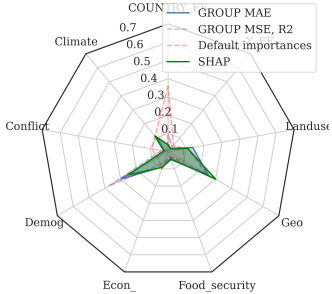
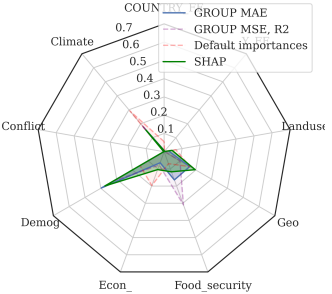


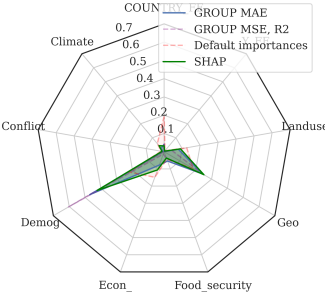
Figure: Permutation Importance and Group Importances



(a) Africa



(b) Asia



(c) LAC

Other heterogeneity

UCLouvain

- Climatic variables have a stronger impact on internal migrants in rural rather than urban destinations.
- Climatic variables also play a greater role in explaining migration flows among younger and middle-aged individuals compared to older age groups. Older individuals appear more sensitive to geographic characteristics.
- Economic variables are particularly important for the unemployed, who also show less sensitivity to geographic and demographic factors than employed and inactive people.
- In the agricultural sector, food security emerges as a significant predictor of migration, while demographic factors appear to be relatively less important.
- In fragile countries, conflict and food security are more salient features.

Conclusions

Policy relevance

- We provide bilateral internal migration data and advise on the best models to predict internal migration.
- Better predicting internal flows matters for **urban planning policy, policy on adaptation to climate change and resilience to shocks.**

What's next?

- Make the data available
- Further research ideas (more academically oriented)
  - Environmental migration and conflict?
  - Migration and conflict risk?
  - Role of soil-matching?

Appendix

Table: Full baseline sample, out of sample performance

Model	*RMSE*		*R2*		*MAE*		*N*	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
PPML-FE only	21813.607	20809.218	0.056	0.088	2942.871	3120.738	53891	6042
PPML	17914.876	16235.775	0.363	0.445	1983.999	2165.772	53891	6042
RF	6377.097	11453.982	0.919	0.724	622.963	1651.119	53891	6042
RF, trimmed	6031.449	11899.08	0.928	0.702	612.831	1661.449	53891	6042
XGBoost	1827.175	10981.404	0.993	0.746	684.641	1565.02	53891	6042
XGBoost; trimmed	1699.7	11598.283	0.994	0.717	652.578	1544.655	53891	6042

Notes: hyperparameters for ML models are calibrated with random grid search and 5-folds cross-validation of the training sample. From the cross validation, we select the best hyperparameters combination based on best MAE scores, constraining R2 values to be above 10 percent to prevent overfitting.

Example of regression tree

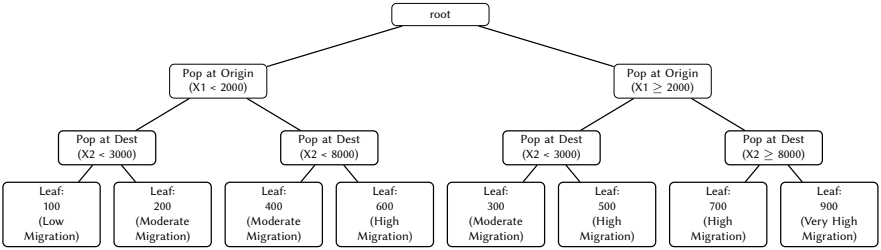
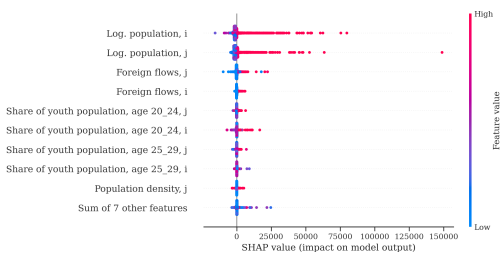
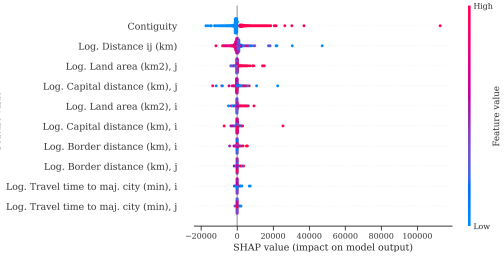


Figure: SHAP contribution of the variables in each group



(a) Demographic



(b) Geographic

SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) take inspiration from game theory and represent the contribution of a feature to the model's prediction. The measure is built to observe what happens to the model prediction, if a variable is excluded from the analysis vs the case that variable is maintained, while considering all possibilities of presence and absence of all other variables. A large negative (positive) value indicates that variable strongly matters for determining predictions much below (above) average.

Final sample, country-year availability, Asia

UCLouvain

	1995	1997	1998	1999	2000	2001	2004	2005	2008	2009	2010	2011	2013	2014	Total
Armenia	0	0	0	0	0	132	0	0	0	0	0	132	0	0	264
Cambodia	0	0	462	0	0	0	462	0	462	0	0	0	462	0	1848
China	0	0	0	0	812	0	0	0	0	0	0	0	0	0	812
India	0	0	0	870	0	0	0	0	0	0	0	0	0	0	870
Indonesia	756	0	0	0	756	0	0	756	0	0	756	0	0	0	3024
Iraq	0	306	0	0	0	0	0	0	0	0	0	0	0	0	306
Jordan	0	0	0	0	0	0	156	0	0	0	0	0	0	0	156
Kyrghzstan	0	0	0	56	0	0	0	0	0	0	0	0	0	0	56
Malaysia	0	0	0	0	156	0	0	0	0	0	0	0	0	0	156
Mongolia	0	0	0	0	420	0	0	0	0	0	0	0	0	0	420
Myanmar	0	0	0	0	0	0	0	0	0	0	0	0	0	210	210
Nepal	0	0	0	0	0	182	0	0	0	0	0	182	0	0	364
Philippines	0	0	0	0	4970	0	0	0	0	0	4970	0	0	0	9940
Thailand	0	0	0	0	4556	0	0	0	0	0	0	0	0	0	4556
Vietnam	0	0	0	1482	0	0	0	0	0	1482	0	0	0	0	2964
Total	756	306	462	2408	11670	314	618	756	462	1482	5726	314	462	210	25946

Final sample, country-year availability, Africa

UCLouvain

	1992	1996	1997	1998	2000	2001	2002	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
Benin	132	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	132	0	396
Botswana	0	0	0	0	0	420	0	0	0	0	0	0	0	0	420	0	0	0	840
Burkina Faso	0	156	0	0	0	0	0	0	0	156	0	0	0	0	0	0	0	0	312
Cameroon	0	0	0	0	0	0	0	0	42	0	0	0	0	0	0	0	0	0	42
Cote D'Ivoire	0	0	0	72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	72
Egypt	0	600	0	0	0	0	0	0	0	600	0	0	0	0	0	0	0	0	1200
Ethiopia	0	0	0	0	0	0	0	0	0	0	110	0	0	0	0	0	0	0	110
Ghana	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0	90
Guinea	0	1056	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1056	2112
Kenya	0	0	0	0	0	0	0	0	0	0	0	0	72	0	0	0	0	0	72
Malawi	0	0	0	0	0	0	0	0	0	0	0	702	0	0	0	0	0	0	702
Mali	0	0	0	56	0	0	0	0	0	0	0	0	56	0	0	0	0	0	112
Mauritius	0	0	0	0	56	0	0	0	0	0	0	0	0	0	56	0	0	0	112
Morocco	0	0	0	0	0	0	0	240	0	0	0	0	0	0	0	0	0	0	240
Mozambique	0	0	132	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	264
Rwanda	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	25
Senegal	0	0	0	0	0	0	56	0	0	0	0	0	0	0	0	0	56	0	112
South Africa	0	0	0	0	0	12	0	0	0	0	12	0	0	0	12	0	0	0	36
Sudan	0	0	0	0	0	0	0	0	0	0	0	600	0	0	0	0	0	0	600
Tanzania	0	0	0	0	0	0	342	0	0	0	0	0	0	0	0	342	0	0	684
Togo	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	6
Uganda	0	0	0	0	0	0	1332	0	0	0	0	0	0	0	0	0	0	1332	2664
Zambia	0	0	0	0	56	0	0	0	0	0	0	0	0	56	0	0	0	0	112
Total	132	1812	132	128	202	432	1862	240	42	756	254	1302	128	62	488	367	188	2388	10915

Final sample, country-year availability, Europe

UCLouvain

	1992	1996	1999	2001	2002	2006	2010	2011	Total
Austria	0	0	0	0	0	0	0	72	72
Belarus	0	0	30	0	0	0	0	0	30
Greece	0	0	0	2256	0	0	0	2256	4512
Ireland	0	30	0	0	30	30	0	30	120
Poland	0	0	0	0	240	0	0	0	240
Portugal	0	0	0	380	0	0	0	380	760
Romania	1482	0	0	0	1482	0	0	0	2964
Russia	0	0	0	0	0	0	6806	0	6806
Slovenia	0	0	0	0	131	0	0	0	131
Spain	0	0	0	272	0	0	0	272	544
United Kingdom	0	0	0	132	0	0	0	0	132
Total	1482	30	30	3040	1883	30	6806	3010	16311

Final sample, country-year availability, Americas

UCLouvain

	1992	1993	1994	1995	1996	2000	2001	2002	2003	2005	2006	2007	2010	2011	2012	Total
Argentina	0	0	0	0	0	0	552	0	0	0	0	0	0	0	0	552
Bolivia	90	0	0	0	0	0	90	0	0	0	0	0	0	0	90	270
Brazil	0	0	0	0	0	600	0	0	0	0	0	0	600	0	0	1200
Canada	0	0	0	0	0	0	110	0	0	0	0	0	0	0	0	110
Chile	1892	0	0	0	0	0	0	1892	0	0	0	0	0	0	0	3784
Colombia	0	420	0	0	0	0	0	0	0	420	0	0	0	0	0	840
Costa Rica	0	0	0	0	0	42	0	0	0	0	0	0	0	42	0	84
Cuba	0	0	0	0	0	0	0	182	0	0	0	0	0	0	182	364
Dominican Republic	0	0	0	0	0	0	0	0	0	0	0	0	552	0	0	552
Ecuador	0	0	0	0	0	0	182	0	0	0	0	0	182	0	0	364
El Salvador	182	0	0	0	0	0	0	0	0	0	0	182	0	0	0	364
Guatemala	0	0	506	0	0	0	0	506	0	0	0	0	0	0	0	1012
Haiti	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	12
Honduras	0	0	0	0	0	0	272	0	0	0	0	0	0	0	0	272
Jamaica	0	0	0	0	0	0	156	0	0	0	0	0	0	0	0	156
Mexico	0	0	0	992	0	992	0	0	0	992	0	0	992	0	0	3968
Nicaragua	0	0	0	156	0	0	0	0	0	156	0	0	0	0	0	312
Paraguay	156	0	0	0	0	0	0	156	0	0	0	0	0	0	0	312
Peru	0	0	0	0	0	0	0	0	0	0	0	650	0	0	0	650
Trinidad and Tobago	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
United States	0	0	0	0	0	2652	0	0	0	2652	0	0	2652	0	0	7956
Uruguay	0	0	0	0	342	0	0	0	0	0	342	0	0	342	0	1026
Venezuela	0	0	0	0	0	0	462	0	0	0	0	0	0	0	0	462
Total	2320	420	506	1148	342	4288	1824	2736	12	4220	342	832	4978	384	272	24624