Chapter 4
Building a Knowledge Base for the Model

Sarah Nurse and Jakub Bijak

In this chapter, after summarising the key conceptual challenges related to the measurement of asylum migration, we briefly outline the history of recent migration flows from Syria to Europe. This case study is intended to guide the development of a model of migration route formation, used throughout this book as an illustration of the proposed model-based research process. Subsequently, for the case study, we offer an overview of the available data types, making a distinction between the sources related to the migration processes, as well as to the context within which migration occurs. We then propose a framework for assessing different aspects of data, based on a review of similar approaches suggested in the literature, and this framework is subsequently applied to a selection of available data sources. The chapter concludes with specific recommendations for using the different forms of data in formal modelling, including in the uncertainty assessment.

4.1 Key Conceptual Challenges of Measuring Asylum Migration and Its Drivers

Motivated by the high uncertainty and complexity of asylum-related migration, discussed in Chap. 2, we aim to illustrate the features of the model-based research process advocated in this book with a model of migration route formation. We have focused on the events that took place in Europe in 2015–16 during the so-called 'asylum crisis', linked mainly to the outcomes of the war in Syria. To remain true to the empirical roots of demography as a social science discipline, a computational model of asylum migration needs to be grounded in the observed social reality (Courgeau et al., 2016).

Given the nature of the challenge, the data requirements for complex migration models are necessarily multi-dimensional, and are not limited to migration processes themselves, additionally including a range of the underpinning features and

J. Bijak, *Towards Bayesian Model-Based Demography*, Methodos Series 17,
https://doi.org/10.1007/978-3-030-83039-7_4

drivers. At the same time, problems with data on asylum migration are manifold and well documented (see Chap. 2). The aim of the work presented in this chapter is to collate as much information as possible on the chosen case study for use in the modelling exercise, and to assess its quality and reliability in a formal way, allowing for an explicit description of data uncertainty. In this way it can be still possible to use all available relevant information while taking into account the relative quality when deciding on the level of importance with which the data should be treated, and the uncertainty that needs to be reflected in the model.

In this context, it was particularly important to choose a migration case study with a large enough number of migrants, and with a broad range of available information and sources of data on different aspects of the flows. This is especially pertinent in order to allow investigation of the different theoretical and methodological dimensions of the migration processes by formally modelling their properties and the underlying migrant behaviour. Consequently, knowledge about the different aspects of data collection and quality of information, and a methodology for reflecting this knowledge in the model, become very important elements of the modelling endeavour in their own right.

In this chapter, we present an assessment of data related to the recent asylum migration from Syria to Europe in 2011–19. As mentioned above, we chose the case study not only due to its humanitarian and policy importance, and the high impact this migration had both on Syria and on the European societies, but also taking into account data availability. This chapter is accompanied by Appendix B, which lists the key sources of data on Syrian migration and its 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 framework introduced in this chapter.

Even though one of the central themes of the computational modelling endeavours is to reflect the complexity of migration, the theoretical context of our understanding of population flows has traditionally been relatively basic. As mentioned in Chap. 2, within a vast majority of the 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) aimed to explain the migration process as a weighing up of factors or ‘drivers’ which influence decisions to migrate, while Zelinsky (1971) described different features of a ‘mobility transition’, which could be directly observed. Most of the traditional theories do not reflect the complexity of migration (Arango, 2000), and typically fail to link the macro- and micro-level features of the migration processes, which is a key gap that needs addressing through modelling.

More recently, there have been attempts to move the conceptual discussion forward and to bridge some of these gaps. 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 general, the disconnection between

the theoretical discussions and their operationalisation – largely limited to survey-based questions on migration intentions – is a standard fixture of much of the conceptual work on migration.

In the context of displacement or forced migration, including asylum-related flows, 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. 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 social networks or media. For that reason, the monitoring of the underlying drivers, such as the conflict intensity, becomes of paramount importance (Bohra-Mishra & Massey, 2011). On the other hand, the measurement of drivers comes with its own set of challenges and limitations, which also need to be formally acknowledged.

Another crucial concept to consider when modelling migration processes is how different elements of the conceptual framework interact, and what that implies for measurement. An example could be the measurement of the difficulty of different routes for migrants undertaking a journey. In this case, 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 individuals have access to the information and resources necessary to navigate all the 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 (see Box 3.3; for a general discussion related to migration, see Stillwell et al., 2016). Friction can include either formal barriers, such as national borders and visa restrictions, or informal barriers, such as geographic distance or physical terrain. These challenges require adopting a flexible and imaginative approach to using data, for example by building synthetic indicators based on several sources, or using model-based reconciliation of data (Willekens, 1994).

4.2 Case Study: Syrian Asylum Migration to Europe 2011–19

In this section, we look at recent Syrian migration to Europe (2011–19) 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. From a historical perspective, recent 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 at the time of writing, more than nine 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 estimated 6.7 million people fleeing Syria since 2011 and a further 6.1 million internally displaced by the end of 2019, according to the UNHCR (2021, see also Fig. 4.1). The humanitarian crisis caused by the Syrian conflict, which had its dramatic peak in 2015–16, has continued throughout the whole decade.

Initial scoping of the modelling 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 Fig. 4.1, 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 many cases, the situation of the Syrian asylum seekers and refugees is also very precarious. By the UNHCR's account, by the end of 2017, nearly 460,000 people still lived in camps, mostly in the region, in need of more 'durable solutions', such as safe repatriation or resettlement. (This number has started to decline, and nearly halved by mid-2019). A further five million were dispersed across the communities in the 'urban, peri-urban and rural areas' of the host countries (UNHCR, 2021). The demographic structure of the Syrian refugee population generates challenges in the destination countries with respect to education provision and labour market participation, with about 53% people of working age (18–59 years), 2% seniors over 60 years, and 45% children and young adults under 18 (UNHCR, 2021).

When it comes to asylum migration journeys to Europe, visible routes and corridors of Syrian migration emerged, in recent years concentrating on the Eastern Mediterranean sea crossing between Turkey and Greece, as well as the secondary land crossings in the Western Balkans, and the Central Mediterranean sea route between Libya and Italy (Frontex, 2018). By the end of 2017, Syrian asylum migrants were still the most numerous group – over 20,000 people – among those apprehended on the external borders of the EU (of whom nearly 14,000 were on the Eastern Mediterranean sea crossing route). However, these numbers were considerably down from the 2015 peak of nearly 600 thousand apprehensions in total, and nearly 500,000 in the Eastern Mediterranean (*idem*, pp. 44–46). These numbers can be supplemented by other sad statistics: the estimated numbers of fatalities, especially referring to people who have drowned while attempting to cross the Mediterranean. The IOM minimum estimates cite over 19,800 drownings in the

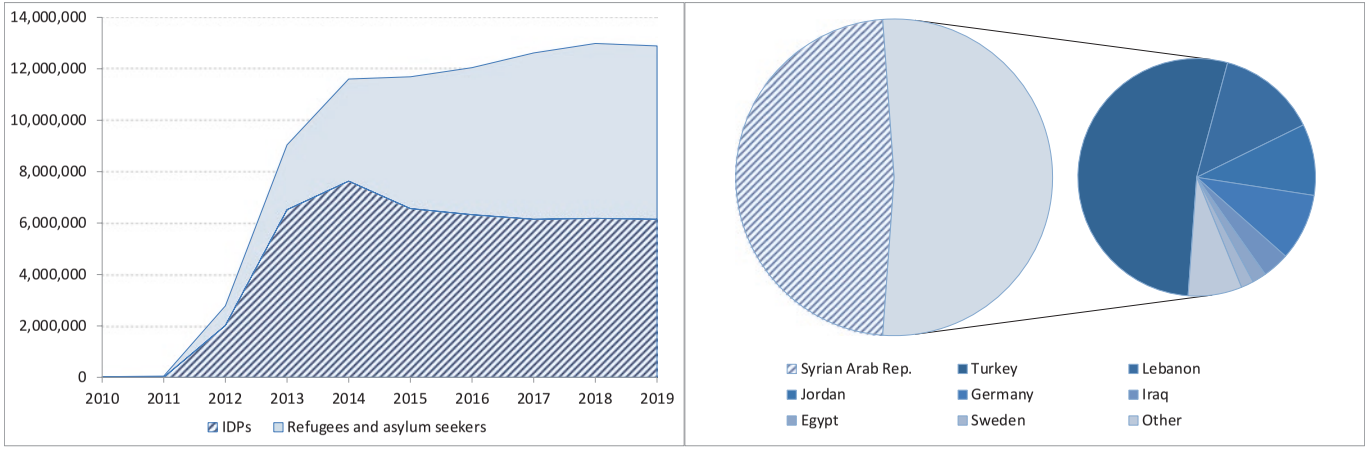

Fig. 4.1 Number of Syrian asylum seekers, refugees, and internally displaced persons (IDPs), 2011–19, and the distribution by country in 2019. (Source: UNHCR, 2021)

period 2014–19, of which 16,300 were in the Central Mediterranean. In about 850 cases, the victims were people who came from the Middle East, a majority presumed to be Syrian (IOM, 2021). In the same period, the relative risk of drowning increased to the current rate of around 1.6%, substantially higher (2.4%) for the Central Mediterranean route (*idem*).

As concerns the destinations themselves, the asylum policies and recognition rates (the proportion of asylum applicants who receive positive decisions granting them refugee status or other form of humanitarian protection) clearly differ across the destination countries, and also play a role in shaping the asylum data. Still, in the case of Syrian asylum seekers, these differences across the European Union are not large. According to the Eurostat data,[1] between 2011 and 2019, over 95% decisions to the applications of Syrian nationals were positive, and these rates were more or less stable across the EU, with the exception of Hungary (with only 36% positive decisions, and a relatively very low number of decisions made). It is worth noting here that administrative data on registrations and decisions have obvious limitations related to the timeliness of registration of new arrivals and processing of the applications, sometimes leading to backlogs, which may take months or even years to clear. Moreover, the EU statistics refer to asylum applications *lodged*, which refers to the final step in the multi-stage asylum application process, consisting of a formal acknowledgement by the relevant authorities that the application is under consideration (European Commission, 2016).

At the same time, besides the official statistics from the registration of Syrian refugees and asylum seekers by national and international authorities, specific operational needs and research objectives have led to the emergence of many other data sources. In this way, in addition to the key official statistics, such as those of the UNHCR, there exist many disparate information sets, which deal with some very specific aspects of Syrian migration flows and their drivers. These sources extend beyond the fact of registration, providing much deeper insights into some aspects of migration processes and their context. Still, the trade-offs of using such sources typically include their narrower coverage and lack of representativeness of the whole refugee and asylum seeker populations. Hence, there is a need for a unified methodology for assessing the different quality aspects of different data sources, which we propose and illustrate in the remainder of this chapter. In addition, we present a more complete survey of these sources in more detail in Appendix B, current as of May 2021, together with an assessment of their suitability for modelling.

[1] All statistics quoted in this paragraph come from the 'Asylum and managed migration' (migr) domain, table 'First instance decisions on applications by citizenship, age and sex' (migr_asydcfsta), extracted on 1 February 2021.

4.3 Data Overview: Process and Context

4.3.1 Key Dimensions of Migration Data

In the proposed approach to data collection and use in modelling, we suggest following a two-stage process of data assessment for modelling. The first stage 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 it can be formally taken into account and incorporated 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 processes, including the characteristics of migrants themselves, their journey and decisions on the one 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 or destination.

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.

The conceptual mapping of the different stages of the migration process and their respective contexts onto a selection of key data sources is presented in Fig. 4.2, with context influencing the different stages of the process, and the process itself being simplified into the origin, journey and destination stages. For each of these stages, several types of sources of information may be typically available, although certain types (surveys, interviews, 'new data' such as information on mobile phone locations or communication exchange, social media networks, or similar) are likely to be more associated with some aspects than with others. From this perspective, it is also worth noting that while the process-related information can be available both at the macro level (populations, flows, events), or at the micro level (individual migrants), the contextual data typically refer to the macro scale.

Hence, to follow the template for the model-building process sketched in Chap. 2, the first step in assessing the availability of data for any migration-related modelling endeavour is to identify the critical aspects of the model, without which the processes could not be properly described, and which can be usefully covered by the existing data sources, with a varying degree of accuracy. Next, we present examples of such process- and context-related aspects.

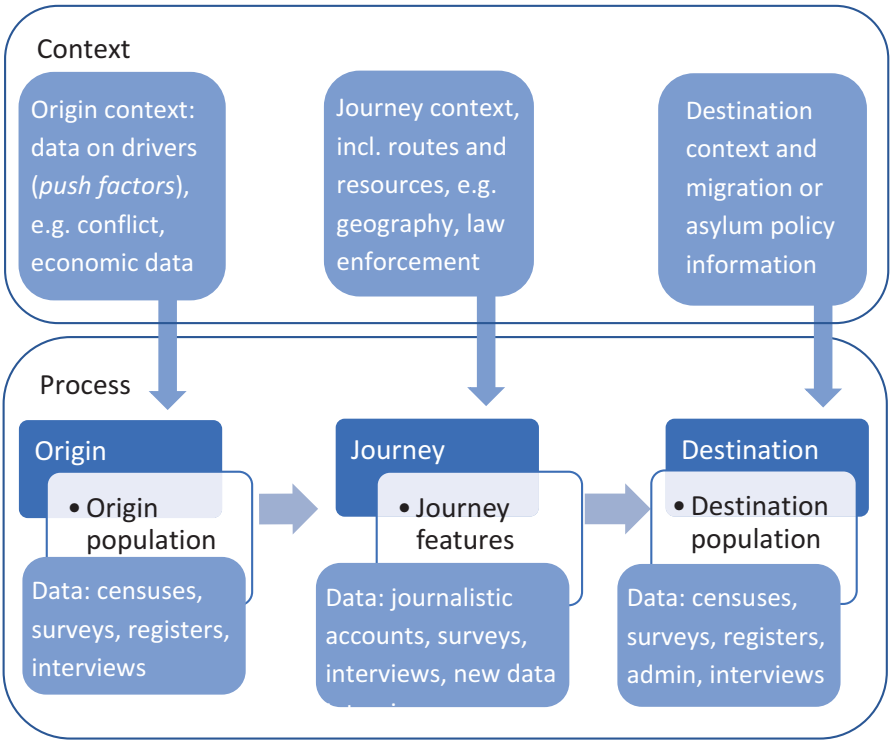

Fig. 4.2 Conceptual relationships between the process and context of migrant journeys and the corresponding data sources. (Source: own elaboration)

4.3.2 Process-Related Data

Among the process-related data, describing the various features of migration flows and migrants, be it for individual actors involved in migration (micro level) or for the whole populations (macro level), the main types of the information can be particularly useful for modelling are listed below.

Origin Populations. Information on the origin country 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.

Destination Populations. A wide range of data on migrant characteristics, economic situation (employment, benefits), access to and use of information, intentions, health and wellbeing 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, and can include a range of sources, from censuses and surveys, through administrative records, to qualitative interviews.

Registrations. Administrative and operational information from destination countries and international or humanitarian organisations, which register the arrival of migrants, can provide particularly timely 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.

Journey. Any information available about the specific features of the journey itself also forms part of the process-related information. This could include data about durations of the different segments of the trip, or distinct features of the process of moving, which can be gauged for example from retrospective accounts or surveys, including qualitative interviews or journalistic accounts. Similarly, information on intermediaries, smugglers, and so on, as long as it is available and even remotely reliable, can be a part of the picture of the migrant journeys.

Information Flows. Availability of information on routes and contextual elements can also impact on migrants' decisions during the migration process. Even though the information itself can be contextual, its availability and trustworthiness are related to the migration process. Insights into the information availability (and its flipside: the uncertainty faced by migrants before, during and after their journeys) can be obtained from surveys, but there is an underutilised potential to use alternative sources ('new data'). The use of such data for analysis requires having appropriate legal and ethical safeguards and protocols in place, in order to ensure that the privacy of the subjects of data collection is stringently protected.

4.3.3 Contextual Data

Formal modelling offers a possibility of incorporating 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, policies, and routes.

Origin Context. Information on the situation in the countries and regions of origin can include such factors as conflict intensity, the presence of specific events or incidents, as well as reports from observers and media, and identify the key drivers related to the decision to migrate (corresponding to push factors in Lee's 1966 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, provides contextual information on the relative attractiveness of various destinations (corresponding to pull factors).

Policies and Institutions. Specifically related to the destination context, but also extending beyond it, 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).

Route Features. Contextual data on, for example, geographic terrain, networks, borders, barriers, transport routes and law enforcement 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 in Lee's framework). Here, 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. This aspect can form a part of the set of route features mentioned above, or feature as a separate category, depending on the importance of the resource aspect for the analysis and modelling.

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.

Different process and context data are characterised by varying degrees of uncertainty, stemming from different features of the data collection processes, varying sample sizes, as well as a range of other quality characteristics. The quality of data itself is a multidimensional concept, which requires adequate formal analysis through a lens of a common assessment framework adopted for a range of different data sources that are to be used in the modelling exercise. We discuss methodological and practical considerations related to the design of such an assessment framework next, illustrated by an application to the case of recent Syrian migration to Europe.

4.4 Quality Assessment Framework for Migration Data

No perfect data exist, let alone concerning 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). As mentioned in Chap. 2, 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 people, for whom multiple events have been recorded, as well as the presence of undercount due to the clandestine nature of many asylum-related flows (Vogel & 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).

More generally, official migration statistics, as with all types of data, are social and political constructs, which strongly reflect the policy and research priorities prevalent at the time (for an example, see Bijak & Koryś, 2009). For this reason, the purpose and mechanisms of data collection also need to be taken into account in the assessment, as different types of information may carry various inherent biases. Given the potential dangers of relying on any single data source, which may be biased, when describing migration flows through modelling, multiple sources ideally need to be used concurrently, and be subject to formal quality assessment, as set out below.

4.4.1 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. Official legal texts, especially more recent ones, include references to data quality – European Regulation 862/2007 on migration and asylum statistics refers to and includes provisions for quality control and for assessing the "quality, comparability and completeness" of data (Art. 9).[2] Similarly, Regulation 763/2008 on population and housing censuses explicitly lists several quality criteria to be applied to the assessment of census data: relevance, accuracy, timeliness, accessibility, clarity, comparability, and coherence (Art. 6).[3]

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, compiled by Isernia et al. (2018), provides a broad overview, as well as listing some specific elements to be considered in the data analysis. Surveys selected for this review highlight definitional issues with identifying the appropriate target population. Aspiring to clarity in definitional issues is an enduring theme in migration studies, asylum migration included (Bijak et al., 2017).

There are also several examples of existing academic studies in related areas, which aim at assessing the quality of sources of information. Specifically in the

[2] Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July 2007 on Community statistics on migration and international protection, OJ L 199, 31.7.2007, p. 23–29, with subsequent amendments.

[3] Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses, OJ L 218, 13.8.2008, p. 14–20.

context of irregular migration, Vogel and Kovacheva (2008) proposed a four-point assessment scale 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 was 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. In addition, similar assessments were carried out in the broader demographic studies of the consequences of armed conflict (GAO, 2006; Tabeau, 2009; Bijak & Lubman, 2016), including additional suggestions for how to address the various challenges of measurement.

4.4.2 Proposed Dimensions of Data Assessment: Example of Syrian Asylum Migration

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 a quality assessment framework and uncertainty measures for different types of data is therefore multidimensional, as set out below. In particular, we propose six generic criteria for data 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 study, Syrian asylum migrants)
6. Transparency of the data collection methods

The need to identify the target population precisely is common for all types of data on migrants, but there are additional quality criteria specific to registers and survey-based sources. Thus, for register-based information an additional criterion relates to its completeness, while for surveys, their design, sampling strategy, sample sizes, and response rates are all aspects that need to be clearly set out in order to be assessed for rigour and good practice in 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 4.1. Finally, for each source, a summary rating is obtained by averaging over the existing classes. This meta-information on data quality can be subsequently used in modelling either by adjusting the raw data, for example when these are known to be biased, or by reflecting the data uncertainty, when there are reasons to believe that they are broadly correct, yet imprecise.

Table 4.1 Proposed framework for formal assessment of the data sources for modelling the recent Syrian asylum migration to Europe

Criteria	Green	Amber	Red
Purpose: Is the purpose for data collection relevant to and appropriate for the aim of modelling?	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 and/or publication delays	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

The result of applying the seven quality criteria to 28 data sources identified as potentially relevant to modelling Syrian migration is summarised in Table 4.2 and presented in detail in Appendix B. The listing in the Appendix additionally

Table 4.2 Summary information on selected data sources related to Syrian migration into Europe

Focus and type	Process data		Context data
	Destination population	Routes and journey	
Macro-level sources			
- Quantitative	Mainly registrations, operational data and large survey data Green/Amber (**10**)	Data from surveys and registrations, as well as operational data Amber (**7**)	Official statistics of the receiving (Green) and sending (Amber/Red) countries (**2**)
- Qualitative			Policy, legal and other secondary information Green/Amber (**1**)
Micro-level sources			
- Quantitative	Large-scale and random surveys Green/Amber (**3**)	Targeted surveys Amber (**1**)	
- Qualitative	Surveys and in-depth interviews. Amber (**1**)	Surveys and in-depth interviews. Amber (**3**)	

Note: Figures in brackets (**0**) indicate the number of sources reviewed in each category. Their details are listed in Appendix B

includes 20 supplementary, general-level sources of information on migration processes, drivers or features, some aspects of which may also be useful for 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.

On the whole, a majority of the data sources on Syrian asylum migration can be potentially useful in the modelling, at least to some degree. Most of the available data rely on registrations, operational data and surveys, and can be directly used to construct, parameterise or benchmark computational models of migration. The key proviso here is to know the limitations of the data and to be able to reflect them formally in the models. Caution needs to be taken when using some specific data sources, such as information from sending countries (in this case, Syria), due to a potential accumulation of several problems with their accuracy and trustworthiness, as detailed in Appendix B, but even for these, some high-level information can prove useful. Some suggestions as to the possible ways in which various data can be included in the models follow.

4.5 The Uses of Data in Simulation 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 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, for example around general concepts such as 'friction' or 'resources', and would allow the modeller to go far beyond standard geographic measures of distance or economic measures of capital, respectively.

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.

Third, the context of data collection has to be always borne in mind. Migration statistics – being to a large extent social and political constructs – are especially prone to becoming 'statistical artefacts' (see e.g. Bijak & Koryś, 2009), being distorted, and sometimes misinterpreted. With that in mind, 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, 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.

A natural formal language for describing the data quality or, in other words, the different dimensions of the uncertainty of the data sources, is provided by probability distributions, which can be easily included in a fully probabilistic (Bayesian) model for analysis. In the probabilistic description, two key aspects of data quality come to the fore: *bias* – by how much the source is over- or under-estimating the real process – which can be modelled by using the location parameters of the relevant distributions (such as mean, median and so on), and *variance* – how accurate the source is – which can be described by scale parameters (such as variance, standard deviation, precision, etc.). As in the statistical analysis of prediction errors, there may be important trade-offs between these two aspects: for example, with sample surveys, increasing the sample size is bound to decrease the variance, but if the sampling frame is mis-specified, this can come at the expense of an increasing bias – the estimates will be more precise, but in the wrong place.

Of the eight quality assessment criteria listed in Table 4.1, the first two (purpose and timeliness) are of a general nature, and – depending on the aim of the modelling endeavours – can be decisive in terms of whether or not a given source can be used at all. The remaining ones can be broadly seen either as contributing to the bias of a source (definitions of the target populations, trustworthiness of data collection, and

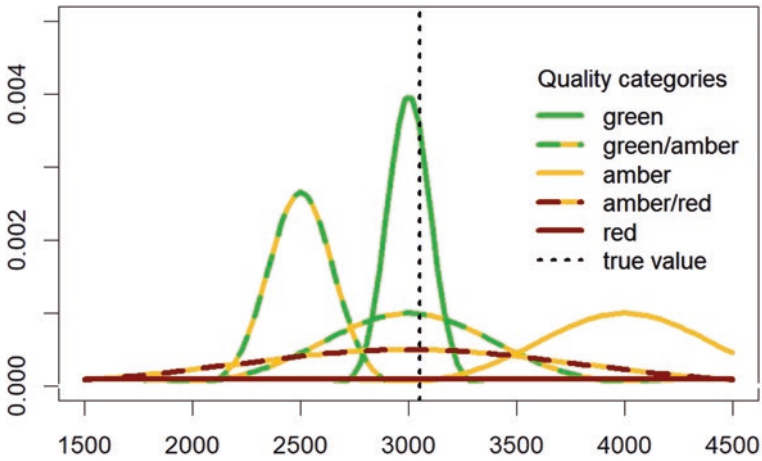

Fig. 4.3 Representing data quality aspects through probability distributions: stylised examples. (Source: own elaboration)

completeness of coverage), or to its variance (level of disaggregation, sample design, and transparency of data collection mechanisms). The interplay between these factors can offer important guidance as to what probabilistic form a given distribution needs to take, and with what parameters.

Figure 4.3 illustrates some stylised possibilities of how data falling into different quality classes can map onto the reality, depicted by the vertical black line. Hence, we would expect a source classified as 'green' to have minimal or negligible bias and relatively small variance. The 'green/amber' sources could either exhibit some bias, the extent of which can be at least approximately assessed, or maybe a somewhat larger variance – although both of these issues together would typically signify the 'amber' quality level and a need for additional care when handling the data. Needless to say, sources falling purely into the 'red' quality category should not be used in the analysis at all, while the data in the 'amber/red' category should only be used with utmost caution, given that they can point to general tendencies, but not much beyond that.

As discussed in Chap. 2, the data can enter into the modelling process at different stages. First, as summarised in Fig. 2.1, modelling starts with observation of the properties of the processes being modelled. What follows, in the inductive step of model construction, is the inclusion of information about the features and structures of the process, as well as the information on the contributing factors and drivers. Hence, at the steps following the principles of the classical inductive approach, all relevant context data need to be included, as well as micro-level data on the building blocks of the process itself. Subsequently, so that the model is validated against the reality, macro-level data on the process can be used for benchmarking. In other words, micro-level process data, as well as context data become model inputs, whereas macro-level process data are used to calibrate model outputs.

A natural way to include the uncertainty assessment of the different types of data sources is then, for the inputs, to feed the data into the model in a probabilistic form (as probability distributions), and, for the outputs, to include in the model an additional error term that is intended to capture the difference between the processes being modelled and their empirical measurements (see Chap. 5). Box 4.1 presents an illustration related to a set of possible data sources, which may serve to augment the Routes and Rumours model introduced in Chap. 3 and to develop it further, together with their key characteristics and overall assessment. More details for these sources are offered in Appendix B.

Box 4.1: Datasets Potentially Useful for Augmenting the Routes and Rumours Model

As described in Chap. 3, temporal detail and spatial information are important for this model in order to understand more about the emergence of migration routes. We focused on the Central Mediterranean route, utilising data on those intercepted leaving Libya or Tunisia, losing their lives during the sea crossing, or being registered upon arrival in Italy. One exception was the retrospective Flight 2.0 survey, carried out in Germany, which looked into the use of information by migrants during their journey. All the data included below are quantitative, reported at the macro-level (although Flight 2.0 recorded micro-level survey data), and relate to the migration process. The available data are listed in Table 4.3 below; for this model monthly totals were used. In addition, OpenStreetMap (see source S02 in Appendix B) data provides real world geographic detail. For a general quality assessment of data sources, see Appendix B, where the more detailed notes for each dataset provide additional relevant information and give some brief explanation of the reasoning behind particular quality ratings.

Table 4.3 Selection of data sources which can inform the Routes and Rumours model, with their key features and quality assessment

Reference in Appendix B	Source	Content focus	Source and time detail	Quality rating	Bias & variance
11	IOM Missing Migrants: Flows	Destination population: Interceptions by Libyan /Tunisian coastguards	Operational & admin, monthly data	**Amber**	Medium undercount & variance
12	IOM Missing Migrants: Deaths	Number of recorded deaths during Central Med crossings	Operational & journalistic, daily data	**Amber**	Medium undercount & variance
13	IOM Displacement Tracker	Destination population: Daily arrivals registered in Italy	Operational, daily data	**Green/ amber**	Small undercount & variance
24	Flight 2.0 / Flucht 2.0	Data on information use and levels of trust *en route* to Germany	One-off survey	**Amber**	Unknown bias, large variance

Source: see Appendix B for details related to individual sources

Of course, there are also other methods for dealing with missing, incomplete or fragmented data, coming from statistics, machine learning and other emerging areas of broader 'data science'. The review of such methods remains beyond the scope of this book, but it suffices to name a few, such as various approaches to imputation, which have been covered extensively e.g. in Kim and Shao (2014), or data matching, which in machine learning is also referred to as data fusion, also covered by a broad literature (e.g. Bishop et al., 1975/2007; D'Orazio et al., 2006; Herzog et al., 2007). A comprehensive recent review of the field was provided by Little and Rubin (2020). In the migration context, some of these methods, such as micro-level matching, are not very feasible, unless individual-level microdata are available with enough personal detail to enable the matching. For ethical reasons, this should not be possible outside of very secure environments under strictly controlled conditions; therefore this may not be the right option for most applied migration research questions. Better, and more realistic options include reconciliation of macro-level data through statistical modelling, such as in the Integrated Modelling of European Migration work (Raymer et al., 2013), producing estimates of migration flows within Europe with a description of uncertainty. Such estimates can then be subject to a quality assessment as well, and be included in the models following the general principles outlined above.

4.6 Towards Better Migration Data: A General Reflection[4]

As discussed before, the various types of contemporary migration data, as well as other associated information on the related factors and drivers, are still far from achieving their potential. The data are typically available only after a time delay, which poses problems for applications requiring timeliness, such as rapid response in the case of asylum migration. Data on migrants, as opposed to counts of migration events, are still relatively scarce, and particularly lacking are longitudinal studies involving migrant populations. The existing data are not harmonised, nor are they exactly 'interoperable' – ready to be used for different purposes or aims, with tensions between particular policy objectives and the information the data can provide.

No matter what practical solutions are adopted for the use of migration data in modelling, several important caveats need to be made when it comes to the interpretation of the meaning of the data. As argued above, the data themselves are

[4] Part of the discussion is inspired by a debate panel on migration modelling, held at the workshop on the uncertainty and complexity of migration, in London on 20–21 November 2018. The discussion, conducted under the Chatham House rule (no individual attribution), covered two main topics: migration knowledge gaps and ways to fill them, and making simulation models useful for policy. We are grateful to (in alphabetical order) Ann Blake, Nico Keilman, Giampaolo Lanzieri, Petra Nahmias, Ann Singleton, Teddy Wilkin and Dominik Zenner for sharing their views.

social constructs and the product of their times, and as such, are not politically neutral. These features put the onus on the modellers and users, who need to be aware of the social and political baggage associated with the data. Besides the need to be conscious of the context of the data collection, there can be a trap associated with bringing in too much of the analysts' and modellers' own life experience to modelling. This, in turn, requires particular attention in the context of modelling of migration processes that are global in nature, or consider different cultural contexts than the modellers' own.

Similar reservations hold from the modelling point of view, especially when dealing with agent-based models attempting to represent human behaviour. Such models often imply making very strong value judgements and assumptions, for example with respect to the objective functions of individual agents, or the constraints under which they operate. The values that are reflected in the models need to be made explicit, also to acknowledge the role of the research stakeholders, for the sake of transparency and to ensure public trust in the data. It has to be clear who defines the research problem underlying the modelling, and what their motivations were.

Another aspect of trust relates to the new forms of data, such as digital traces from social media or mobile phones, where their analytical potential needs to be counterbalanced by strong ethical precautions related to ensuring privacy. This is especially crucial in the context of individual-level data linking, where many different sources of data taken together can reveal more about individuals than is justified by the research needs, or than should be ethically admissible. This also constitutes a very important challenge for traditional data providers and custodians, such as national and international statistical offices and other parts of the system of official statistics, whose future mission can include acting as legal, ethical and methodological safeguards of the highest professional standards with respect to migration data collection, processing, storage and dissemination.

Another important point is that the modelling process, especially if employed in an iterative manner, as argued in Chap. 2 and throughout this book, can act as an important pathway towards discovering further gaps in the existing knowledge and data. This is a more readily attainable aim than a precise description or explanation of migration processes, not to mention their prediction. Additionally, this is the place for a continuous dialogue between the modellers and stakeholders, as long as the underpinning ideas and concepts are well defined, simple, clear and transparent, and the expectations as to what the data and models can and cannot deliver are realistic.

To achieve these aims, open communication about the strengths and limitations of data and models is crucial, which is one of the key arguments behind an explicit treatment of different aspects of data quality, as discussed above. These features can help both the data producers and users better navigate the different guises of the uncertainty and complexity of migration processes, by setting the minimum quality standards – or even requirements – that should be expected from the data and

models alike. A prerequisite for that is a high level of statistical and scientific literacy, not only of the users and producers of data and models, but also ideally among the general public. To that end, while the focus of this chapter is on the limitations of various sources of data, and what aspects of information they are able to provide, the next one looks specifically at the ways in which the formal model analysis can help shed light on information gaps in the model, and also utilise empirical information at different stages of the modelling process.