Bayesian estimation and projections of internal migration flows in Italy

Emanuela Furfaro[1], Federico Castelletti[2], and Adrian E. Raftery[3]

[1]University of California, Davis
[2]Università Cattolica del Sacro Cuore, Milan
[3]University of Washington, Seattle

1 Introduction

The study of human migration has always attracted interest among researchers in many disciplines as well as among policy makers. In fact, migrations impact the size and the age structure of populations producing consequences on labour force resources, impacting labour markets, local economies and social policies (Greenwood, 1997; White & Lindstrom, 2005; Bijak, 2010a). While international migrations have become a major contributor to the national population change in many countries, migration flows within administrative units of the same country contribute to population changes at local scales, triggering subnational dynamics and being of interest to local governments (Congdon, 2010; Rees et al., 2017). Understanding and forecasting both international and internal migration flows is hence essential in population projections as well as in policy making. In this paper, we focus on migration flows within administrative units of the same country referring to them as *internal migrations*.

Some authors have argued that internal migration flows may be important for understanding other subnational demographic behaviors, suggesting, for instance, that they may foster subnational convergence in fertility rates (Daudin et al., 2019), or stressing the linkage between mobility, which mainly involves people in their reproductive age, and life cycle (White & Lindstrom, 2005; Gabrielli et al., 2007; Banougnin et al., 2018). Internal migrations can thus be of interest for understanding other subnational dynamics which can be important for subnational population projections.

Moreover, there is an ongoing debate on whether the globalization process has been inducing similarities between internal and international migrations (Bijak, 2010b), making the study of the former important for the latter. The attention on movements within a country has also recently been brought up by the IMAGE (Internal Migration Around the GlobE) project, which has established an inventory of internal migration data collections across the 193 UN member states, together with data repository and tools to compare internal migrations across countries (Bell et al., 2015; Bernard et al., 2014b,a). Comparative research studies showed that in many advanced economies, such as the United States or Australia, a decline in internal migration flows in the last thirty years has been observed (Bell et al., 2017). In this context, Italy represents an interesting exception. Moreover, while poor data quality often undermines studies on migration, the Italian population registers provide a stable and consistent source of data. For these reasons, in this paper we focus on internal migrations in Italy.

Theories that attempt to explain human internal migrations do not substantially differ from those of international migrations (Willekens, 1994). These include sociological theories, including those that date back to the concept of intervening opportunities (Stouffer, 1940, 1960) and more recent ones based on the concept of *push and pull* factors, migration networks (Massey et al., 1993) and social capital (Faist, 2000), as well as economic theories, which attribute the migration process to expected wages and the labour market (among others, see Bijak (2010a) for a review of the main human migration theories). The so-called gravity theory of migration, which is based on human geography and demography theories, highlights the role of distance and that of the size of the populations involved, arguing that closer regions with larger populations are characterised by larger flows (Zipf, 1946). While these theories may fail in explaining international migrations, they were argued to be particularly suitable for explaining internal migrations, since they do not take account of state borders, visa restrictions, rivalries between regions, etc., which, instead, play an intrinsic role in international migrations (Oberg & Wils, 1992). In the context of international migration, Billari & Dalla-Zuanna (2012) also argued that replacement migration has been partially taking place in low fertility countries, drawing attention on Total Fertility Rate (TFR) for explaining migration flows.

Being able to project the size and direction of flows has been the object of another large body of literature (see Bijak (2010a) for a review). Econometric models have been used not only for detecting pull and push factors of origin and destination areas, but also to forecast flows. Gravity models, strategies that rely on population distributions by age, or that combine population and economic aspects of social development, have also been proposed to forecast migration flows. However, these models tend to produce large errors in forecasting since prediction of explanatory variables is required first, and the predictors may not exhaustively explain the variability in the size of the flows. For these reasons, time series approaches, which make use of past data for making predictions on future migration levels, are usually preferred. Historical data implicitly include the information on push and pull factors, population size, etc, and they also include information that may not be considered by gravity models. Forecasting models of human migration flows have recently been developed in the Bayesian framework (Bijak & Wiśniowski, 2010; Congdon, 2010; Abel et al., 2013; Azose & Raftery, 2015) since the methodology is particularly well suited not only for estimation but also for projections (Alkema et al., 2011). In particular, Bayesian methods provide a coherent quantification of the uncertainty around estimated quantities, a task which is of primary interest in forecasting migration flows.

In order to explain the direction and intensity of internal migration flows in Italy, in this paper we propose a Bayesian hierarchical model rooted in the human geography and demography migration theories. In addition, within the framework of Bayesian population projections, we extend the usage of Bayesian hierarchical models to forecast internal migration flows. The rest of the paper is organized as follows. In Section 2 we introduce our data source and describe the dataset; in Section 3 we outline the proposed method; in Section 4 we present our results, while Section 5 closes the paper with a brief discussion.

2 Data

2.1 Data source

Sources of data on internal migration flows may be substantially different across countries. They generally include population and administrative registers, censuses and national surveys, while sometimes hybrid strategies are considered (Bell et al., 2015; Coleman, 2013). In many European countries, most demographic data come from population registers. In Italy, the "Anagrafe" is a municipality-based registration system that keeps track of dates of birth, death, marriage, divorce and changes in residential addresses. The Anagrafe is hence a valuable data source for a number of population quantities, including birth rate, death rate, fertility rate, population age structure, and migrations. In particular, Italian population register define a migration as a change in residential address. In the context of internal migrations, the migration consists in a cancellation from a municipality (origin) and an inscription in a new different municipality (destination). For each internal movement, the origin address along with the destination address is available, allowing to identify the origin and destination of each migration flow. Individuals' structural information, such as gender, age and citizenship, is also registered along with the movement. Data collected at municipality level are then aggregated at NUTS 3 (Nomenclature of Territorial Units for Statistics), NUTS 2 and NUTS 1 level and made available by the Italian Institute of Statistics (ISTAT).

Internal migration flows between provinces (NUTS 3) are freely available from 2002 to 2018 (from 1955 upon request).

The Anagrafe is the official source of data on migration in Italy and population registers are generally preferred to census data because they allow the monitoring of internal migration flows on a continuous basis (Bell et al., 2015). However, the Anagrafe reflects changes in administrative procedures and may contain anomalies that are worth mentioning. In 2012, for instance, the introduction of a law to speed up the registration process produced an anomalous increase in internal migrations[1]. Other anomalies, although less relevant for the purpose of this paper, concern international migrations. For instance, the inclusion of some eastern European countries in the European Union produced an increase in the flows of foreign citizens that was probably due to the registration of people already present on the territory rather than actual new arrivals (ISTAT, 2017).

2.2 Data description

In this paper, we consider migration flows between all pairs of Italian provinces in years 2002 to 2018 [2]. Italy is divided into about 100 provinces [3] (NUTS 3), twenty regions (NUTS 2) and five macro-regions (NUTS 1). Provinces' size varies from 84,379 people to over 3 million people, and population density varies from 50 people *per* square kilometer up to over 2,000 inhabitants *per* square kilometer. It is worth noticing that the number of provinces varies across time, starting with 103 until 2005, 107 until 2010, then 110 and again 107 since 2017.

We excluded within province flows, so that for each year the dataset consists of a number of observations ranging between $10,506$ (103×102) and $11,990$ (110×109) observations according to the number of existing provinces.

2.2.1 Migration flows

We model the number of people moving from province i to province j at time t. In the following, we then consider the migration count as the response variable of interest. The distribution of the lowest 98.5% flows in 2018 is summarized in Figure 1, which reveals a high concentration of flows towards the zero value.

[1] 4 Decree-Law February 9, 2012 N. 5, converted into law April 4, 2012 N. 35 regarding urgent provisions on simplification and development—procedures for the application of Art. 5 Cambio di residenza in tempo reale (Change of residence in real time)

[2] data source: http://dati.istat.it/

[3] Since 2015, provinces have been classified with the broader term of "institutional bodies of second level", which include administrative units classified as "provinces", along with others classified as autonomous provinces, or metropolitan cities. However, this is merely an administrative/formal distinction, and we will refer to all of them as provinces for simplicity and for consistency among years.

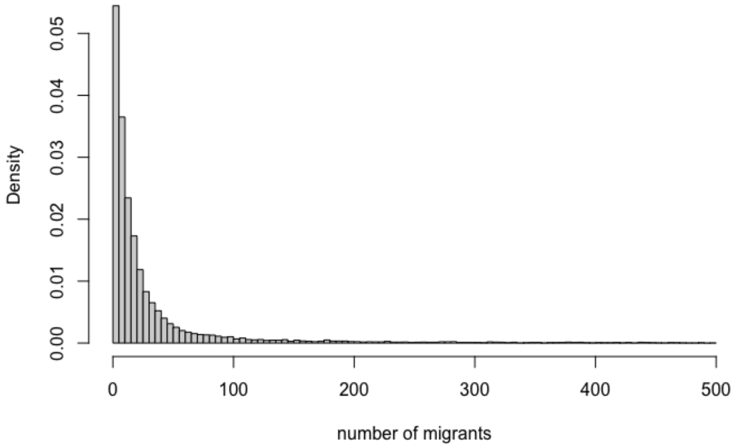

Figure 1: Distribution of the number of migrants between provinces (year 2018).

Internal mobility nowadays mainly involves Italian citizens, but the number of foreign citizens moving within the Italian territory has been growing. In particular, in the last 10 years the percentage of foreign citizens moving within Italy has been estimated between 15% and 19% of the total flows. This is due to both an increase in the number of foreigners in Italy and to a higher propensity of the foreign population of moving out (ISTAT, 2017).

In order to illustrate the intensity and direction of internal migration flows, along with a representation of the administrative divisions considered, Figure 2 represents the net migration rates in each considered province as of 2018. The net migration rate is defined as the difference between the number of people immigrating and the number of people emigrating over the population at mid-year. The map shows a clear tendency of southern regions to have negative net migration rates, and of northern regions to have positive net migration rates. A similar trend was observed for the previous years.

2.2.2 Covariates

We include in the proposed model a number of covariates which have been established to relate with migration flows. Specifically, we consider both geographic variables, which are not time-dependent, and demographic variables. The former include: *(a)* a dummy variable indicating the absence/presence of a border in common between the two provinces, *(b)* the distance between provinces' centroids (in kilometers), *(c)* a dummy variable indicating whether the two provinces belong to the same region (NUTS 2), and *(d)* a dummy variable indicating whether the considered province is the main one of its region.

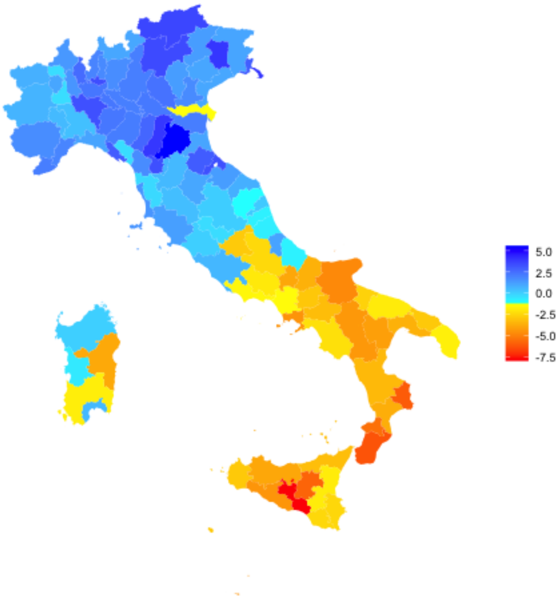

Figure 2: Internal net migration rates per 1,000 people (year 2018).

Regarding demographic variables, we consider *(e)* the population size, *(f)* the percentage of people age 20-30 year, *(g)* the average Total Fertility Rate (TFR) with a time-lag of 20 to 30 years and *(h)* the percentage of foreigners. Notice that variables *(f)*-*(h)* are highly mutually correlated, and also highly correlated with the youth unemployment rate (see also Table 1). Although we do not consider the latter in our model, it is included in Table 1 in order to show its correlation with the demographic variables of interest, in three selected years, 2004, 2012 and 2018.

	Year: 2004				Year: 2012				2018			
	TFR	% FP.	% YP	YU	TFR	% FP.	% YP	YU	TFR	% FP.	% YP	YU
TFR	1.00	-0.78	0.87	0.84	1.00	-0.79	0.65	0.76	1.00	-0.72	0.61	0.77
% FP	-0.78	1.00	-0.59	-0.77	-0.79	1.00	-0.51	-.76	-0.72	1.00	-0.48	-0.67
% YP	0.87	-0.59	1.00	0.7	0.65	-0.51	1.00	0.41	0.61	-0.48	1.00	0.41
YU	0.84	-0.77	0.7	1.00	0.76	-0.76	0.41	1.00	0.77	-0.67	0.41	1.00

Table 1: Correlation coefficients between lagged Total Fertility Rate (TFR), percentage of foreign population (% FP), percentage of young population (%YP), and youth unemployment (YU).

The strength and sign of the relationship between the four mentioned variables is fairly stable throughout the years, although it seems to weaken in more recent years. Beside the positive correlation between TFR and the percentage of young people, which is fairly natural, the negative correlation between TFR and the percentage of foreign citizens suggests that some international replacement migration has been taking place in Italy, especially in lowest fertility areas. The stable positive correlation between youth unemployment rate and TFR also reinforces the relationship between demographic variables and economic ones.

Regarding the possibility of using the above variables as auxiliary in migration flows projections, we finally emphasize that differently from the percentage of foreign population, the percentage of young people, and the unemployment rates, lagged TFR does not require projections as long as we do not forecast more than 20-30 years ahead.

3 Method

3.1 Testing migration theories

In this section we introduce the Bayesian model that we adopt for the analysis of migration flows.

Let $Y_{i,j}^{(t)} \in \{0, 1, \dots\}$, $i \neq j$, $t = 1, \dots, T$ be the random variable describing the migration count from origin i to destination j at time t. Let also $X_1, \dots, X_p$ be a collection of p covariates; see Section 2.2.2 for details. We assume that, conditionally on a flow-specific parameter $\lambda_{i,j}$, $Y_{i,j}^{(t)}$ are independent with Poisson distribution, namely

$$Y_{i,j}^{(t)} \mid \lambda_{i,j} \overset{iid}{\sim} \text{Pois}\left(\lambda_{i,j}\right), \quad \lambda_{i,j} \in \Re^{+}. \tag{1}$$

$$\begin{aligned} \log\left(\lambda_{i,j}\right) \mid \mu_{i,j} &\overset{iid}{\sim} \mathcal{N}\left(\mu_{i,j}, \sigma^2\right), && (2) \\ \mu_{i,j} &= \boldsymbol{\beta}^{\top} \boldsymbol{x}_{i,j}^{(t)}, && (3) \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^{\top}$ is a $(p+1, 1)$ vector of parameters associated with the p covariates (including an intercept term β_0) and $\boldsymbol{x}_{i,j} = (1, x_{i,j,1}^{(t)}, \dots, x_{i,j,p}^{(t)})^{\top}$ is the $(p+1, 1)$ vector of observed covariates relative to flow $Y_{(i,j)}^{(t)}$. We adopt weakly informative priors for both $\boldsymbol{\beta}$ and σ^2, by assigning $\beta_s \overset{iid}{\sim} \mathcal{N}(0, 100)$, $s = 0, \dots, p$, $1/\sigma^2 \sim \text{Gamma}(0.01, 0.01)$. We refer to model defined by (1):(3) as the *Extended Gravity (EG) Model*, which is implemented for each year from 2002 to 2018.

We evaluate the model's calibration using the Probability Integral Transform (PIT) histogram (Czado et al., 2009). This allows to compare the predictive cumulative distribution with the observed data. Since we work with a discrete distribution, we compute the *adjusted* PIT, defined as

$$PIT_{ijt} = P(y_{ijt}^{new} < y_{ijt} | \boldsymbol{y}_{-ijt}) + 0.5 \cdot P(y_{ijt}^{new} = y_{ijt} | \boldsymbol{y}_{-ijt}),$$

where $\boldsymbol{y}_{-ijt}$ is the observation vector with the observation corresponding to flow (i, j) at time t omitted. Hence, PIT represents the value that the predictive cumulative distribution function attains at the observation y_{ijt}. An approximate uniform distribution of the PIT suggests that the model assumption is appropriate for the data.

3.2 Forecasting

The construction of a gravity-like model, especially if the included covariates do not need projections, may help in forecasting migration flows. We will hence apply a small modification to the above model to allow forecasting. If regression coefficients are "stable" across time, meaning that the effect of covariates on the response is constant w.r.t. time, we propose to use the EG model, after including an additional covariate indexing the time, to make forecasts of migration flows.

As a further point, we notice that forecasting models based on autoregressive structures have been shown to produce more adequate predictions in many contexts (Bijak, 2010a). In particular, past values of the response variable may provide information, not fully captured by the available predictors. Accordingly, we propose the use of first-order autoregressive, AR(1) models. In particular, we propose two different models.

Following Zhang & Bryant (2020), our first proposal assumes that, conditionally on a flow-specific parameter $\lambda_{i,j}$, migration counts from province i to province j, $Y_{i,j}^{(t)}$ $(i \neq j, t = 1, \dots, T)$ are independent with Poisson distribution,

$$Y_{i,j}^{(t)} \mid \lambda_{i,j}^{(t)} \overset{iid}{\sim} \text{Pois}\left(\lambda_{i,j}^{(t)} \cdot P_i^{(t)}\right). \tag{4}$$

Parameter $\lambda_{ij}^{(t)}$ represents the migration rate relative to flow (i, j) at time t; we further assume

$$\begin{aligned} \log\left(\lambda_{i,j}^{(t)}\right) \mid \mu_{i,j}^{(t)} &\overset{iid}{\sim} \mathcal{N}\left(\mu_{i,j}^{(t)}, \sigma_{ij}^2\right), & (5)\\ \mu_{i,j}^{(t)} &= \log\left(y_{i,j}^{(t-1)} + 1\right), & (6)\end{aligned}$$

$$\begin{aligned} 1/\sigma_{ij}^2 &\sim \text{Gamma}(\alpha_1, \alpha_2), & (7)\\ \alpha_1 &\sim \text{Unif}(0, 20),\\ \alpha_2 &\sim \text{Unif}(0, 20).\end{aligned}$$

The resulting model is called *BHPM 1*.

Drawing on the literature on Bayesian forecasting of international migration flows, we propose a second model which is an adaptation of the model in Azose & Raftery (2015) to the Poisson case (here simply referred to as *BHPM 2*). In particular we assume

$$Y_{ij}^{(t)} | \lambda_{ij}^{(t)} \sim \text{Pois}(\lambda_{ij}^{(t)} \cdot P_i^{(t)}), \tag{8}$$

$$\log(\lambda_{ij}^{(t)}) \sim \mathcal{N}(\mu_{ij}^{(t)}, \sigma^2), \tag{9}$$

$$\mu_{ij}^{(t)} - \alpha_{ij} = \phi_{ij}(\log(y_{ij}^{(t-1)} + 1) - \alpha_{ij}), \tag{10}$$

with priors on σ^2, α_{ij} and ϕ_{ij}:

$$\begin{aligned} 1/\sigma^2 &\sim \text{Gamma}(1,1), \\ \alpha_{ij} &\overset{iid}{\sim} \mathcal{N}(\lambda_\alpha, \tau_\alpha), \end{aligned} \tag{11}$$

and on the hyperparameters:

$$\begin{aligned} \phi_{ij} &\overset{iid}{\sim} \text{Unif}(0,1), \\ \lambda_\alpha &\sim \text{Unif}(-5,5), \\ 1/\tau_\alpha^2 &\sim \text{Unif}(0,5). \end{aligned} \tag{12}$$

We perform out-of-sample evaluation by holding out the 7 most recent data points for each pair of countries and producing posterior predictive distributions using the remaining 10 time points. As a point estimate of each predicted flow, $\widehat{y}_{ij}^{(t)}$, we adopt the median value of the posterior predictive distribution. We finally compare $\widehat{y}_{ij}^{(t)}$ with the corresponding true observed flow by means of the Mean Absolute Error (MAE)

$$MAE^{(t)} = \frac{1}{n} \sum_{ij} |y_{ij}^{(t)} - \widehat{y}_{ij}^{(t)}|,$$

where $y_{ij}^{(t)}$ is the observed count from province i to province j at time t and n is the number of different flows to be predicted in year t.

4 Results

The posterior distribution of the model parameters is obtained by combining prior information with the data. Since in general the proposed models are non-standard, meaning that no closed-form expression for the posterior distribution of parameters is available, Markov Chain Monte Carlo methods are implemented to obtain samples which are approximately drawn from the posterior of each model parameter. All analyses are performed using the `R` package `rjags`.

Parameters of interest consist of the regression coefficients for the EG models, solely the flow-specific standard deviations for BHPM 1, and the flow-specific intercepts and slopes for BHPM 2.

4.1 Testing migration theories

For each year t, we consider three alternative EG models which differ by the inclusion of one demographic variable among *(f)* share of people aged 20-30 year, *(g)* lagged TFR, and *(h)* percentage of foreigners. All other covariates (see Section 2.2.2) are instead common to the three models.

Figures 3 summarizes the posterior distributions of the regression coefficients relative to the EG model including the TFR *(g)*, across time. A similar behavior for these coefficients was observed under the other two EG models. In particular, for each coefficient we report the posterior mean (dark thick line) together with a 95% credible interval represented as a grey area. In addition, each dashed line represents an average (w.r.t. time) posterior mean of the coefficient, which then provides a graphical synthesis of the parameter fluctuations across time.

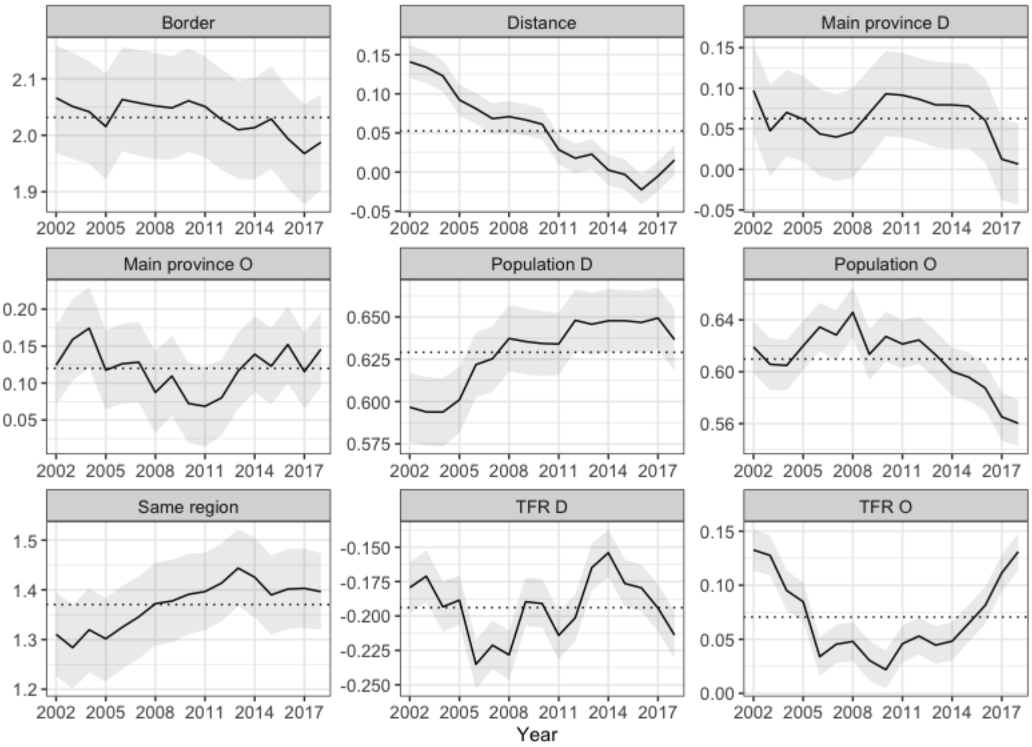

Figure 3: Posterior means of regression coefficients across time (solid black line) and 95% credible intervals. The average over the 17 considered years is represented as a dashed line.

The estimated coefficients for the geographical variables and the population size are fairly stable across time and across the three models. Note that, parameters associated with covariates "belonging to the same region" and "being contiguous provinces" are significantly different from zero and positive, suggesting that nearby provinces are characterized by more intense flows.

We now focus on coefficients associated with the three model-specific covariates, namely the "average Total Fertility Rate with a time-lag of 20 to 30 years" (Figure 3), the "share of people

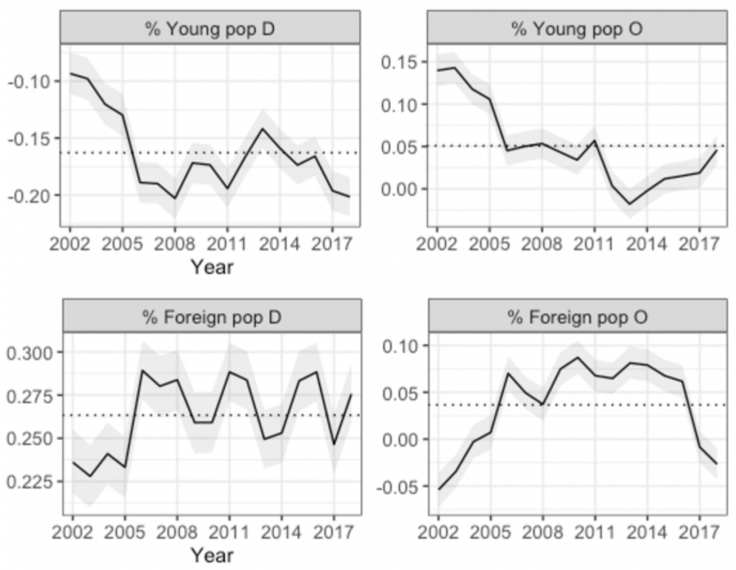

Figure 4: Posterior means of percentage of young population and percentage of foreign born population (for Origin and Destination provinces) and 95% credible intervals. The average over the 17 considered years is represented as a dashed line.

aged 20-30 year" (Figure 4, top two panels) and the "percentage of foreigners" (Figure 4, bottom two panels).

Results in Figure 3 underline the existence of a subnational replacement migration at province level, with larger internal migration flows targeting provinces with lower TFR. The top two panels of Figure 4 provide a different way of interpreting the same results, with larger flows targeting provinces with a lower shares of young population. Results in the bottom panels of Figure 4 highlight the importance of considering the level of foreign composition of the population. In fact, the percentage of foreign population seems to increase the attractiveness of a province with larger migration flows targeting provinces with a larger presence of foreigners.

Finally, we perform model checking by constructing the PIT histograms; see also Section 3. Results in Figure 5 refer to the EG model including the TFR *(g)*. The approximate uniform distribution of the histograms suggests a good level of calibration. The same behavior was observed for the alternative EG models.

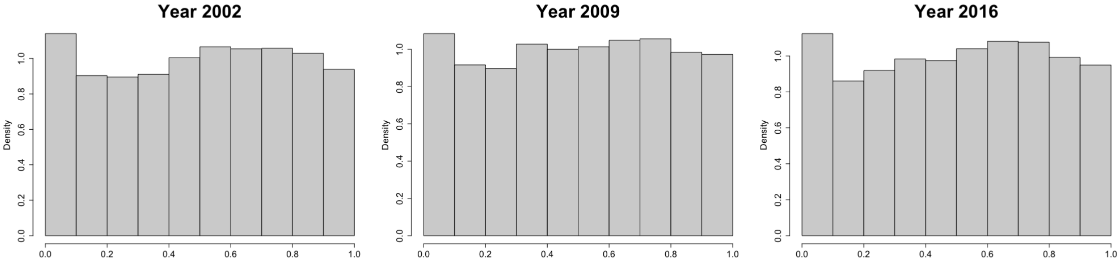

Figure 5: In-sample PIT histogram for the EG model including TFR and selected years.

4.2 Migration flows projections

For forecasting purposes, we consider the following models: the above EG model with the inclusion of a covariate representing the time, and the two AR(1) models BHPM1 and BHPM2. For comparison purposes, we also implement a constant model which project the migration flows using their historic mean and a persistence model which predicts the flow using the most recent observed value. We implement all the above models on years 2002-2011 and evaluate the models' performance w.r.t. years 2012-2018. For the latter purpose, we consider the out-of-sample mean absolute error (Table 2) as a measure of the quality of point forecasts; see also Section 3.2. In addition, to evaluate the accuracy of interval predictions, we adopt the interval coverage computed w.r.t. predictive distributions (Tables 3 and 4). The coverage at level $(1-\alpha)$, $\alpha \in [0, 1]$, of the prediction interval is computed as the number of observations which are contained in the $(1-\alpha)$-predictive credible interval. The closer to their nominal values, the better. The interval projections seem to achieve a good calibration, with coverages of the 80% and 95% prediction intervals being close to their nominal values (Table 3 and Table 4 respectively).

By looking at Table 2, we notice that the poorest performance is that of the EG model. In particular, although providing estimates which are close to the observed data points, it tends to underestimate the migration count. Differently, BHPM 1 and BHPM 2 show performances comparable with that of the persistence and constant models. The surprisingly good performance of the two competing models is probably due to the fact that the time-frame of the validation period is quite short, and changes in flows' size are relatively slow. This makes a prediction based on the most recent flow a reasonable estimate for the subsequent years.

Model	2012	2013	2014	2015	2016	2017	2018
Extended Gravity	25.96	23.45	23.09	22.80	23.22	24.67	25.14
BHPM 1	11.63	9.62	10.01	10.63	11.03	11.92	12.83
BHPM 2	11.4	8.81	9.34	10.11	10.93	12.2	13.77
Historic mean	13.01	11.03	10.98	11.11	11.43	11.43	11.45
Persistence	11.76	9.23	9.27	9.40	9.65	9.99	10.62

Table 2: Predictive performance of different methods: Mean Absolute Errors (MAE).

Figure 6 displays the observed values along with the forecast values for year 2012 to 2018 produced using model BHPM 2 for four selected flows. The red bars represent 80% predictive intervals, while the red dots represent the medians of the posterior predictive distribution. The dashed line represents the prediction given by the persistence approach. It is easy to see that BHPM 2 model follows the trend observed in the data, being able to capture the inversion in trends.

Model	2012	2013	2014	2015	2016	2017	2018
Extended Gravity	0.78	0.81	0.80	0.82	0.79	0.81	0.82
BHPM 1	0.80	0.89	0.91	0.91	0.91	0.9	0.9
BHPM 2	0.74	0.81	0.81	0.82	0.82	0.82	0.83
Historic mean	-	-	-	-	-	-	-
Persistence	-	-	-	-	-	-	-

Table 3: Prediction interval coverage (80%) for the Bayesian Extended Gravity model, and the BHPM.

Model	2012	2013	2014	2015	2016	2017	2018
Extended Gravity	0.93	0.92	0.95	0.96	0.94	0.97	0.97
BHPM 1	0.92	0.97	0.97	0.97	0.98	0.97	0.98
BHPM 2	0.91	0.94	0.93	0.93	0.94	0.93	0.93
Historic mean	-	-	-	-	-	-	-
Persistence	-	-	-	-	-	-	-

Table 4: Prediction interval coverage (95%) for the Bayesian Extended Gravity model, and the BHPM.

Since often migration flows data are not available on a yearly basis, we aggregate the flows by 3 years. We consider only BHPM 1 and the two baseline methods since they have the best fit. BHPM 2 is excluded from this second analysis because once we aggregate the years, we are left with very few data points. Table 5 shows the results for years 2013-2018 aggregated in three-year intervals. In 5, the persistence model is computed using the sum of years 2010 to 2012. Notice that the performance of the persistence model has worsened. This supports the idea that using year 2012 to predict up to year 2018 would provide a good prediction, but if we include older data points, the performance worsens.

Model	2013-2015	2016-2018
BHPM 1	20.35	23.84
Historic mean	25.5	27.03
Persistence:	20.17	23.55

Table 5: Predictive performance of different methods on three-year intervals: Mean Absolute Errors (MAE).

5 Discussion

Most of the studies on human migration flows focus on the migration from one country to another, however migrations within the same country may be important for understanding subnational dynamics. In this paper, we proposed the use of a Bayesian Hierarchical model to explain internal migration flows in Italy using a set of demographic and geographic covariates. We also proposed

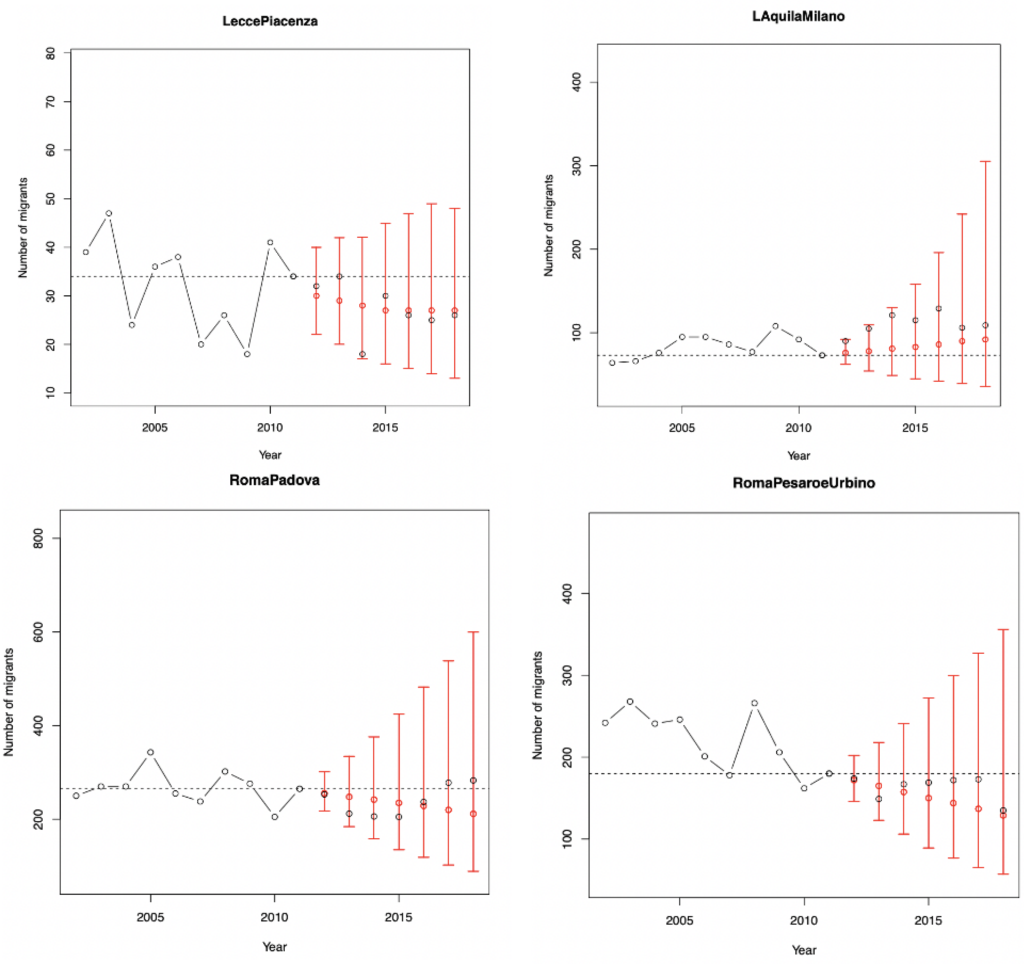

Figure 6: Selected pairs of provinces: true values vs. predicted values. Red bars represent 80% predictive intervals and red dots represent medians (obtained with BHPM 2). Black dots represent observed values, and the dashed line represents the prediction using the persistence model.

the use of autoregressive models to forecast internal migration flows.

Regarding the Extended gravity model, we notice that while most of the coefficients remain stable throughout time or do not show specific trends, distance dramatically loses significance throughout time, and it goes from being positively associated with flows to being close to zero. Our results on geographic variables suggest that, rather than distance, an important factor in internal migrations in Italy is constituted by the administrative boundaries of regions, with movements within regions being significantly larger than those outside regions. The use of variables which highlight cultural or administrative similarities rather than the mere use of a measure of geographical distance has been highlighted in other contexts as well. For example, contiguity, having had a common colonizer after 1945 and belonging to the same UN region was found to be a better predictor than distance in studying the correlation between TFR in different countries (Fosdick & Raftery, 2014).

We also find that TFR is a relevant predictor in internal migration flows. This finding is in line with the concept of replacement migration, highlighted in international migration literature (Billari & Dalla-Zuanna, 2012). The fluctuations of the TFR coefficients across time could be attributed to international migration, which is negatively correlated with TFR (the lower the TFR the higher the international migration), and positively correlated with the size of internal migration flows. We can therefore highlight two types of replacement migration taking place in Italy: the international replacement migration, which is not the focus of this paper, and a subnational replacement migration which we try to exploit in internal migration flows projections. Unfortunately, TFR is not available at the province level, but only at the region level, thus we consider it constant across provinces of the same region.

Moreover, our results, which show that a high presence of foreigners at destination is associated with larger flows, is in line with the literature on internal mobility of foreigners in Italy. In fact, Rimoldi et al. (2020) highlight that internal migrations of foreigners is positively associated with family commitment, scarce knowledge about the first destination, and no welcoming network, suggesting that foreign born population tend to move where there is more foreign population.

The proposed gravity model with TFR could be used for forecasting since the covariates in this setup do not need forecasting, being either time invariant or have being already forecast (population) or observed (TFR). However, the model performance suggests that if historic data is available, using them produces more accurate estimates. While keeping the structure of a Bayesian Hierachical Poisson model, based on the literature and on our results, we then developed several AR(1) models for forecasting migration flows.

BHPM 1 and BHPM 2 show out-of-sample performance which is comparable with that of the persistence and historic mean models, and they both perform better than the gravity model. However, autoregressive models require the existence of the time series of flows between two provinces, which not always is available.

The performance of the persistence and historic mean models remain slightly better than that of BHPM 1 and BHPM 2. This could be explained by the fact that migration dynamics change slowly, and we consider a relatively short time series, therefore using observations from 2011 to predict 2018 could indeed represent a good approximation. However, while this may be the case for Italy, other countries may benefit more from the proposed approach. The results presented at the aggregated level in Section 4, Table 5, do show a poorer performance of the persistence model, supporting this explanation.

Other approaches to the study and projection of internal migration flows, include the multiplicative component approach, which is based on in and out migration counts. These approaches have been used for forecasting internal migration flows by age and gender for projecting interregional migration in Italy and in Australia (Raymer et al., 2006, 2017, 2020).

Future research will address the extension of the BHP models to other countries, and the implications for sub-national population projections.

Acknowledgments

Raftery's research was supported by NICHD grant no. R01 HD070936.

References

Abel, G. J., Bijak, J., Forster, J. J., Raymer, J., Smith, P. W. & Wong, J. S. (2013). Integrating uncertainty in time series population forecasts: An illustration using a simple projection model. *Demographic Research* 29 1187–1226.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T. & Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography* 48 815–839.

Azose, J. J. & Raftery, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography* 52 1627–1650.

Banougnin, B. H., Adekunle, A. O., Oladokun, A. & Sanni, M. A. (2018). Impact of internal migration on fertility in Cotonou, Benin Republic. *African Population Studies* 32.

Bell, M., Charles-Edwards, E., Bernard, A. & Ueffing, P. (2017). Global trends in internal migration. In *Internal Migration in the Developed World.* Routledge, 76–97.

Bell, M., Charles-Edwards, E., Kupiszewska, D., Kupiszewski, M., Stillwell, J. & Zhu, Y. (2015). Internal migration data around the world: Assessing contemporary practice. *Population, Space and Place* 21 1–17.

Bernard, A., Bell, M. & Charles-Edwards, E. (2014a). Improved measures for the cross-national comparison of age profiles of internal migration. *Population Studies* 68 179–195.

Bernard, A., Bell, M. & Charles-Edwards, E. (2014b). Life-course transitions and the age profile of internal migration. *Population and Development Review* 40 213–239.

Bijak, J. (2010a). *Forecasting International Migration in Europe: A Bayesian View.* Dordrecht: Springer Netherlands.

Bijak, J. (2010b). Theory in migration forecasting: A global outlook. In *Forecasting International Migration in Europe: A Bayesian View*. Dordrecht: Springer Netherlands, 53–87.

Bijak, J. & Wiśniowski, A. (2010). Bayesian forecasting of immigration to selected european countries by using expert knowledge. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173 775–796.

Billari, F. C. & Dalla-Zuanna, G. (2012). Is replacement migration actually taking place in low fertility countries? *Genus* 67.

Coleman, D. (2013). The twilight of the census. *PoPulation and develoPment review* 38 334–351.

Congdon, P. (2010). Random-effects models for migration attractivity and retentivity: a bayesian methodology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173 755–774.

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data. *Biometrics* 65 1254–1261.

Daudin, G., Franck, R. & Rapoport, H. (2019). Can internal migration foster the convergence in regional fertility rates? evidence from 19th century france. *The Economic Journal* 129 1618–1692.

Faist, T. (2000). The volume and dynamics of international migration and transnational social spaces .

Fosdick, B. K. & Raftery, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demographic Research* 30 1011.

Gabrielli, G., Paterno, A. & White, M. (2007). The impact of origin region and internal migration on italian fertility. *Demographic Research* 17 705–740.

Greenwood, M. J. (1997). Internal migration in developed countries. *Handbook of population and family economics* 1 647–720.

ISTAT (2017). Migrazioni internazionali e interne della popolazione residente.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A. & Taylor, J. E. (1993). Theories of international migration: A review and appraisal. *Population and development review* 431–466.

Oberg, S. & Wils, A. (1992). East-west migration in europe: can migration theories help estimate the numbers? *Popnet* 1–7.

Raymer, J., Bai, X. & Smith, P. W. (2020). Forecasting origin-destination-age-sex migration flow tables with multiplicative components. In *Developments in Demographic Forecasting.* Springer, Cham, 217–242.

Raymer, J., Biddle, N. & Campbell, P. (2017). Analysing and projecting indigenous migration in australia. *Applied Spatial Analysis and Policy* 10 211–232.

Raymer, J., Bonaguidi, A. & Valentini, A. (2006). Describing and projecting the age and spatial structures of interregional migration in italy. *Population, Space and Place* 12 371–388.

Rees, P., Bell, M., Kupiszewski, M., Kupiszewska, D., Ueffing, P., Bernard, A., Charles-Edwards, E. & Stillwell, J. (2017). The impact of internal migration on population redistribution: An international comparison. *Population, Space and Place* 23 e2036.

Rimoldi, S., Barbiano di Belgiojoso, E. & Terzera, L. (2020). Internal mobility and family commitment of foreigners in italy. *International Migration* 58 168–183.

Stouffer, S. A. (1940). Intervening opportunities: a theory relating mobility and distance. *American sociological review* 5 845–867.

Stouffer, S. A. (1960). Intervening opportunities and competing migrants. *Journal of regional science* 2 1–26.

White, M. J. & Lindstrom, D. P. (2005). Internal migration. In *Handbook of population.* Springer, 311–346.

Willekens, F. (1994). Monitoring international migration flows in europe. *European Journal of Population/Revue européenne de Démographie* 10 1–42.

Zhang, J. L. & Bryant, J. (2020). Bayesian disaggregated forecasts: internal migration in iceland. In *Developments in Demographic Forecasting.* Springer, Cham, 193–215.

Zipf, G. K. (1946). The p1*p2/d hypothesis: on the intercity movement of persons. *American sociological review* 11 677–686.