**Probabilistic Methods For Combining Traditional and Social Media Bilateral Migration Data**

Dilek Yildiz, Arkadiusz Wisniowski, Guy Abel, Cloé Gendronneau, Martin Stepanek, Ingmar Weber, Emilio Zagheni, Stijn Hoorens

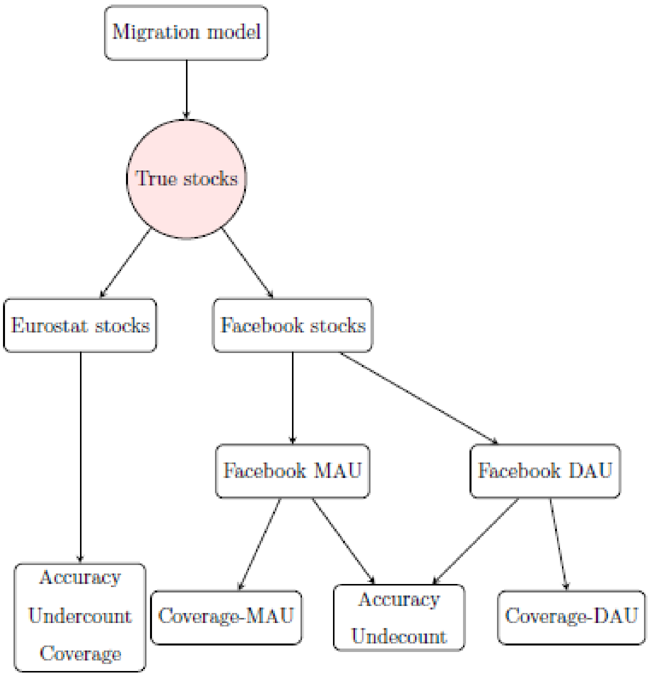
Bayesian methods offer a powerful mechanism to combine data sources. Previously models have been developed for solely combining traditional migration data sources using the prior models (see e.g. Bijak et al., 2010; Raymer et al. 2013; Wiśniowski et al. 2013; Wisniowski, 2017; Wisniowski et al., 2016). We adapt the basic methodologies of these former models to combine migration data from both traditional and new data sources derived from social media.

We first focus on models to estimate true bilateral migrant stocks. This choice is motivated by a number of factors. First, migrant stocks are more easily measured by national statistical agencies, and are more widely available for each migration corridor. Second, migrant stock data is far more uniformly defined in comparison to flow estimates. Third, non-traditional data sources on migrant stocks are more easily obtained than flow estimates and with reduced biases. For example, proxies for estimates of migrant stocks can be obtained directly from Facebook advertising platform.

Our modelling framework to combine migration stock data is shown in

Figure 1. Each layer of the Figure illustrates a hierarchy of our Bayesian model. Towards the bottom is a level based on the reported data from Eurostat and Facebook, the data inputs into the model. Below each data source are factors that drive the level and variation of the reported data through systematic and random errors respectively.

Figure 1 Conceptual framework for Bayesian hierarchical model for EU migration stocks



The level of migration in reported data tends to be systematically lower than the true level of migration. In order to obtain an estimate of the true migration quantity, our model adjusts reported migration flows using meta-data on the undercount and coverage of the reported migrant stocks. Coverage provides a fraction of the total population that believed to be covered by each data source.

The coverage quantity in Eurostat stocks reflects the share of total population that were targeted by national statistical authorities. For Facebook data the coverage parameter relates to the penetration of the social platform on the general population. Undercount provides a fraction of the migrant stocks that is believed to be missing for each data source. For Eurostat data this reflects the share of the migrants that were not enumerated. For Facebook data this reflects the share of Facebook users living in a different country than their origin country and who do not provide location details.

The variation in reported migration can be driven by a number of random errors. The size of the noise of the measure of reported migration quantities in each data source can vary according to a number of factors. For example with traditional migration data, reported data from Census or administrative data bases tend to be more accurate than reported data based on surveys. For Facebook migrant measures the accuracy levels vary much greater in reported data based on Daily Active Users (DAU) than for reported data based on Monthly Active Users (MAU). In all data sources, the variation in the reported data is related to the size of the underlying migration quantity, for example, larger migrant stocks have greater associated margins of error.

Each measurement parameters (the undercount, coverage and accuracy) require prior distributions. Current model (initial results plotted in Figure 2 – 4) assumes that all parameters are equal and negligible, to allow us to first concentrate on implementing an initial computation framework. We recently began to include meta information on data sources to provide a far more robust estimate of true migration quantities. The current formulation of prior distributions derives averages of migrant stocks in each migration corridor, as no data source is weighted (via the measurement parameter) more heavily than another.

The reported data, influenced by their measurement parameters are driven by the unobserved migrant stocks that sit above them in Figure 1 of the Bayesian hierarchical modelling framework. The migrant stocks themselves are driven by a migration model that sits at the top of the hierarchy. The migration model reflects migration theory on migrant stocks. In our initial model we assumed a migration model that allows the level of migrant stocks to vary in each migration corridor. The parameters in this model, that drive the estimation of the true stocks are ultimately obtained from information in the reported data and prior distributions for the measurement parameters, with their information fed up the hierarchy.

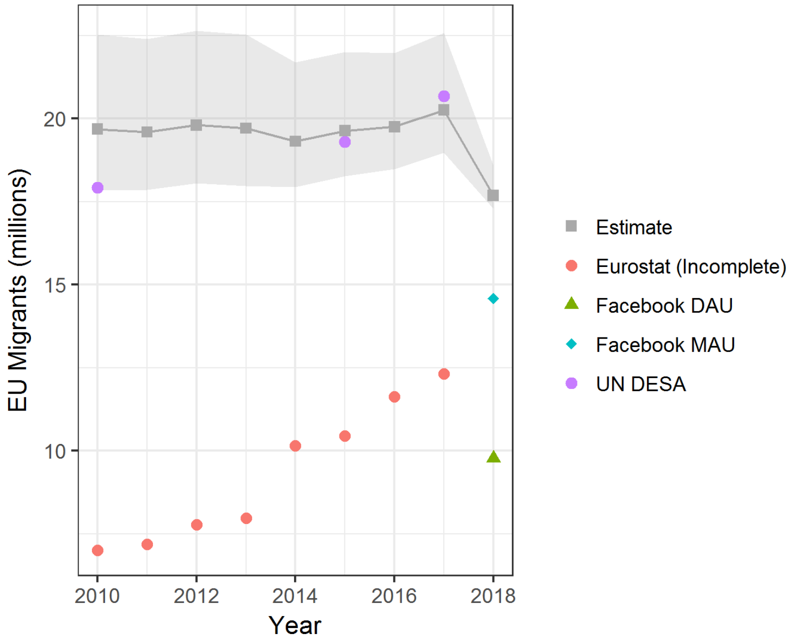
The full modelling framework has been implemented in open source Bayesian software, JAGS, operated from within the “R” environment (Plummer et al, 2003). JAGS split each part of the model into its hierarchical components and underlying sub-model components. The migration model parameters, that are used to derive the estimates of true migrants, as well as the measurement parameters, that are highly influenced by the prior distributions base on meta data, are simultaneously estimated using Markov chain Monte Carlo (MCMC) methods.

## Initial Results

In our initial model we used a simple migration model that allows variation in the level of migrants in each migration corridor to vary alongside equal prior distributions for the measurement parameters for the undercount, coverage, undercount and accuracy.

Figure 2 plots the overall level of EU migrants in another EU country. We used reported data from Eurostat (2010-2017) and Facebook data for 2018. The totals based on the Eurostat reported stocks do not include reported data on many migration corridors – these are not reported to Eurostat by the countries, and hence a direct comparison of the Eurostat reported data and the estimates from the model should not be made.

**Figure 2 Estimate of total EU migrants living in EU countries using Eurostat, Facebook and UN data with 60% Prediction Interval**



Included in the plot, but not in the model, are the reported data from the UN DESA[1] on the number of EU migrants in EU countries (data available in 2010, 2015 and 2017). Each of these observations is within our prediction interval for the true flow shown by the grey region (between 20th and 80th quantile). Totals from both Facebook data sources are much lower than our estimates. This is primarily due to their lower coverage in comparison to Eurostat reported data, where meta information on such quantities has not yet been incorporated into the model. These lower values drive the dips in the estimated flows for the final year (2018).

Figure 3 shows the estimated number of EU migrant totals in each EU country alongside reported data. The location of each country is arranged by broad geographic location. In countries such as Germany, where there is no reported Eurostat data on the number of foreign born migrants (by birthplace) the estimated migrants in earlier years closely match the levels of the Facebook data. In countries such as France where earlier Eurostat data is not available, there is greater uncertainty in the estimated migrant levels.

**Figure** 4 shows estimates for the number of EU migrants abroad by each EU country of birth (arranged by broad geographic location of the country of birth). As Eurostat data is collected in the country of residence, not birth, the totals for the foreign born populations is incomplete in all countries, as there is at least one country in each year that does not collect data). As with Figure 3, estimates are relatively constant over time – due to the current lack of temporal parameters in our migration model - except with dips in the last period where Facebook data is available.

EU countries with large numbers of their population elsewhere in the EU, such as Poland, Romania and Portugal are clearly visible from the estimates. Their numbers are greater than those reported in the partially available Eurostat data as the estimates of the true migrant stocks are based on information over all time periods. As the model combines data, it takes advantage of the completeness in availability of Facebook MAU data for migrant stock estimates in all corridors.

[1] http://www.un.org/en/development/desa/population/migration/data/estimates2/estimates17.shtml

**Figure 3 Total EU migrants in each EU country. Displays arranged by broad geographic location**

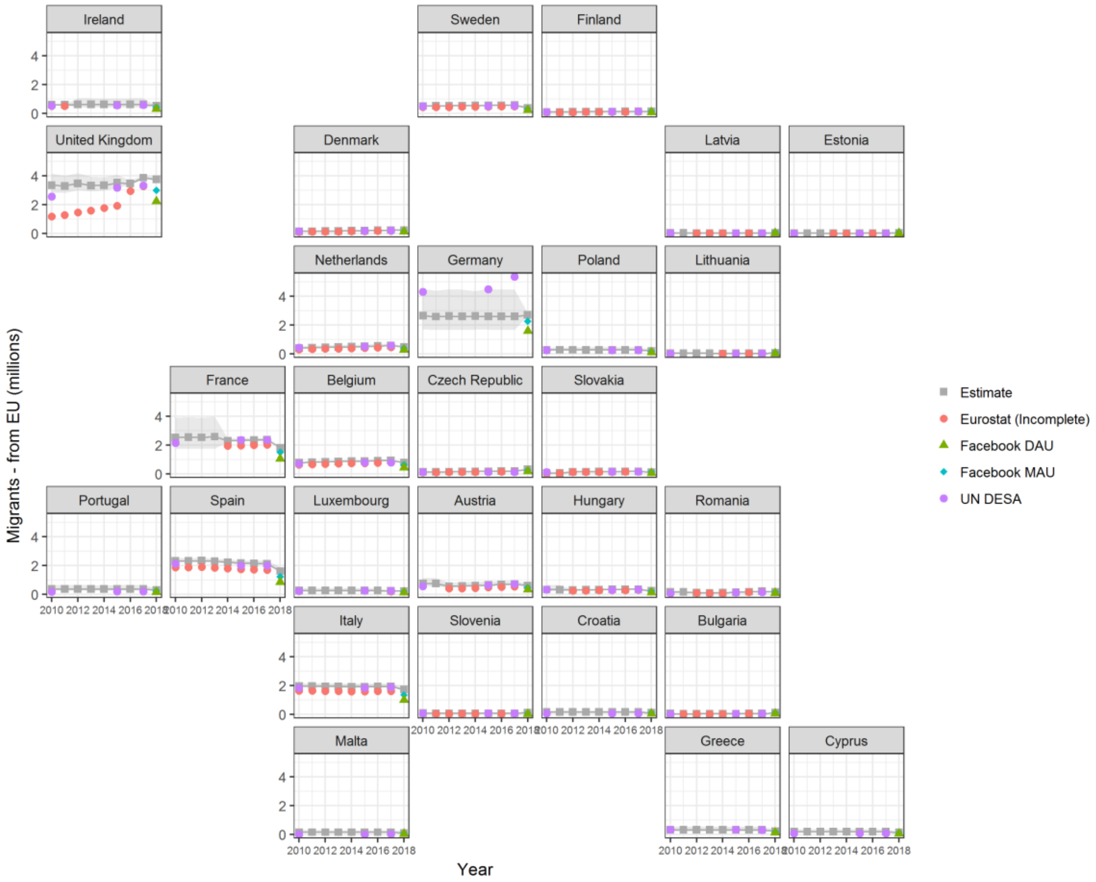
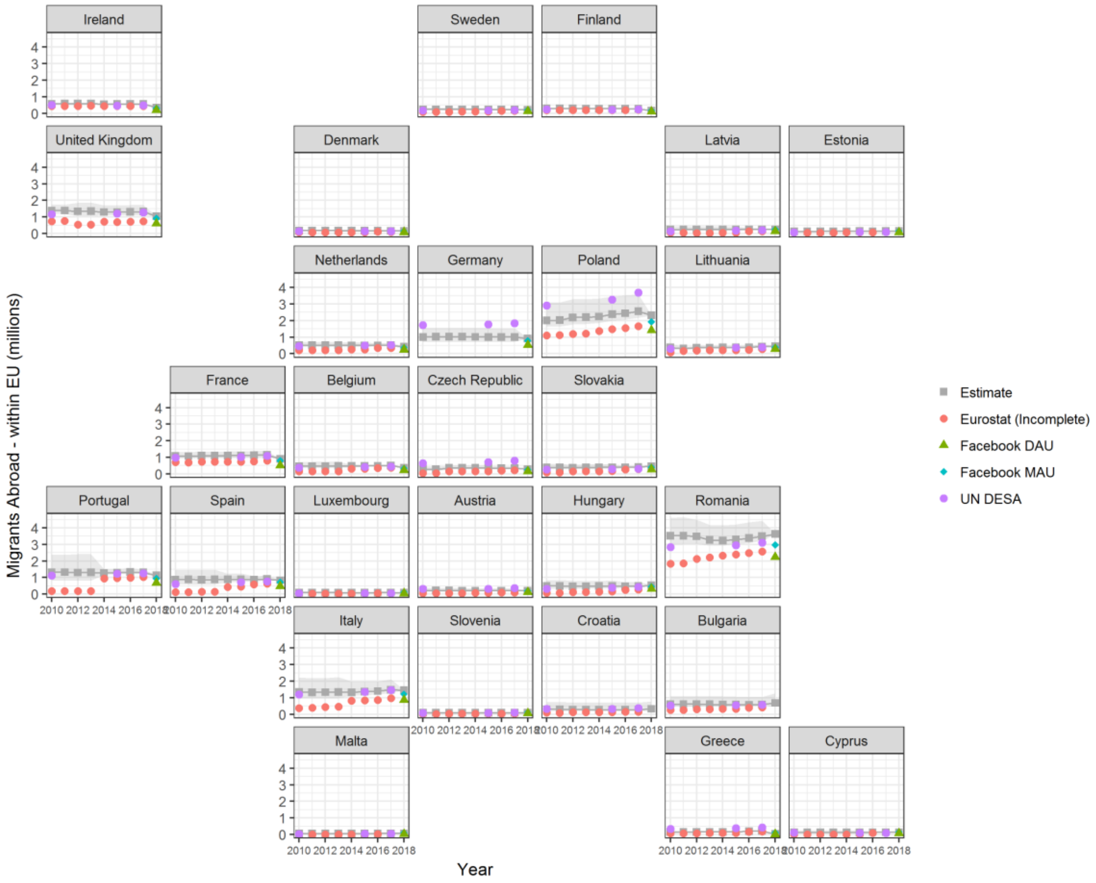


**Figure 4 Total EU migrants abroad from each EU country.**


Ireland
Sweden
Finland
United Kingdom
Denmark
Latvia
Estonia
Netherlands
Germany
Poland
Lithuania
France
Belgium
Czech Republic
Slovakia
Portugal
Spain
Luxembourg
Austria
Hungary
Romania
Italy
Slovenia
Croatia
Bulgaria
Malta
Greece
Cyprus
Migrants Abroad - within EU (millions)
Year
Estimate
Eurostat (Incomplete)
Facebook DAU
Facebook MAU
UN DESA