Understanding social data: An example of migration

Dr Sarah Nurse and Prof. Jakub Bijak, University of Southampton

Data on social processes are generally fraught with problems, just because they aim to describe our very messy social reality. In this video, we will use an example of international migration to discuss some of the more general issues with social data. We will focus on the limitations of measurement, but also present some ways in which we can overcome, or at least acknowledge these limitations. Finally, we will give practical advice on how the imperfect data can be used in modelling.

So, why migration? This is one of the trickiest demographic processes, which involves a lot of human agency in making decisions to move or not, where to, and when. Migration is linked to many drivers or factors, both at the macro level – so sending and receiving countries – as well as at the micro level – individual migrants and other actors. All these drivers typically interact: a decision to migrate is linked to the opportunities at the destination that are not available at origin, and these opportunities are viewed through the lens of personal preferences. Besides, there are also other important elements intervening in the migration process, such as transit routes, or information exchange between the agents.

These many different aspects of migration, or many other social processes, are very difficult to describe and measure by using the data that we have on our disposal – if this is at all possible. Over the next few minutes, my colleague, Dr Sarah Nurse, will introduce you to specific challenges related to different aspect of measuring migration – or to be precise, one of the most uncertain migration processes, which is forced migration and asylum.

As Jakub has just described, dealing with data quality is an essential part of the modelling process. To do this we begin by identifying the types of data from a variety of sources which can be used in different parts of the model.

Broadly speaking, the data sources can be viewed as providing either process-related or contextual information, depending on how they may be used in the model. Some of the data directly relate to the migration process and tell us for example about the characteristics of individual migrants or their journeys and decisions. Other useful data relate to the broader context within which the migration occurs both at the origin and destination as well as transit countries. For example the macro-economic indicators of destination countries are relevant for the decisions made by migrants as they weigh up their possible options.

It is also useful to classify the data by the level of aggregation. Noting whether the data is collected and available at the micro or macro level allows us to easily identify how and where we can appropriately use it in the model. An additional label of whether the data is quantitative or qualitative is useful for comparing similar sources during the quality assessment, as well as categorisation by type (for example is it a register, a survey or census data). These are the key components of meta-information related to individual data sources that we need for the quality assessment.

To take a step back for a moment, before we delve deeper into how we can assess data quality and look at some examples it is helpful to first recognise the challenges we face, both conceptually and in the measurement of this complex social process.

While it is true that one of the central themes of computational modelling is to try and reflect the complexity of migration, traditional theories which attempt to explain population flows are generally relatively basic, often limited to describing the effect of structural differences, such as employment rates, on the resulting overall flows. They also fail to link the macro- and micro-level features of the migration processes. More recently there have been attempts made to address these gaps, but in general, the disconnect between the theoretical discussions and their potential operationalisation with the empirical reality remains.

In terms of the challenges with measurement: consider for example how we measure the relative difficulty of different potential routes between a point of origin and destination; would you measure it by physical distance of a route? What if you are on foot and there is a mountain range or a desert between you and your destination, not to mention an expanse of water? On top of these informal barriers, the difficulty will inevitably increase if there are formal barriers such as a national border to cross and visa restrictions, but mitigated where there is a smuggling network already operating and if an individual has access to the information and resources necessary to navigate these. An overall summary measure of this can be thought of as a route's 'friction', and migrants' perception of this will feed into their decision making about which destinations and routes to take.

So how can we lay the groundwork for addressing some of these challenges and using the data in the modelling process? We do this by carrying out a systematic process of quality assessment. We have identified six aspects that need to be considered for each data source, with each one being graded as green, amber or red.

Firstly… Is the purpose for collecting the data relevant to and appropriate for our use? A source could be labelled green if this is the case, and the data is collected in order to estimate or understand our particular migration flow of interest, amber if it is collected for a different purpose but is still relevant, or red if the data has a different purpose which impacts on its usefulness for our analysis.

Secondly, are the data published at sufficiently frequent intervals? Also, is the source free from obvious biases or stated political aims? Fourthly, is there a sufficient level of detail in the geographic and country of origin information being reported? Next, does the data include our target population of migrants from the specified time period? Finally, are the data collection methods transparent with clearly stated aims and description of their design and implementation?

In addition, for population registers there needs to be an assessment of how complete the source is and for surveys, whether there is evidence of appropriate sampling strategy, sample sizes and response rates. These are all aspects which need to be clearly set out in order to be assessed for rigour and good practice in the data collection.

In order to fully explore the multi-dimensional nature of the migration process, its features and drivers, it is important to choose a migration case study with a large enough flow of migrants, and with a broad range of available information and sources of data on different aspects of that flow. Therefore, we have chosen to present an assessment of the data related to the recent asylum migration from Syria to Europe between 2011 and 2017. In addition to the availability of data for this case study, it has also been chosen for its humanitarian and policy importance, and the high impact this migration has had both on Syria and on the European societies.

Here we can see a summary of the quality indicators for UNHCR registration data. At the top are labels describing the data type, topic and level of aggregation. The coloured shading corresponds to a traffic-light assessments of the individual quality criteria. Green indicates good quality, amber suggests it is acceptable and red signals problems. We can see this here for the timeliness of a one off survey of Syrian refugees living in Austria. Each of the quality dimensions, as well as a global quality score for a given data source, is classified into one of the three categories: green, amber and red, or two in-between classes, green-amber and amber-red, reflecting the various aspects of variation and bias inherent in the source. This process is repeated for each source, creating an inventory of all available data relevant to the case study of interest, alongside an assessment of their quality issues.

So, what are the ways forward? As data-free models have obvious limitations, we need to use at least some empirical data in some form, if the models we build are to bear any resemblance to the reality they are trying to describe. The challenge becomes, what to do with all the known and unknown problems with different aspects of data quality. Based on the discussion so far, we can make three observations, which can help us use the data in modelling more consciously.

First, there are no perfect data, so we need to have realistic expectations on how any data can be used in modelling. Looking across the range of the assessment criteria mentioned by Sarah, different sources have different strenghts and weaknesses. For this reason, to understand the data better, we need to carry out a comprehensive data assessment looking at different criteria at the same time. What we intend to model will have a bearing on the criteria that we will treat as more important: for example, if we are building a model to provide early warnings about changes in migration flows, timeliness will be a crucial consideration, while the original purpose of data collection may be of secondary concern.

Second, we need to acknowledge any source of uncertainty, ambiguity or other imperfections in the data. Ideally, we would like to do it formally, through a description in probabilistic terms, as probability distributions can then get propagated into the simulation model. The language of probabilities is a natural choice for expressing different types of uncertainty.

Third, the context of data collection has to be borne in mind. Statistics on migration or many other social processes are to a large extent social and political constructs: they are collected with specific purposes in mind. This needs to be acknowledged, and the use of particular data needs to be ideally driven by particular research questions rather than convenience. There are a few additional concerns here: for example, new forms of data, such as digital traces from social media or mobile phones, need to come with strong ethical protections related to data privacy.

How then to describe data quality in the language of probabilities? Of the eight criteria mentioned by Sarah, some are general, such as purpose or timeliness. They can help decide, whether a given source can be used at all in our model, or not. The remaining ones can be broadly seen either as contributing to the bias of a source, or to its variance. This graph shows some possibilities of how data falling into different quality classes can map onto the reality, which depicted by the vertical black line.

Hence, we would expect a ‘green’ source to have minimal or negligible bias and relatively small variance. The ‘green-to-amber’ sources could either exhibit some bias, or maybe a somewhat larger variance. If both problems are present together, this would typically signify the ‘amber’ quality level and a need for additional care when handling the data. Sources falling purely into the ‘red’ quality category should not be used in the analysis at all. Data in the ‘amber-to-red’ class should only be used with utmost caution: they can point to general tendencies, but not much beyond that.

How to include the data and their quality assessments in agent-based models? Here, the distinctions between process and context, as well as micro and macro data can come handy. If our aim is to build a model of a social process, such as migration, then all context data can easily become model inputs. The same holds for micro-level process data, such as decision rules, which are used to generate the process behaviour. At the same time, the macro-level process data can be used to calibrate model outputs. If we repeat the modelling exercise several times, based on different sets of data, the modelling process can also help us illuminate the gaps in the existing data. In any case, modellers need to be transparent about the strengths and limitations of the data they use, which is why we suggested our framework for assessing different aspects of data quality in an open way.