Data and Knowledge for Modelling Asylum Migration

Sarah Nurse, Jakub Bijak

21st May 2019

1 Introduction

The overarching aim of this project is to develop a simulation model of a real migration flow, based on a population of interacting cognitive agents, which would integrate different aspects of behavioural and social theory with formal modelling. To remain true to the empirical roots of demography as a social science discipline, the model would need to be as much grounded in the observed social reality as possible (Courgeau et al., 2016). In this context, it is important to choose a migration test case with a large enough flow of migrants and a broad range of available information sources of data in order to allow investigation of the different theoretical and methodological aspects of the migration processes by formally modelling their behaviour. Knowing the different aspects of data collection and quality of information, and reflecting them in the model become important elements of the modelling endeavour.

In this paper, we look at recent Syrian migration to Europe (2011–2017) through the lens of the available data sources, and propose a unified framework to assess the different aspects in which the data may be useful for modelling. Given the nature of the task, the data requirements for complex migration modelling are necessarily multi-dimensional, and are not limited to migration processes themselves, additionally including a range of the underpinning features and drivers. Our aim is to collate as much information as possible on the chosen case study for use in the modelling, and to assess its quality and reliability in a formal way, allowing for an explicit description of data uncertainty. In this way it is possible to make use of all available relevant information whilst taking into account the relative quality when deciding on the level of importance with which the data should be treated.

In this paper, after briefly introducing the recent asylum-related migration from Syria (Section 2), we summarise the key conceptual challenges related to asylum migration and its measurement (Section 3). Subsequently, a brief overview of the available data is provided, with a distinction between the sources related to the migration *process*, as well as to the *context* within which migration occurs (Section 4). Subsequently, a framework for assessing different aspects of data is proposed, based on a review of similar approaches suggested in the literature (Section 5). The paper concludes by making specific recommendations for using the different forms of data in formal modelling, including in the formal uncertainty assessment (Section 6).

Working paper No. 1 of the project "Bayesian Agent-based Population Studies", prepared for the workshop on Uncertainty and Complexity of Migration, London, 20-21 November 2019. The project is funded by the European Research Council under the grant ERC-CoG-2016-725232.

The paper is accompanied by a Metadata document, which lists the key sources of data on Syrian migration and its main drivers. The listing includes details on the data types, content and availability, as well as a multidimensional assessment of their usefulness for migration models, following the analytical framework introduced in this paper.

2 Background: Asylum migration from Syria to Europe

Large-scale Syrian migration has a distinct start, following the widespread protests in 2011 and the outbreak of the civil war. After more than a year of unrest, in June 2012 the UN declared the Syrian Arab Republic to be in a state of civil war, which continues seven years later. Whereas previous levels of Syrian emigration remained relatively low, the nature of the conflict involving multiple armed groups, government forces and external nations has resulted in an estimate of 6.4 million people fleeing Syria since 2011 and a further 6.8 million internally displaced according to the UN High Commissioner for Refugees (UNHCR), at the end of 2017 (UNHCR, 2019).

Initial scoping of the work suggests the availability of a wide range of different types of data that have been collected on the recent Syrian migration into Europe. In particular, the key UNHCR datasets show the number of Syrians who were displaced each year, as measured by the number of registered asylum seekers, refugees and other "persons of concern", and the main destinations of asylum seekers and refugees who have either registered with the UNHCR or applied for asylum. The information is broken down by basic characteristics, including age and sex and location of registration, distinguishing people located within refugee camps and outside.

As shown in Figure 1, neighbouring countries in the region (chiefly Turkey, Lebanon and Jordan, as well as Iraq and Egypt) feature heavily as countries of asylum, together with a number of European destinations (in particular, Germany and Sweden). The scale of the flows, as well as the level of international interest and media coverage means that the development of migrant routes and strategies have often been observed and recorded as they occur. In addition to the main official sources, such as the UNHCR, many other pieces of information deal with some very specific aspects of Syrian flows and their drivers. These sources are presented in more detailed in the Metadata document, together with an assessment of their suitability for modelling.

3 Asylum migration: Some conceptual challenges

Even though one of the central themes of the computational modelling endeavour is to reflect the complexity of migration, the theoretical context of our understanding of population flows has traditionally been relatively basic. Within the most existing frameworks, decisions are based on structural differentials, such as employment rates, resulting in observed overall migration flows (for reviews, see e.g. Massey et al., 1993; Bijak, 2010). In his classical work, Lee (1966) aims to explain the migration process as a weighing up of factors or 'drivers' which influence decisions to migrate, while Zelinsky

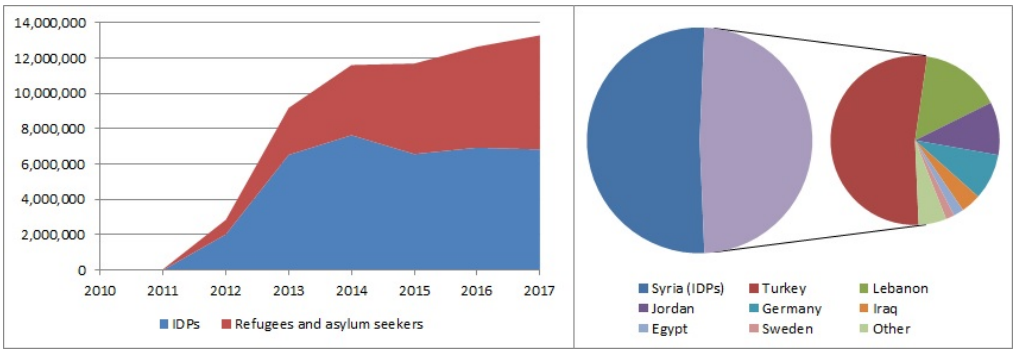

Figure 1: Number of Syrian asylum seekers, refugees, and IDPs, 2011–2017, and the distribution by country in 2017. Source: UHNCR

(1971) describes different features of a 'mobility transition', which can be directly observed. Most of the traditional theories do not reflect the complexity of migration, are largely fragmented along disciplinary boundaries (Arango, 2000), and in most cases fail to link the macro- and micro-level features of the migration processes, which is a key gap that needs addressing.

More recently, there have been attempts to move the conceptual discussion forward. A contemporary 'push-pull plus' model (Van Hear et al., 2018) adds complexity to the original theory of Lee (1966), but fails to provide a framework that can be operationalised in an applied empirical context. The capability framework of Carling and Schewel (2018) stresses the importance of individual aspirations and ability to migrate, but again fails to map the concepts clearly onto the empirical reality.

In the context of displacement or forced migration, the conceptual challenges only get amplified. As noted by Suriyakumaran and Tamura (2016) and Bijak et al. (2017), operationalisation of the conceptually-complex theories of asylum migration is typically reduced to identifying a selection of available drivers to include in explanatory models. Here, it has to be noted that the presence of underlying structural factors or 'pre-conditions' for migration is itself not a sufficient driver of migration; very often, migration occurs following accumulation of adverse circumstances, and some trigger events, either experienced or learnt about through networks or media. For that reason, the monitoring of the underlying drivers, such as the conflict intensity, becomes of paramount importance (Bohra-Mishra and Massey, 2011).

Another crucial concept to consider when modelling the migration *process* is how difficult different routes are for migrants undertaking a journey. In particular, it is important, whether a prospective route includes crossing national borders, whether those borders are patrolled, whether there is a smuggling network already operating and whether an individual has access to the information and resources necessary to navigate these are all examples of barriers that can exist for migrants. As an overall summary measure or perception for decision-making this can be thought of as a route's 'friction' (for a general discussion related to migration, see Stillwell et al., 2016). This can include both formal barriers, such as national borders and visa restrictions, or informal barriers, such as geographic distance or physical terrain.

4 Data overview

In the proposed approach, we suggest to follow a two-stage process of data assessment for modelling. The first stage of this process is to identify all available data relevant to the different elements involved in the decision making and migration flows being modelled. The second stage is then to introduce an assessment of uncertainty so that this can be formally taken into account and incorporates into the model.

Depending on the purpose and the intended use in different parts of the model, the data sources can be classified by type; broadly these can be viewed as providing either process-related or contextual information. The distinction here is made between data relating specifically to the migration process, including the characteristics of migrants themselves, their journey and decisions on the on hand, and contextual information, which covers the wider situation at the origin, destination and transit countries, on the other. Relevant data on context can include for example macro-economic conditions, the policy environment and the conflict situation in the country of origin.

In addition, in order to allow the data to be easily accessed and appropriately utilised in the model, the sources can be further classified depending on the level of aggregation (macro or micro), as well as paradigm under which they were collected (quantitative or qualitative). These categories, alongside a description of source type (for example, registers, surveys, censuses, administrative or operational data, journalistic accounts, or legal texts) are the key components of meta-information related to individual data sources, and are useful for comparing similar sources during the quality assessment stage.

4.1 Process-related data

Among the process-related data, describing the various features of migration flows, be it for individual migrants (micro level) or for the whole populations (macro level), the following type of information can be particularly useful for modelling:

Origin population Information on the pre-conflict Syrian population, such as data from a census or health surveys can be used for benchmarking. Data on age and sex distributions as well as other social and economic characteristics can be helpful in identifying specific subpopulations of interest, as well as in allowing for heterogeneity in the populations of migrants and stayers.

Registrations Administrative and operational information from destination countries and international or humanitarian organisations, which register the arrival of migrants, can provide data on numbers and characteristics as well as the timing of arrivals. These data also have clearly specified definitions due to their explicit collection purposes.

Destination surveys Wide range of data on migrant characteristics, economic situation (employment, benefits), access to and use of information, intentions, health and well-being at the destination countries can be used for reconstructing various elements of migrant journeys, and assessing the situation of migrants at the destination. Note that with respect to migration processes, these data are typically retrospective.

Resources Information on the level of resources that are required for the journey, including availability of humanitarian aid, or intricacies of the smuggling market, as well as information on migrant access to resources, can provide additional insights into the migration routes and trajectories. Resources typically deplete over time and journey, which again impacts on decisions by determining the route, destination choice, and so on.

Information Finally, availability of information on routes and contextual elements can also impact on migrants' decisions. Even though the information itself can be contextual, its availability and trustworthiness are related to the migration process.

4.2 Contextual data

Formal modelling offers a possibility to incorporate a wide range of different types of contextual data shaping the migration decisions through the environment in which the migration processes take place. The list below is by no means exhaustive, and it concentrates on the four main aspects of the context – related to the origin, destination, routes, and policies.

Origin context Information on the situation in the countries and regions of origin, including such factors like conflict intensity, the presence of specific events or incidents, as well as reports from observers and media, identify the key drivers related to the decision to migrate (corresponding to *push factors* in Lee's theoretical framework).

Destination context At the other end of the journey, information on destination countries, such as macro-economic data, attitudes and asylum acceptance rates, provide contextual information on the relative attractiveness of various destinations (corresponding to *pull factors*).

Routes and journey Contextual data on terrain, networks, barriers and transport routes can be used to assess different and variable levels of friction of distance, which can have long- and short-term impact on migration decisions and on actual flows (corresponding to *intervening obstacles*).

Policies Specifically related to the destination context, but also extending beyond it, the information on various aspect of migration policy and law enforcement, including visa, asylum and settlement policies in destination and transit countries, as well as their changes in response to migration, additionally helps paint a more complete picture of the dynamic legal context of migrant decisions and of their possible interactions with those of other actors (border agents, policy makers, and so on).

The multidimensionality of migration results in a patchwork of sources of information covering different aspects of the flows and the context in which they are taking place, often involving different populations and varying accuracy of measurement, which can be combined with the help of formal modelling (Willekens, 1994). At the same time, it implies the need for greater rigour and transparency, and a careful consideration of the data quality and their usefulness for a particular purpose, such as modelling. The key tenets of the proposed quality assessment exercise are discussed next.

5 Quality assessment

No perfect data exist, not least on migration processes. The measurement of asylum migration requires particular care, going beyond the otherwise challenging measurement of other forms of human mobility (see e.g. Willekens, 1994). The most widespread ways to measure asylum migration processes involve administrative data on events, which include very limited information about the context (Singleton, 2016). Other, well-known issues with the statistics involve duplicated records of the same persons, for whom multiple events have been recorded, as well as the presence of undercount due to clandestine nature of many asylum-related flows (Vogel and Kovacheva, 2008). The use of asylum statistics for political purposes adds another layer of complexity, and necessitates extra care when interpreting the data (Bakewell, 1999). For all these reasons, when describing migration flows through modelling, multiple data sources ideally need to be used concurrently, and be subject to formal quality assessment, as set forth below.

5.1 Existing frameworks

Assessing the quality of sources can allow us to make use of a greater range of information that may otherwise be discarded. Trustworthiness and transparency of data are particularly important for a politically sensitive topic of migration against the backdrop of armed conflict at the origin, and political controversies at the destination.

Existing studies indicate several important aspects in assessing the quality of data from different sources. A key recent review of survey data specifically targeting asylum migrants has been compiled by Isernia et al. (2018) and gives a broad overview, as well as some specific elements to be considered. An initial discussion of how surveys were selected for this review highlights definitional issues with identifying datasets with the appropriate target population. Aspiring to clarity in definitional issues is an enduring theme in migration studies, to which asylum migration is no exception (Bijak et al., 2017).

There are several examples of existing studies in related areas, which aim at assessing the quality of sources of information. Specifically in the context of irregular migration, Vogel and Kovacheva (2008) proposed a four-point assessment scales for various available estimates, broadly following the 'traffic lights' convention (*green, amber, red*), but with the red category split into two subgroups, depending on whether the estimates were of any use or not. Recently, the traffic-lights approach has been used by Bijak et al. (2017) for asylum migration, and was based on six main assessment criteria: (1) Frequency of measurement; (2) Fit with the definitions; (3) Coverage in terms of time and space; (4) Accuracy, uncertainty and the presence of any biases; (5) Timeliness of data release; and (6) Evidence of quality assurance processes. Similar assessments have been also previously carried out in the broader demographic studies of the consequences of armed conflict (GAO, 2006; Tabeau, 2009; Bijak and Lubman, 2016), presenting additional suggestions for how to address the various challenges of measurement.

5.2 Proposed dimensions of data assessment

The aim and nature of the modelling process imply that, while clarity of definitions is important, it is also possible to encompass a wider range of information sources and to assign different relative importance to these sources in the model. Our proposal for an assessment framework and uncertainty measures for different types of data is presented below. In particular, we propose six generic criteria for assessment:

1. Purpose for data collection and its relevance for modelling;
2. Timeliness and frequency of data collection and publication;
3. Trustworthiness and absence of biases;
4. Sufficient levels of disaggregation;
5. Target population and definitions including the population of interest (in our case, Syrian migrants);
6. Transparency of the data collection methods.

The need to identify the target population clearly is common for all types of data on migrants, but there are additional quality criteria specific to register- and survey-based sources. Thus, for register-based information an additional criterion is related to its completeness, while for surveys, their design, sampling strategy, sample sizes, and response rates are all aspects which need to be clearly set out in order to be assessed for rigour and good practice in the data collection (Isernia et al., 2018).

In our framework, all criteria are evaluated according to a five-point scale, based on the traffic-lights approach (green, amber, red), but also including half-way categories (green-amber and amber-red). The specific classification descriptors for assigning a particular source to a given class across all the criteria are listed in Table 1. Finally, for each source, a summary rating is obtained by averaging over the existing classes.

The result of an application of these seven criteria to 24 data sources potentially relevant to modelling Syrian migration is presented in detail in the Metadata document. The listing additionally includes 19 supplementary, general-level sources of information on migration processes, drivers or features, some aspects of which may also be useful of modelling, but which are unlikely to be at the core of the modelling exercise, and therefore have not been assessed following the same framework. For the latter group of sources, only generic information about source type and the purpose of collection is provided, alongside a basic description and access information.

6 Recommendations for modelling

One important consideration when choosing data to aid modelling is that the information used needs to be subsidiary to the research or policy questions that will be answered through models. For example, consider the questions about the journey (*process*), such as on whether migrants choose the route with the shortest geographic distance, or is it mitigated by resources, networks and access to information? Exploring possible answers to this question would require gathering different sources of data around the general concept

of 'friction', and would allow the modeller to go far beyond standard geographic measures of distance.

The arguments presented above lead to three main recommendations regarding the use of data in the practice of formal modelling.

First, there are no perfect data, so the expectations related to using them need to be realistic. There may be important trade-offs between different sources in terms of various evaluation criteria. For this reason, any data assessment has to be multidimensional, as different purposes may imply focus on different desired features of the data.

Second, any source of uncertainty, ambiguity or other imperfection in the data has to be formally reflected and propagated into the model. A natural language for expressing this uncertainty is one of probabilities, such as in the Bayesian statistical framework (see e.g. Bijak et al., 2017);

Third, the context of data collection has to be always borne in mind, and the use of particular data needs to be ideally driven by the specific research and policy requirements rather than mere convenience.

One key extension of the formal evaluation of various data sources is to investigate the importance of the different pieces of knowledge on migration processes and context, and to address the challenge of coherently incorporating the data on both micro and macro level processes, as well as the contextual information, together with their uncertainty assessment, in a migration model. If that could be successfully achieved, the results of the modelling can additionally help identify the future directions of data collection, strengthening the evidence base behind asylum migration and helping shape more realistic policy responses.

References

Arango, J. (2000). Explaining Migration: A Critical View. *International Social Science Journal*, 52:283–296.

Bakewell, O. (1999). Can we ever rely on refugee statistics? *Radical Statistics*, 72(1):`http://www.radstats.org.uk/no072/article1.htm`.

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

Bijak, J., Forster, J. J., and Hilton, J. (2017). Quantitative assessment of asylum-related migration: A survey of methodology. Report, European Asylum Support Office, Malta.

Bijak, J. and Lubman, S. (2016). The Disputed Numbers: In Search of the Demographic Basis for Studies of Armenian Population Losses, 1915–1923. In Demirdjian, A., editor, *The Armenian Genocide Legacy*, pages 26–43. Palgrave, London.

Bohra-Mishra, P. and Massey, D. S. (2011). Individual decisions to migrate during civil conflict. *Demography*, 48(2):401–424.

Carling, J. and Schewel, K. (2018). Revisiting aspiration and ability in international migration. *Journal of Ethnic & Migration Studies*, 44(6):945–963.

Courgeau, D., Bijak, J., Franck, R., and Silverman, E. (2016). Model-Based Demography: Towards a Research Agenda. In Van Bavel, J. and Grow, A., editors, *Agent-Based Modelling in Population Studies: Concepts, Methods, and Applications*, pages 29–51. Springer, Cham.

GAO (2006). Darfur Crisis: Death Estimates Demonstrate Severity of Crisis, but Their Accuracy and Credibility Could Be Enhanced. Report to Congressional Requesters GAO-07-24, US Government Accountability Office, Washington DC.

Isernia, P., Urso, O., Gyuzalyan, H., and Wilczyńska, A. (2018). A Review of Empirical Surveys of Asylum-Related Migrants. Report, European Asylum Support Office, Malta.

Lee, E. S. (1966). A Theory of Migration. *Demography*, 3(1):47–57.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1993). Theories of International Migration: Review and Appraisal. *Population and Development Review*, 19(3):431–466.

Singleton, A. (2016). Migration and Asylum Data for Policy-Making in the European Union – The Problem with Numbers. CEPS Paper 89, Centre for Europe and Policy Studies, Brussels.

Stillwell, J., Bell, M., Ueffing, P., Daras, K., Charles-Edwards, E., Kupiszewski, M., and Kupiszewska, D. (2016). Internal migration around the world: comparing distance travelled and its frictional effect. *Environment and Planning A: Economy and Space*, 48(8):1657–1675.

Suriyakumaran, A. and Tamura, Y. (2016). Asylum provision: A review of economic theories. *International Migration*, 54(4):18–30.

Tabeau, E. (2009). Victims of the Khmer Rouge Regime in Cambodia, April 1975 to January 1979: A Critical Assessment of Existing Estimates and Recommendations for Court. Expert report, Extraordinary Chambers of the Courts of Cambodia, Phnom Penh.

UNHCR (2019). UNHCR Population Statistics: Persons of Concern. `http://popstats.unhcr.org/en/persons_of_concern`. Accessed on: 1 May 2019.

Van Hear, N., Bakewell, O., and Long, K. (2018). Push-pull plus: Reconsidering drivers of migration. *Journal of Ethnic & Migration Studies*, 44(6):927–944.

Vogel, D. and Kovacheva, V. (2008). Classification report: Quality assessment of estimates on stocks of irregular migrants. Report of the CLANDESTINO project, Hamburg Institute of International Economics, Hamburg.

Willekens, F. (1994). Monitoring International Migration Flows in Europe: Towards a Statistical Data Base Combining Data from Different Sources. *European Journal of Population*, 10(1):1–42.

Zelinsky, W. (1971). The hypothesis of the mobility transition. *Geographical Review*, 61(2):219–249.

Criteria	Green	Amber	Red
Purpose: Is the purpose for collecting the data relevant to and appropriate for our use?	Yes: aim is to estimate and/or understand migration from Syria	May be different purpose but still relevant	No: data collection for different purpose, impacting usefulness
Timeliness: Are the data published at sufficiently frequent intervals?	Yes: repeated measures published regularly	May be repeated measures but with long gaps/publication delay	No: one-off collection or long delay in publication
Trustworthiness: Is the source free from obvious biases or stated political aims?	Yes: evidence of impartiality	Unclear or unstated	No: clear evidence of bias
Disaggregation: Is there sufficient geographic and country of origin detail?	Yes: country of origin and destination fully disaggregated	Partial disaggregation e.g. for some variables of interest	No: not possible to identify sufficient detail
Target population and definitions: Are they Syrian migrants from specified time period?	Yes	May be a dataset including Syrian migrants	May be dataset of migrants but incorrect time period or nationality
Transparency: Is there a clearly stated purpose, design and methodology?	Yes, thorough	Yes, partial	No
Completeness: [1] Is there evidence of rigorous processes to capture and report the entire population?	Yes: stated aim and explicit strategies to achieve this	May not be sufficiently addressed but without evidence of gaps	No: evidence of gaps in dataset
Sample design: [2] Is there an appropriate sampling strategy and attempt to achieve sufficient sample size and response rate?	Yes, thoroughly described	Yes, partial	No or unclear

[1] Criterion specific to population registers

[2] Criterion specific to survey data and qualitative sources

Table 1: Proposed framework for formal assessment of the data sources for modelling the recent Syrian asylum migration to Europe