Bayesian Prediction of Migration Flows

StatApp Project

By:

Louise · Romain · Ishagh · Varnel

December 2025

Report Outline

---

1 Introduction and Literature Review

Predicting international migration is critical for global demographics, but complex. Unlike birth or death rates, migration data depend on volatile geopolitical contexts and lack global reliability: emigration figures from country A to B rarely match immigration figures of B coming from A, for example.

This project first replicates the Gravity Model by Welch and Raftery (2022) as a benchmark, before comparing it to a Poisson Hurdle model. Finally, we discuss the hierarchical Bayesian model (Azose and Raftery, 2015), which currently stands out for the quality of its predictions.

The Data Challenge: Estimation vs. Observation

The main obstacle is the absence of direct measurement of global bilateral flows. The variable m_{ijt} used here (flow from i to j at time t) is not a raw observation but a statistical estimate derived from variations in migrant *stocks* (census data). These flows are deduced by Azose and Raftery (2019) via a pseudo-Bayesian method satisfying the demographic balancing equation:

$$P_{t+1} = P_t + \text{Births}_t - \text{Deaths}_t + \text{Net Migration}_t$$

The data for the gravity model are therefore themselves derived from a prior modeling process filling the gaps in official registers.

1.1 The Gravity Model: The Benchmark

Inspired by Newton's law, this model posits that migratory flow is proportional to the origin and destination populations (P_i, P_j) and inversely proportional to the distance between them. Its log-linear form (Welch and Raftery, 2022) is specified as follows:

$$\log(m_{ijt}) = \beta_0 + \beta_1 \log(P_{it}) + \beta_2 \log(P_{jt}) + \beta_3 \log(D_{ij}) + \mathbf{X}_{ij}\gamma + \epsilon_{ijt}$$

where $\mathbf{X}_{ij}$ includes dummy variables (border, language, colonial ties) and socio-economic variables (infant mortality, urban population, etc.)[1].

The Problem of Zero Flows

The major limitation of this model is the log(0) term, as flow matrices contain many zeros. Welch and Raftery (2022) bypass the problem by removing zero observations, which allows for OLS estimation but leads to significant information loss (see Fig. A1 in the appendix): a zero flow is not useless information.

Alternatives exist, such as the Poisson Hurdle model which accepts zeros. However, Welch and Raftery (2022) conclude that while this model explains past data better, it fails to correctly predict future flows (the prediction error explodes, see Appendix A).

1.2 The Hierarchical Bayesian Approach

To address the predictive limitations of linear models like the gravity model, Azose and Raftery (2015) model **net migration rates** $(r_{c,t})$ via a hierarchical Bayesian auto-regressive process (AR(1)):

$$(r_{c,t} - \mu_c) = \phi_c(r_{c,t-1} - \mu_c) + \epsilon_{c,t}$$

This specification captures migratory *inertia* (strong dependence on the past) while guaranteeing stationarity ($|\phi_c| < 1$) to avoid unrealistic long-term divergence. The **hierarchical structure** allows countries with little data to "borrow information" from global trends. As the high dimensionality of parameters ($\sim$600 for 200 countries) makes the exact calculation of the posterior distribution impossible, estimation relies on **MCMC sampling (Gibbs Sampler)** (Gelman et al., 2013).

Finally, Welch and Raftery (2022) adapt this approach to bilateral flows (m_{ijt}): the net migration rates determine the total outflow, which is then allocated across destinations according to pairwise patterns (κ_{ij}). These parameters κ_{ij} are derived from a centred logarithmic transformation, which reflects the stable, long-term preference of migrants from country i to settle in country j.

[1] Welch & Raftery have excluded volatile economic predictors like GDP per capita. With $N \approx 87,000$, overfitting is definitely not the concern; the constraint is predictive. Forecasting migration based on GDP would require forecasting GDP itself, which is probably the main difficulty, and the main reason why it is absent from the study of the authors (Welch and Raftery, 2022).

2 Gravity Model Estimation and Results

2.1 Estimation Strategy.

To estimate the drivers of migration, we applied the standard log-linear gravity specification derived from Welch and Raftery (2022).

2.2 Challenges in Covariate Consistency and Missing Data.

A significant methodological constraint in estimating the gravity model lies in the structural incompleteness of historical demographic series for key independent variables, specifically the Infant Mortality Rate (IMR) and the Potential Support Ratio (PSR). Unlike the migration flow matrix, which provides comprehensive coverage for all 200 most populated territories, these development indicators exhibit systematic attrition due to geopolitical and administrative factors.

While Welch and Raftery (2022) define the theoretical specification of the gravity model, the specific protocols used to handle historical covariate discontinuities—such as the absence of World Bank data for nations like South Sudan prior to independence—are not explicitly detailed in the published methodology.

2.3 Regression Results.

The model was fitted using Ordinary Least Squares (OLS) with Heteroscedasticity-Consistent (HC1) standard errors. **Crucially, this estimation was performed exclusively on complete observations; no data filling or imputation was applied to the training set for this regression.** The analysis yields an R^2 of **0.42**, indicating that the baseline gravity variables explain approximately 42% of the variation in global migration flows.

Table 1: Résultats MCO (Var. dép : $\log m_{ijt}$)

Variable	Coef.	t-stat	Variable	Coef.	t-stat
Constante	-1.14***	-6.0	$\log(\text{IMR}_d)$	-0.64***	-45.6
log(Dist)	-1.20***	-124.9	$\log(\text{PSR}_o)$	0.27***	14.9
Contiguïté	1.99***	38.9	$\log(\text{PSR}_d)$	-0.10***	-5.3
Langue Com.	1.33***	59.3	$\log(\text{Urb}_o)$	0.04*	1.9
Lien Colonial	2.35***	41.0	$\log(\text{Urb}_d)$	0.17***	7.7
$\log(\text{Pop}_o)$	0.49***	73.7	$\log(\text{Surf}_o)$	0.06***	11.6
$\log(\text{Pop}_d)$	0.41***	63.4	$\log(\text{Surf}_d)$	0.13***	24.8
$\log(\text{IMR}_o)$	-0.43***	-31.9	Temps	-0.03***	-28.2
			Temps^2	-0.00**	-2.1
$N = 87\,834$	$R^2 = 0.426$	$F = 3\,903$			

* p<0.1, ** p<0.05, *** p<0.01 (HC1)

The estimated coefficients align with theoretical expectations: distance acts as a significant deterrent to migration ($\beta_{dist} < 0$), while origin and destination populations are strong positive drivers ($\beta_{pop} > 0$). Notably, the development indicators (IMR, PSR) remain significant even in this restricted sample, confirming their utility as predictors for future flows.

3 Next Step

Our immediate future work following this report focuses on:

- **The search for more robust data.** Key variables (e.g., infant mortality) are missing for several countries in UN/CEPII databases. Currently, it leads to significant information loss (an issue Welch & Raftery also encoutered).
- **The New Variables Dilemma.** The 'Gravity' dataset from CEPII have more than 90 explanatory variables and we plan to test the power of some of them. For instance, GDP per capita improves in-sample fit, but we suspect Welch & Raftery excluded it to avoid "forecasting the predictors" (relying on uncertain future GDP to predict migration). We will test if the predictive gain justifies such added uncertainty. We also plan to integrate a binary variable indicating whether country i is in a tropical zone (the most vulnerable areas if the Paris Agreement limits are exceeded). However, as these climate concerns are recent, they likely will not explain flows prior to 2015 or 2020.
- **Smart Variable Selection (LASSO).** Given our large dataset, overfitting is not a primary concern. However, we will use **LASSO regression** to efficiently identify the most relevant predictors among many candidates, ensuring we capture true signal rather than noise.
- **Next Steps.** We will finalize the validation of our Gravity and PPML benchmarks against Welch & Raftery's results, before moving to the Hierarchical Bayesian model to better capture temporal dynamics.

A Supplementary Material

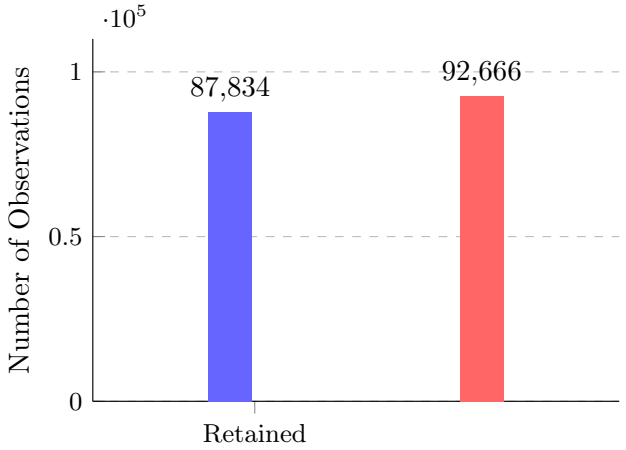

Figure A1: Data Selection: Retained vs. Removed Lines (Total theoretical pairs: $190 \times 190 \times 5$)

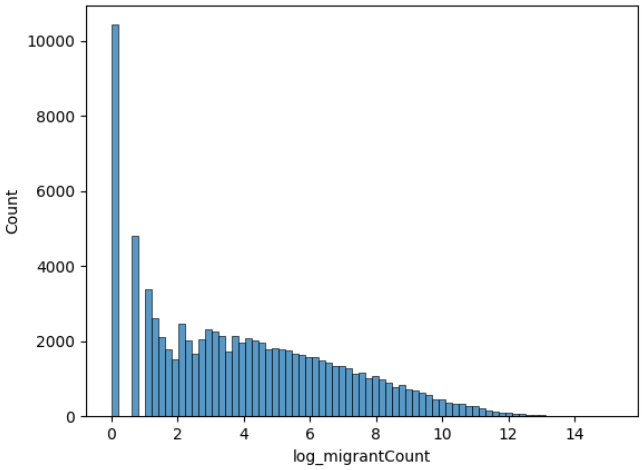

Figure A2: Distribution of Log-Flows ($m_{ij} > 0$)

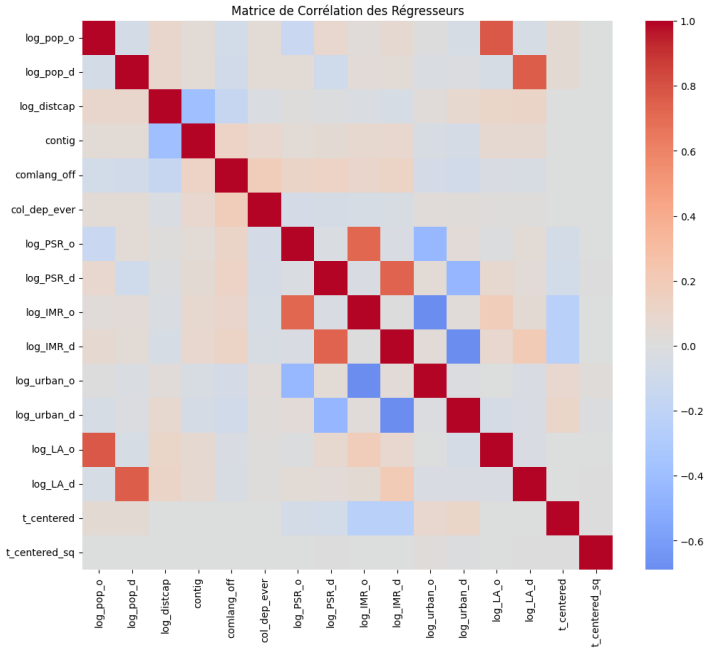

Figure A3: Covariate Correlation Matrix

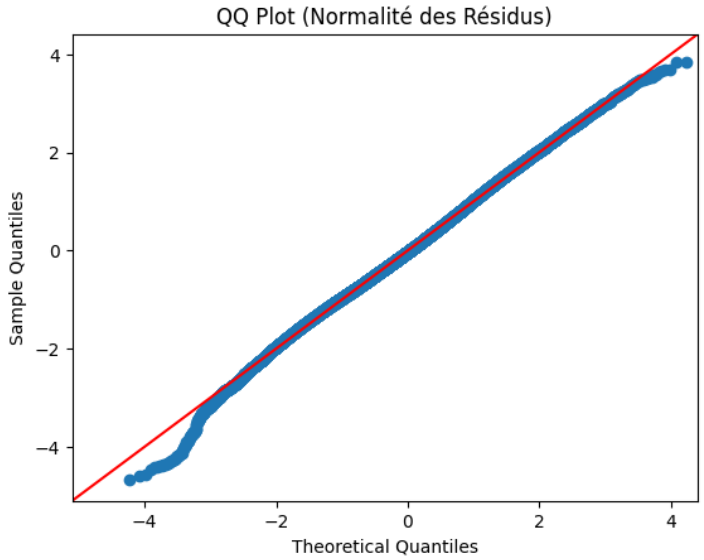

Figure A4: Q-Q Plot of Residuals

Figure A5: Gravity Model Diagnostics: Multicollinearity and Normality Checks.

Table 1. Out-of-sample MAE in thousands of migrants per period, MAPE, and 95% prediction interval (PI) coverage for models fitted to all 1990 to 1995 through 2010 to 2015 migration flows and tested on all 39,800 2015 to 2020 flows

			95% PI			
Method	MAE	MAPE	Flow	In	Out	Net
Historic mean flow	1.2	139	—	—	—	—
Persistence	**1.0**	79	—	—	—	—
Gravity model	3.0	1,565	86	77	80	99
Poisson hurdle model	10.0	25,649	90	66	65	48
Bayesian flow model	1.2	**76**	**93**	**87**	**92**	**94**

A bold-face entry indicates the most accurate number for that metric.

References

Jonathan J Azose and Adrian E Raftery. Bayesian probabilistic projection of international migration. *Demography*, 52(5): 1627–1650, 2015. Version publiée du Working Paper de 2013.

Jonathan J Azose and Adrian E Raftery. Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proceedings of the National Academy of Sciences*, 116(1):116–122, 2019.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, third edition, 2013. Updated electronic version (2025).

Nathan G Welch and Adrian E Raftery. Probabilistic forecasts of international bilateral migration flows. *Proceedings of the National Academy of Sciences*, 119(35):e2203822119, 2022.