Bayesian hierarchical modelling for migration flow size estimation: application to Italian data

Charlotte Taglioni, Brunero Liseo

Migration is a critical point for population estimation even in countries with reliable registers. Migration related problems are important to address as population mobility is constantly increasing, and registration of such movements is not always and not evenly reported. The differences in registering in and out migration is not the only one, there are also problems for specific sub-populations, like students or migrants within the European Community. Also, the phenomenon of illegal migration has relevant political implications and it is the centre of a significant part of the public and political debates in the recent years. Despite there are studies and estimations of the number of illegal migrant coming to Italy and to other European countries, this issue is out of the scope of this work. These applications only consider official data on resident population of Italy. The complexity of illegal migration is high, data quality is difficult to evaluate and results from application of the model would be difficult to interpret especially because the model investigation has not been completed yet and still needs improvements. In addition to the difficult data quality assessment and accuracy variability depending on time and countries, migration is a complex phenomenon also from other points of view. Especially for countries experimenting a zero or negative natural population growth, migration is an important issue from demographic but also sociological, political and economic perspectives. Therefore, despite migration is “the most complex and most difficult to predict component of population change, bearing high levels of forecast errors" (Kupiszewski, 2002) it is essential, especially for so-called “developed countries", to find methods to estimate and predict migration flows. Attempts to estimate or forecast migrations flows can be found, in several works adopting different approaches, and Bijak (2010) dedicated a book to migration in Europe. Raymer et al. (2013) address the problem of incoherence in migration flows registrations between countries. Their aim is to harmonise and estimate migration flows among 31 countries in the European Union and European Free Trade Association from 2002 until 2008. They integrate a theory-based migration model and a

measurement models from both sending and receiving countries. Using Eurostat data and a set of covariates, they model measurement errors considering imbalance between in- and outmigration and estimate under-counting country levels. Expert opinions are used for building prior distributions. Tests on sensitivity to prior information and to partial removal of the data are also accomplished along with a comparison with other approaches. In Congdon (2008) the Author compares the estimation of migration flows through a fully Bayesian and an estimation approach. He applies the method to the migration flows from Scotland to England during the 1990s. He uses the software “WinBugs" for the analysis and comments on benefits of the Bayesian approach and, specifically, on the random effect approach. The comparison between parametric and non-parametric approaches reveals how the first one performs well and gives good results for preliminary smoothing analysis, whereas the second one, having fewer constraints, is able to reveal details that are not so strongly empathised in the first one. An example of projections on net migration with very few data is Azose and Raftery (2016). They perform a Bayesian estimation of correlation matrices with informative priors and show how it outperforms Pearson correlation matrix and simple shrinkage estimators especially when the correlation matrix to estimate is sparse. Putting interpretable and simple priors on correlations is the main innovation of the method. An extension they suggest is to consider a matrix of bilateral migration instead of net migration.

Data on migration for Italy come from Istat and different datasets show differences in counts and dimensions. A first distinction is between migration from or to other countries, international migration, and migration within the country, internal migration. For internal migration, data have different level of detail from migration between municipalities to migration between regions. Municipalities collect data on registrations and cancellations from their registers and communicate them to Istat which publishes customised tables. Data on Istat website are complete in the sense that there is international and internal migration at different levels. There are data about origin and destination, but they only come with large age class groups ("0-17", "18-39", "40-64", "65+"). In order to have more details about migrants age, Eurostat data provide yearly age classes but only for international migration at national level. All datasets have data on migrant sex, and time span is 2006-2015. Note that registrations and cancellations refer to permanent

residences, it is then very likely that actual data on migration are much higher than what data report.

As mentioned, difficulties in estimating international flows are a common problem in many countries. Even in a country like New Zealand which is an island with better immigration records than in most of the countries, accuracy on migration data is considered “moderate" (Bryant and Zhang, 2018) whereas births and deaths registration have excellent accuracy. Italian data on migration refer to registrations and cancellations from municipality registers. International migration, especially at European level, is difficult to estimate. European laws and increasing mobility especially for students and workers make available data only partially trustworthy whereas illegal migration topic is not even addressed despite it is a major topic in Italy and Europe nowadays.

Internal migration analysis is similar to the international one but, apart from age effect, all the other effects and interactions do not show any clear pattern and, if they do, magnitude is quite low. Unlike for international migration which has clearer characteristics, a preliminary analysis easing the choice of system model for internal migration is difficult to provide. Clearly preliminary analyses give a glimpse of what could be the driving effects of a phenomenon, but they do not replace the proper estimation procedures.

Another aspect of migration is the format to describe it. There are four formats explained in Bryant and Zhang (2018) each one providing a different level of information. The most complete is the origin-destination format, all the movements are recorded in a square matrix with all the regions of origin and destination. This model provides information about both sending and receiving regions but it is a computationally demanding format. For example, considering the twenty Italian regions a matrix of four hundreds cells would be needed. Another way is the pool structure where only “total outward movements and total inward movements are shown for each status" (Bryant and Zhang, 2018). In this way the number of cells for Italian internal migration would be 40. A third format is the net format, it is efficient for population size estimation and only requires as many cells as status, i.e. twenty for Italy. Net migration

only gives the balance between immigration and emigration but it does not provide information about the size of the flows and, as net flows are usually much smaller than inward and outward flows, even small percentage changes in separate flows could produce large percentage changes in net flows. Data collected from the Istat website have pool format, they do not link origin and destination but only provide the number of registrations and cancellations. An origin-destination format can be obtained but the pool format, more parsimonious, has been chosen. Unlike net migration, pool format allows for separate immigration and emigration estimation but is not as computationally demanding as an origin-destination model. Migration is a complex phenomenon and to only estimate these series based only on the datasets available provide very partial results, especially because they are known not to be very accurate. For this reason migration is only estimated within the demographic account, where demographic balance consistence is always checked and hence provide results that ensure internal consistency. When only estimating migration series results tend stay closer to the data, but they might not reflect the actual situation as balance equation is not considered. The estimation of the flow is performed with a dedicated R package '*demest*' still under development.

Essential bibliography

Azose, J. J. and Raftery, A. (2016) Estimating large correlation matrices for international migration. arXiv.org 1605.08759.

Bijak, J. (2010) Forecasting international migration in Europe: a Bayesian view. 24. Springer.

Bryant, J. and Zhang, J. (2018) Bayesian demographic estimation and forecasting. Chapman and Hall/CRC.

Bryant, J. R. and Graham, P. J. (2013) Bayesian demographic accounts: subnational population estimation using multiple data sources. Bayesian Analysis 8, 591-622.

Congdon, P. (2008) Models for migration age schedules: a Bayesian perspective with an application to flows between Scotland and England. In International migration in Europe: data, models and estimates, eds J. Raymer and F. Willekens. Wiley and Sons.

Kupiszewski, M. (2002) The role of international migration in the modelling of population dynamics. Warsaw: Institute of Geography and Spatial Organisation, Polish Academy of Sciences.

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