BRIEF REPORT

REVISED Mapping secondary data gaps for social simulation modelling: A case study of the journeys of Syrian asylum seekers to Europe [version 2; peer review: 3 approved, 1 approved with reservations]

Sarah Nurse[1], Martin Hinsch[2], Jakub Bijak[1]

[1]University of Southampton, Southampton, SO17 1BJ, UK
[2]University of Glasgow, Glasgow, G12 8RZ, UK

v2 **First published:** 08 Dec 2023, **3**:216
https://doi.org/10.12688/openreseurope.15583.1
Latest published: 07 Oct 2024, **3**:216
https://doi.org/10.12688/openreseurope.15583.2

Abstract

Simulation models of social processes may require data that are not readily available, have low accuracy, are incomplete or biased. The paper presents a formal process for collating, assessing, selecting, and using secondary data as part of creating, validating, and documenting an agent-based simulation model of a complex social process, in this case, asylum seekers' journeys to Europe. The process starts by creating an inventory of data sources, and the associated metadata, followed by assessing different aspects of data quality according to pre-defined criteria. As a result, based on the typology of available data, we are able to produce a thematic map of the area under study, and assess the uncertainty of key data sources, at least qualitatively. We illustrate the process by looking at the data on Syrian migration to Europe in 2011–21.

In parallel, successive stages of the development of a simulation model allow for identifying key types of information which are needed as input into empirically grounded modelling analysis. Juxtaposing the available evidence and model requirements allows for identifying knowledge gaps that need filling, preferably by collecting additional primary data, or, failing that, by carrying out a sensitivity analysis for the assumptions made. By doing so, we offer a way of formalising the data collection process in the context of model-building endeavours, while allowing the modelling to be predominantly question-driven rather than purely data-driven. The paper concludes with recommendations with respect to data and evidence, both for

Open Peer Review

Approval Status ✓ ? ✓ ✓

	1	2	3	4
version 2 (revision) 07 Oct 2024	✓ view			
version 1 08 Dec 2023	? view	? view	✓ view	✓ view

1. **Alejandra Rodriguez-Sánchez**, University of Potsdam, Potsdam, Germany
2. **Edgar Scrase**, United Nations High Commissioner for Refugees (UNHCR), Geneva, Switzerland
   Zaruhi Mkrtchyan, UNHCR, Copenhagen, Denmark
3. **Alireza Jahani**, Brunel University London, London, UK
4. **Denis Kierans**, University of Oxford, Oxford, UK

Any reports and responses or comments on the article can be found at the end of the article.

modellers, as well as model users in practice-oriented applications.

Keywords
Agent-based modelling, Asylum migration, Data quality, Empirical evidence, Knowledge gaps, Modelling process, Secondary data, Syrian migration

This article is included in the European Research Council (ERC) gateway.

This article is included in the Horizon 2020 gateway.

This article is included in the Migration collection.

Corresponding author: Jakub Bijak (j.bijak@soton.ac.uk)

Author roles: Nurse S: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Hinsch M**: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **Bijak J**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 725232). This article reflects the authors' views, and the Research Executive Agency of the European Commission is not responsible for any use that may be made of the information it contains.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Nurse S, Hinsch M and Bijak J. **Mapping secondary data gaps for social simulation modelling: A case study of the journeys of Syrian asylum seekers to Europe [version 2; peer review: 3 approved, 1 approved with reservations]** Open Research Europe 2024, **3**:216 https://doi.org/10.12688/openreseurope.15583.2

First published: 08 Dec 2023, **3**:216 https://doi.org/10.12688/openreseurope.15583.1

REVISED **Amendments from Version 1**

Version 2 of the paper has been amended to directly address some of the very helpful reviewers' comments to Version 1, for which we are very grateful. The key changes included: explaining the data assessment process better, including some examples, explaining the way in which the selected data sources ended up being used in the modelling, expanding the review of existing practices and guidelines for quality assurance in official statistics (including on humanitarian matters), and sharpening the discussion along the lines suggested by the reviewers. The details of the revisions made are listed in our responses to the individual reviews.

Any further responses from the reviewers can be found at the end of the article

Introduction

The aim of this brief report is to propose and reflect on an approach for collating, assessing, and selecting appropriate secondary data for use in an agent-based simulation model of a complex social process. The discussion is illustrated by a case study related to creating, validating, and documenting a model of migration route formation. Our intended contribution is to propose a template for critical assessment of secondary social data and for identifying key knowledge gaps, while remaining honest about the overall uncertainty of social simulation models, and the particular role of data uncertainty in it.

The work is motivated by the need to provide more realistic and reliable input to meet the ever-increasing policy demand for a better understanding of the patterns and drivers of migration. Our particular focus is on the rapidly-evolving asylum flows, where the relevant policies can be aimed at improving the preparedness of the transit and destination countries[1]. The discussion is illustrated with an example of modelling Syrian asylum asylum seekers travelling to Europe in the 2010s, with the modelling process itself presented in more detail in Bijak *et al.* (2021). At the same time, in social simulation studies, data quality aspects rarely receive the attention they deserve.

This paper is structured as follows. First, in the Methods section, we discuss the interplay and tensions between the supply of available secondary data and the demand for such information to meet the specific needs and research objectives of social simulation models. We focus here particularly on the ways to identify the key information gaps. Subsequently, in the Results section, we present an example of data for a simulation model of migration route formation, based on the journeys made by Syrian asylum seekers travelling to Europe in the 2010s. The final section concludes with a discussion of key remaining challenges, possible solutions, and practical recommendations for model builders and users.

[1] See for example the EU migration 'Blueprint': Commission Recommendation (EU) 2020/1366 of 23 September 2020 on an EU mechanism for preparedness and management of crises related to migration, OJ L 317, 1.10.2020, p. 26–38, http://data.europa.eu/eli/reco/2020/1366/o

Methods: Assessing data supply and demand

In empirically-grounded social modelling, a systematic review of the knowledge base typically begins with an assessment of the available secondary information on the topic in question. The foundations for that can be laid by formally creating a **comprehensive inventory of data supply**, as complete as possible, which aims to assemble key information and meta-information about the different types of data available, the key themes they refer to, and sources they come from. Such an inventory should ideally include meta-information about the classification of data according to a range of characteristics, such as whether the data source is qualitative or quantitative, is it a survey, census, register, observation, or interview, whether the level of aggregation is micro (individuals) or macro (groups), and so on. This meta-information can be either taken from the sources of the data themselves or imputed by analysts.

In the example concerning Syrian asylum seekers' journeys to Europe in 2011–21, we have considered all publicly available datasets that could be freely accessible online and could potentially provide useful input into the simulation modelling of cross-Mediterranean migration routes. In particular, for data on migration flows and the context in which they occur, we have included a range of databases, online registers, provided by international and national organisations, such as UNHCR, Eurostat, Frontex, or national statistical offices, as well as those resulting from individual research studies, purposefully searched via Google Scholar. The meta-data collection took place in the summer of 2019, and was updated in the autumn of 2021, initially yielding 28 directly-relevant sources (after update increasing to 32), as well as 20 auxiliary ones, listed in Bijak *et al.* (2021).

The analysis of the limitations of the datasets, including of any biases, or other sources of uncertainty and errors, formed an inherent part of the next steps in evaluating the supply of available data. These steps consisted in establishing **a set of quality assessment criteria** suitable for the problem at hand. As data quality is a multidimensional construct, with different aspects having varying importance for different users and application areas, the assessment needs to be multidimensional, too. A simple and intuitive quality assessment scheme can use a variation of a traffic-lights approach (green for good data, amber for data with some problems, and red for poor data) to assess each quality dimension according to professional judgement (for related examples, see GAO, 2006, or Vogel and Kovacheva, 2008, and for a recent overview of issues with migration statistics generally, see Kraler and Reichel, 2022).[1]

More broadly, international guidelines for data quality assessment in the context of the production of official statistics can be found e.g. in the documents of the United Nations Statistical Division (2018) and Eurostat (2019). For data quality, the latter guidance proposes to use five quality criteria (relevance; accuracy and reliability; timeliness and punctuality; coherence and comparability; and accessibility), while the former treats timeliness and punctuality separately, and additionally includes interpretability. However, these are generic

data quality criteria, not specifically focused on modelling applications. The United Nations Statistical Division (2018) also explicitly notes that there are important trade-offs between different quality dimensions, the most notable of which is the tension between timeliness and accuracy – the more rapid the data production, the less space for thorough quality assurance.

In the example of data for modelling Syrian asylum journeys, we used a **five-point traffic-lights scale** inspired by the traffic-lights approach, with green-amber and amber-red as interim points. We use one threshold criterion, to what extent the data may be suitable for the problem at hand, followed by six specific criteria. Three of them concern the data generation process as such (its timeliness, the level of data disaggregation, as well as population coverage and adherence to definitions), two criteria refer to the level of trust in data (trustworthiness of the source and transparency of documentation), and one either to completeness (for register-based sources) or sample design (for surveys and other sample-based studies). Finally, we also provide a summary score averaging across the available quality dimensions – with individual scores treated numerically (from one for red to five for green), the summary score is an equally-weighted average across all the individual criteria, rounded to the nearest integer (Bijak *et al*. 2021).

An initial stages of assembling a data inventory should also include a comprehensive documentation of all sources and classification decisions, as well as rationale for data assessment. The **analysis of metadata and data quality dimensions** can already shed light on the real extent of data supply potentially usable for modelling purposes. A summary of themes covered by different types of available data, filtered by those with a positive assessment of the relevant quality aspects, determines the actual, rather than potential data supply for the problem at hand. For example, by limiting the analysis to the theme of migrant journeys and considering only the sources falling between the green and amber categories overall, with a green rating for transparency and trustworthiness, already considerably limits the number of sources that can be used in the subsequent analysis. This exercise also helps illuminate the data gaps that would ideally need to be covered to meet the needs of modelling, if it is to be empirically grounded.

The other side of the process is related to the assessment of the **demand for secondary data** to be used in a simulation model. There are two main aspects of this. First, an analysis of data needs, initially at the conceptual level, can provide a rough idea what information would be ideally required for the model to have full empirical basis. Clearly, it is not possible to expect data on all aspects of the model, but this exercise already gives the first approximation about the areas, in which empirical grounding would be ideally needed. It also helps reflect on how the data can be operationalised—qualitatively or quantitatively, and through which variables—and how they can be used in the modelling process—whether to calibrate model parameters or other inputs, or to externally validate model outputs.

Once the initial assessment and modelling has taken place, the second step in assessing the demand for data can consist of an iterative process for **identifying the data gaps** that can be filled. From the point of view of the modelling process, the ultimate aim is either filling these gaps with secondary data, if available, or through a dedicated collection of primary data on a specific topic. The analytical mechanism which allows this involves *sensitivity analysis*, especially in the statistical sense, which identifies the model inputs that contribute the most to the model results (Oakley & O'Hagan, 2004). Such inputs are primary targets for additional data collection. This process can be aided by formal modelling of the *provenance* (origins) of different data sources used in a model, and the relationships between the data and different modules and aspects of the model, identifying the data sources, on which many other elements of the modelled social reality are dependent (Reinhardt *et al.*, 2023).

The identification of data gaps followed by collection of additional information can **proceed in an iterative manner** to reduce the gap between the demand for data and the available supply, while expanding the latter through addition of new, dedicated datasets. In the end, however, there are always some models' aspects which cannot be calibrated to data, and some free parameters, for which empirical grounding is not possible. They are the key sources of residual uncertainty, which is an important indicator of model performance in its own right, and thus merits a separate analysis (Bijak *et al.*, 2021).

Results: Data on Syrian asylum seekers and refugees

In the presented example, we aim to study a migration process modelled on asylum flows from Syria to Europe during the 2010s. We use an agent-based approach, simulating individual migrants (agents), their journeys and decisions. The decisions are based on a range of factors known to the agents, but are made under uncertainty, resulting from incomplete information about the world, as well as knowledge and decisions of other agents (for details, see Hinsch & Bijak, 2023). The creation of the data inventory therefore had to encompass both individual-level data on Syrian asylum seekers and factors affecting their decisions, as well as macro-level information about the variables that could enter into the modelling process, either as input to decisions, or as a way of calibrating the model results. The inventory, covering 32 entries and including metadata for individual data sources, as well as their quality assessment, is available in Nurse and Bijak (2023) and in a searchable form from Nurse and Bijak (2021).

The data included in the inventory include both quantitative and qualitative sources that either pertain to the process of migrant journeys or the relevant context in which the journeys took place. The data could relate to individuals (micro-level), or be available at the aggregate, population levels (macro). **The inventory is purpose-specific** and covers only those sources that were publicly available at the time and that were deemed potentially relevant for modelling purposes. For this reason, the

inventory may not cover those sources of data on Syrian migration that were either unavailable for research purposes or were found to be unrelated to the modelling objectives. What has been included, typically clustered around four main themes: data on routes and journeys as such, including on migrant decisions; data on the migrant populations, typically in destination countries; data on availability of information during journeys; and key data on relevant policies. The thematic map of the case study data is presented in Figure 1, displaying a word cloud created from the descriptions of all sources included in the inventory. The map is intended as a broad qualitative summary of the content of the inventory. This high-level exploratory overview points out to several key themes emerging from the data description: good availability of the population-level (macro) quantitative data on migration processes, but also the presence of at least some surveys and other micro-level data focused on journeys, including qualitative sources (interviews).

A more detailed **summary of the inventory** is offered in Figure 2, which shows the frequencies (counts) of data sources included in the inventory. It is clear that the dominant groups of data sources covered migrant routes and journeys, as well as migrant populations in the destination countries. The data would typically come from either registers or surveys, routine or bespoke, closely followed by interviews with individual migrants. Most data included in the inventory are quantitative, related to migration process as such, and available at the aggregate level, although with a visible minority of qualitative, contextual and micro-level data as well.

The quality assessment for the sources included in the inventory is summarised in Figure 3. For all individual quality criteria, the expert assessment was carried out within the research team, and included an initial evaluation followed by independent moderation of the scores through a deliberative discussion and thought exchange. By definition of the inventory, all sources included had to be at least partially relevant for the modelling purposes. A clear majority of sources have also scored highly on the transparency and trustworthiness dimensions of the evaluation, as well as, although to a slightly smaller extent, on adherence to the established concepts and definitions. The quality assessment for source-specific criteria (sample design for surveys and completeness for registers) was mixed, with various sources located in different places along the five-point scale. The availability and quality of disaggregated data—for example by age or geographic detail—was typically average. The only more problematic dimension was related to timeliness: many data sources were one-off, resulting from specific projects rather than established collection efforts.

The overall quality rating, based on average scores summarising all quality dimensions with equal weights, indicated that the **available data supply is potentially useful for modelling**, especially after some corrections and adjustments. In addition to assessing the usefulness of available data for modelling, the quality dimensions also directly indicated the existing data gaps: what is potentially missing in the current body of evidence, in this case about Syrian asylum journeys. That timeliness was found particularly problematic points to an important data gap that could be filled through a dedicated asylum monitoring system, while issues with sample design or register completeness suggest the need for a methodological reflection on data collection mechanisms.

From the demand side, both the conceptual analysis of the model, as well as the formal statistical sensitivity analysis gave unequivocal results, identifying *information* as one of the primary factors shaping migrant journeys (Bijak *et al.*, 2021; Hinsch & Bijak, 2023). At the same time, **information remains one of the key data gaps** and sources of model uncertainty: empirical data about information use by migrants and exchange between migrants are scarce. Usable micro-level

Figure 1. Textual description of the Syrian migration data inventory in a word cloud. Source: Nurse and Bijak (2021), created with WordItOut.com.

Figure 2. Focus, data types, and collection modes in the Syrian migration data inventory (numbers of sources). Source: Nurse and Bijak (2021).

data on information are limited to a single source: a comprehensive, albeit just one-off German survey *Flucht 2.0* (Emmer *et al.*, 2016). In addition, modelling made use of aggregate estimates of the numbers of arrivals, border apprehensions, missing migrants and fatalities during journeys, originating from international organisations, such as UNHCR, International Organization for Migration, or Frontex (see Bijak *et al.*, 2021). In our models, the micro-level survey data ended up informing the assumptions about behavioural rules driving the agents' behaviour, while the macro-level estimates were used for calibrating the model outputs. To correct for identified data quality issues, especially the biases (underestimation), the aggregate numbers had to be transformed into relative measures and annual rates of change.

Including empirical data enabled some reduction in the model uncertainty (Bijak *et al.*, 2021). Still, bridging the data gap further would require additional collection of dedicated data, which in the example of modelling Syrian asylum seekers' journeys has been attempted in follow-up work, both qualitative and quantitative (Belabbas *et al.*, 2022; Bijak *et al.*, 2023), and verifying the remaining modelling assumptions through an extensive sensitivity analysis. In future work, the model construction can be also revisited, in order to provide a closer match to the available empirical basis, and **the process can be iterated, for as long as there are any information gains** and reduction in the model uncertainty.

Discussion and conclusions

In this note, we have described a process we propose for collating, assessing, selecting, and using secondary data for social simulation modelling purposes. The process, summarised in Figure 4, aims to reconcile the demand for data from the modelling side, with the available supply of reasonable-quality data that can be used to inform the model inputs or calibrate the outputs. The process is model specific, in that it always needs to begin by defining the needs defined by the particular modelling questions, with the data evaluated not only in their own right, but also with respect how well they serve the modelling objectives. The process then involves creating a dedicated data inventory, carrying out a quality assessment of the data, and analysis of data gaps both directly, from the quality assessment, as well as indirectly, through the lens of the modelling results. The identified gaps allow focusing on specific areas of further primary data collection and sensitivity analysis.

Figure 3. Quality dimensions of data sources included in the Syrian migration data inventory (numbers of sources). Source: Nurse and Bijak (2021).

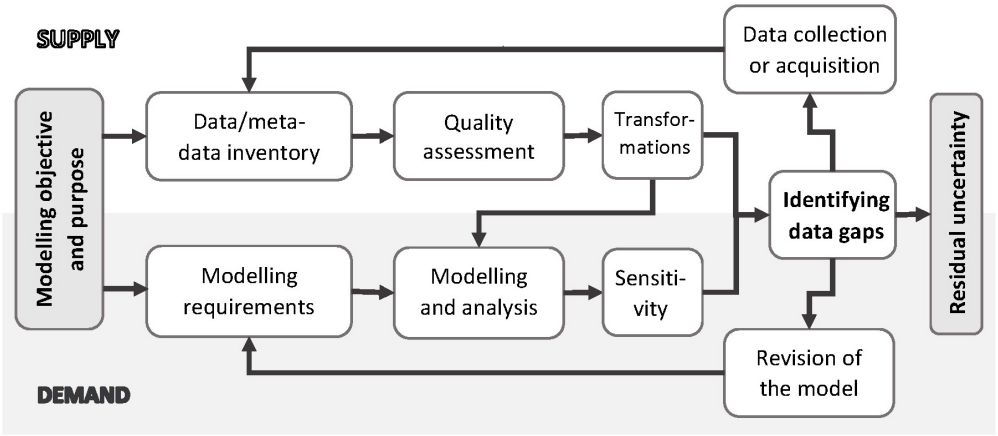

Figure 4. Iterative process for assembling and evaluating of secondary data for social simulation modelling. Source: Own elaboration.

The example of our modelling exercise, discussed in the previous section, can illustrate the workings of this process. Having defined the purpose and objective of data collection (to inform the modelling Syrian asylum journeys to Europe), enabled to set out the initial modelling requirements in terms of data on migrant decisions, information, and macro-level processes. At the same time, from the available data, collated and assessed for quality by following the template discussed above, we were able to identify a few sources that could be potentially useful for modelling. The macro-data, after transforming, were used for model calibration, and the *Flight 2.0* survey data provided important insights into the role of information in migration decisions. These data sources were used directly in the modelling process, but also leaving important gaps, reflected through residual model uncertainty. The sensitivity analysis showed information to be one of key gaps, which suggests a potentially fruitful area for further data collection in the future.

Any limitations notwithstanding, our findings point out to several practical implications for social simulation modellers with respect to the gathering of data and evidence for modelling. First, as summarised in Figure 4, the process of modelling enquiry needs to be driven by a specific research objective and purpose, and clearly acknowledge the limitations and uncertainty of empirical data. Second, data quality and importance in a given model needs to be formally analysed, both conceptually and through statistical means. Third, to facilitate that, data need to be thoroughly documented, for example by using provenance standards (Reinhardt *et al.*, 2023).

Finally, there are important trade-offs between the degree of empirical grounding and the level of detail of mechanistic realism of social simulation models. The more complicated the models, the fewer model elements can have reliable empirical basis, and the greater need for more disaggregated data – even though the aggregates are still likely to remain crucial, if only for calibration. At the same time, any users of social simulation models, especially in policy and practice-related applications, need to be made explicitly aware of data-related limitations. The need to take data uncertainty formally and openly into account when modelling is paramount both for transparency and with respect to the limits of scientific knowledge. This is especially important for such complex and politically charged topics as migration.

As an important caveat, the presented quality assessment of data sources has been inevitably based on expert judgement and has been arrived at through a deliberative process within the research team. For future extensions, one possible way of making the assessment more robust, would be to give it more formal structure, for example through a Delphi survey amongst many experts. Another could involve adding more specific classification rules for every criterion, for example accompanied by decision trees where a certain set of answers to questions about data would imply the final quality rating. In our exercise, however, we have found the process presented in this paper to be sufficiently fine-tuned to be able to identify the potentially-useful data sources and point to their shortcomings, such as biases. This in turn enabled correcting these shortcomings through appropriate data transformations and by making realistic assumptions about the errors.

Given that our quality assessment criteria are generic, and align well with the professional standards in official statistics (e.g. United Nations Statistics Division, 2018; Eurostat, 2019), the proposed framework should be broadly transferrable to other contexts, possibly with only minor modifications. The purpose can, of course, extend beyond modelling, and in other situations various criteria may be given different weighting, but the broad principles remain the same. Still, one additional aspect which may become important especially in some areas of application is the responsible use, including – but not limited to – data protection and security (IASC, 2023). It is easy to imagine contexts in which the analysis of especially individual-level data, if done carelessly or by malevolent actors, can result in harm to people or communities. In such situations, including additional criteria related to robustness of the data against possible misuses would be critical to safeguard the key ethical imperative of any well-intentioned social analysis: do no harm.

Ethics and consent

Ethical approval and consent were not required.

Data availability

Zenodo: Syrian Migration to Europe, 2011–21: Data Inventory. https://doi.org/10.5281/zenodo.7586826 (Nurse & Bijak, 2023).

The project contains the following underlying data:

- Syrian_migration_2011–21_data_inventory.tsv
- Syrian_migration_2011–21_data_inventory.xlsx

No primary data are associated with this article, as the analysis presented herein is based on a conceptual analysis of publicly-available metadata of various migration data sources, summarised in an online inventory https://baps-project.eu/inventory/data_inventory (Nurse & Bijak, 2021). The description of the agent-based simulation model of migration route formation used in this paper is available via https://baps-project.eu/inventory/project_outputs and the model itself is discussed in more detail in Hinsch and Bijak (2023).

Author contributions

Sarah Nurse: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Review & Editing

Martin Hinsch: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing

Jakub Bijak: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Acknowledgments

We are grateful to the GeoData Institute team at the University of Southampton, Ian Waldock, Sam Goddard and Jason Sadler, for making the metadata inventory available online in a searchable form via https://baps-project.eu/inventory/data_inventory.

References

Belabbas S, Bijak J, Modirrousta-Galian A, *et al.*: **From conflict zones to Europe: Syrian and Afghan refugees' journeys, stories, and strategies.** *Soc Incl.* 2022; **10**(4): 211–221.
Publisher Full Text

Bijak J, Higham PA, Hilton J, *et al.*: **Towards Bayesian model-based demography. Agency, complexity and uncertainty in migration studies.** *Methodos Series (volume 17)*. Cham: Springer, 2021.
Publisher Full Text

Bijak J, Modirrousta-Galian A, Higham PA, *et al.*: **Investigating immersion and migration decisions for agent-based modelling: a cautionary tale [version 2; peer review: 2 approved with reservations].** Brief Report. *Open Res Eur.* 2023; **3**: 34.
Publisher Full Text

Emmer M, Richter C, Kunst M: **Flucht 2.0: mediennutzung durch Flüchtlinge vor, während und nach der Flucht.** Institut für Publizistik, Freie Universität Berlin, 2016.

Eurostat: **Quality assurance framework of the European statistical system**. 2019.
Reference Source

GAO: **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, 2006.
Reference Source

Hinsch M, Bijak J: **The effects of information on the formation of migration routes and the dynamics of migration.** *Artif Life.* 2023; **29**(1): 3–20.
PubMed Abstract | **Publisher Full Text**

IASC: **IASC operational guidance on data responsibility in humanitarian action**. Geneva: Inter-Agency Standing Committee, 2023.
Reference Source

Kraler A, Reichel D: **Migration statistics.** In: Scholten P (ed.) *Introduction to Migration Studies.* IMISCOE Research Series. Cham: Springer, 2022; 439–462.
Publisher Full Text

Nurse S, Bijak J: **Syrian migration to Europe, 2011–21: Data Inventory.** Online resource, 2021.
Reference Source

Nurse S, Bijak J: **Syrian migration to Europe, 2011–21: Data Inventory (1.1).** [Data set]. *Zenodo*. 2023.
http://www.doi.org/10.5281/zenodo.7586826

Oakley JE, O'Hagan A: **Probabilistic sensitivity analysis of complex models: a Bayesian approach.** *J R Stat Soc Series B.* 2004; **66**(3): 751–769.
Publisher Full Text

Reinhardt O, Prike T, Hinsch M, *et al.*: **Simulation studies of social systems – telling the story based on provenance patterns.** *TexRxiv.* Preprint, v.2.0 (28 April 2023). 2023.
Publisher Full Text

United Nations Statistics Division: **UN statistics quality assurance framework. including a generic statistical quality assurance framework for a UN agency**. New York: United Nations, 2018.
Reference Source

Vogel D, Kovacheva V: **Classification report: quality assessment of estimates on stocks of irregular migrants (Report of the Clandestino project).** Hamburg Institute of International Economics, 2008.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ? ✔ ✔

Version 2

Reviewer Report 14 October 2024

https://doi.org/10.21956/openreseurope.20171.r44897

✔ **Alejandra Rodriguez-Sánchez**
University of Potsdam, Potsdam, Germany

I have read the manuscript and author responses to my previous comments and I find the paper is now in a better form.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 16 February 2024

https://doi.org/10.21956/openreseurope.16843.r37144

✔ **Denis Kierans**
University of Oxford, Oxford, England, UK

The article outlines a carefully thought out approach to making sense of data that are of variable quality but carry high levels of political and operational significance. Of particular interest is the transparency the authors provide by reflecting on the types of data under review; the creation of categories that fall between the typical green, amber and red; the decision to prioritise high levels of trustworthiness and transparency among the various criteria. The article is clear about the

limits of modelling in this area, which is also well received. For these reasons and more this succinct article is a fine contribution and will hopefully lead to further well-documented, sober reflections on data quality (etc.) in other areas of social science research.

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Areas of research: migration, migration data, migration policy, integration.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 26 Sep 2024

Jakub Bijak

Thank you very much for your review and very kind words: we are delighted to hear you find the material useful!

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 February 2024

https://doi.org/10.21956/openreseurope.16843.r37139

Alireza Jahani
Brunel University London, London, England, UK

The paper provides a detailed and well-structured methodology for collating, assessing, and selecting secondary data for social simulation modelling, with a specific focus on Syrian asylum migration. The paper begins by outlining the need for more realistic and reliable data to meet the increasing demand for policy-oriented insights into migration patterns. It emphasizes the importance of assessing data quality and identifying knowledge gaps in the context of modelling endeavours. The structured approach involves creating a comprehensive inventory of data sources, assessing data quality using predefined criteria, and mapping the thematic areas to visualize data uncertainty.

The methodology presented in the "Methods" section is clear and well-defined. The authors rightly highlight the multidimensional nature of data quality, covering aspects such as timeliness, level of aggregation, trustworthiness, and transparency. The use of a traffic-light approach and a summary score provides a practical and intuitive way to evaluate data quality. The inclusion of a case study on Syrian asylum migration enhances the paper's applicability and demonstrates the methodology in action.
As an expert in the field that modelled the Syrian conflict forced migration before, I find the approach to be comprehensive, and the paper addresses a critical aspect often overlooked in social simulation studies—the quality and availability of secondary data. I remember we wanted to couple our model with telecommunication dataset as secondary data and faced a lot of challenges.

In the "Results" section, I expected to see more details of the case study, with a focus on the data inventory, thematic mapping, and quality assessment of available data. However, I believe that the identification of data gaps and the iterative process of data collection are particularly valuable insights for researchers and practitioners alike.

The paper concludes with a thoughtful discussion on the challenges, solutions, and recommendations for social simulation model builders and users. It emphasizes the importance of a question-driven rather than purely data-driven approach, acknowledging the inherent uncertainty in social simulation models. The iterative process outlined in the discussion offers a practical way to bridge data gaps and improve model reliability.
Overall, the paper is a valuable contribution to the field of social simulation modelling. It provides a systematic and transparent approach to handling secondary data in modelling endeavours, making it accessible to both experts and non-experts in the field. While the paper is comprehensive, providing more illustrative examples or specific insights from the case study could further enhance its impact and clarity for a broader audience.

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

Are all the source data and materials underlying the results available?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Forced Displacement Modelling and Simulation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 26 Sep 2024

Jakub Bijak

Thank you very much for your review and the endorsement of the paper. In the initial version, we were constrained by the limits of the Research Brief format, but we have now added some more details on the case study, with particular focus on the assessment process, mapping of the available data, and how the results of the exercise were subsequently used in the modelling. We hope this helps explain the process and our intentions behind it clearer.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 January 2024

https://doi.org/10.21956/openreseurope.16843.r36789

?

Edgar Scrase
United Nations High Commissioner for Refugees (UNHCR), Geneva, Switzerland
Zaruhi Mkrtchyan
Global Data Service, UNHCR, Copenhagen, Denmark

We welcome this research – systematizing the review of secondary data is useful, and relevant. There are many research teams that could benefit from such a structured approach. The

approach is well presented and the structure and flow of the paper are good with a logical progression from presenting the framework and then unpacking it as a case study. The Syrian refugee example is relevant and serves as a practical example of how the theoretical model can be implemented and operationalized.

The framework described to assess the quality of the data is useful. We would welcome greater explanation of how the traffic light colours were assessed – i.e. are there a series of criteria in each quality dimension that lead to a general score? Or is it an expert assessment of each dataset? Additional detail on this would help readers to operationalize the approach in their own research.

There is also no mention of statistical quality assurance frameworks such as the work by UN stats division (https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf) and the European Statistic System (https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646). These are well researched and describe a broader set of quality dimensions that are relevant to this research. The UNSD SQAF in particular also describes the inherent compromises that need to be made between quality dimensions (i.e. prioritising timeliness over accuracy etc.) which could usefully be mentioned in this paper. The broader set of quality dimensions could also help to provide a set of criteria that could be more broadly applicable to other research. For example, greater mention could be made about using data responsibly. It's implicit in this research example as only public data is used, but in other research some of the qualitative individual data may contain personal data, for which the researchers and ultimately the data owners must take responsibility in assessing whether it is appropriate to publish such data or even to use it in the research. This is well covered in this IASC paper on data responsibility (https://interagencystandingcommittee.org/operational-response/iasc-operational-guidance-data-responsibility-humanitarian-action).

A broader set of quality dimensions would also cover data availability / data gaps as well, which are mentioned in the results section. I.e. for the quality dimensions to be truly useful they would ideally help to assess both the available data as well as what is potentially missing.

In terms of terminology, we would suggest to avoid "asylum migration" as a term as it conflates two reasons for people moving. Either the Syrians are seeking asylum in order to become refugees or they are migrating for other reasons. We would suggest "Syrian refugees arriving in Europe" for the article.

Thanks again for the opportunity to review, all work like that expressed in this paper, to operationalise quality improvements to data and data processing are very welcome and we would be interested to collaborate in future activities in this domain.

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Forced displacement population statistics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 26 Sep 2024

Jakub Bijak

Thank you for very careful reading of the paper and for very helpful comments. In the first version, we opted for brevity, to meet the constraints of the Research Brief submission, but we agree that additional details would be helpful, including on the aggregation of quality scores, as pointed out also by the other reviewers. We hope the new additions in the Methods and Results sections are helpful, and we summarise the key changes below.

- Many thanks for the excellent references to the official statistics standards on data quality and the associated documentation (which we have now added in the Methods section), and to the point about responsible use of data, with which we fully concur, and are now mentioning in the Discussion and Conclusions. We allow ourselves to credit you for the idea, but agree wholeheartedly about the importance of the ethical reflection for data use, especially at the individual level.
- Thank you also for the suggestions of additional quality dimensions, which we have now mentioned in the Results and Discussion and Conclusions of the paper as the direct result of the application of the evaluation framework (as opposed to indirect, following the modelling results).
- As to the terminology, your point is well taken: even though contemporary migration scholarship would point to multiple motives for different types of mobility, including asylum, we agree that a more precise term would help. We have thus pivoted to "Syrian asylum seekers" throughout, as this is about people who have not (yet) been granted refugee status, although we also mention "refugees" in the context of the database, which may include data relevant to this population.

Competing Interests: No competing interests were disclosed.

Reviewer Report 24 January 2024

https://doi.org/10.21956/openreseurope.16843.r37138

?

Alejandra Rodriguez-Sánchez
University of Potsdam, Potsdam, Germany

This paper presents an approach for collating assessing and selecting appropriate secondary data to build agent-based simulation models of migration. It is applied to the specific case of the Syrian asylum migration to Europ during the 2010s.

Evaluating the information available for modelling analysis is a fundamental aspect of model development. My main take from the paper is on the importance of incorporating data critically in the building of an ABM for migration.

However, here are some recommendations and questions that might improve the manuscript and the exposition of the method:

- The paper could make clearer how the discussion of the data quality led to decisions in the modeling part. This is left implicit, but readers could benefit from that.
- Figure 1, a word cloud, is rather uninformative and includes terms that aren't relevant (e.g., rather, since, other, detail etc.). I am not sure if this was the starting point for doing something else on the basis of word frequency, like classifying the types of data sources.
- The quality assessment of the data sources seems rather ad hoc. Here is perhaps where some notion of a standard could be applicable or at least determined. What is a high quality data source? What is a bad quality? Isn't a bad quality data source useful for modelling in the absence of any other information?
- On a similar note, depending on the type of data, an evaluation of its quality will have to follow different criteria. How were criteria differently reconciled here? And more specifically, how did you arrive at the score in the table? Would it be possible to describe, at least for some case, how did you come to evaluate a data source as green-amber v. amber? In general, the quality of a data source should be evaluated on the basis of a given model (how useful is it to model that specific element of the phenomenon to be simulated). Is it something like the more detailed the model, the less useful aggregate data might be?
- If this piece is to serve as a template, as described by the authors, it should have such a structure. The diagram in figure 4 is, in this sense, quite illustrative, but I miss a reflection of each of the elements, perhaps in a separate paragraph, that walks the reader through an example of a data source and how it affected the modelling.
- What factors would we have to consider if we were interested in modeling other forms of migration or migration in other parts of the world? How specific or transferable are the different steps in this template?

All in all, authors are attempting to provide guidance on how to incorporate data into the building of ABMs, which would be a great contribution.

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Migration, family demography, causal inference

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 26 Sep 2024

Jakub Bijak

Thank you for the careful reading of the paper and for your feedback. We have now, also in response to the other reviewers, expanded on the description of the process of incorporating the data quality assessment in the modelling, in the Results section, which we hope will help the readers see not only the details of the process, but also its end-point. Our responses to your specific comments are listed below:

- Regarding the word cloud map in Figure 1, it was meant to visualise the thematic map of the inventory, which we now state explicitly in the text. You are right that some terms may be common words, although would disagree about some specifics (e.g. to us 'other' carries extra meaning in that 'other' means 'outside of pre-defined categories'). Still, we have cleaned the figure by removing some of the obviously irrelevant terms and have expanded the discussion, to include the exploratory character of this de facto qualitative analysis of the database content.
- As to the assessment of data sources, it is inevitably based on some dose of judgement, which does not necessary mean ad hoc, as it has been arrived through a deliberative process within the team – here, one possible way of making the assessment more robust, would be to give it more formal structure, for example

through a Delphi survey amongst experts, or similar. In addition, We now also explain the process of aggregating the quality scores more clearly.
- Thank you also for the request to explain the aggregation of individual criteria clearer, which we have now attempted in the current version of the paper. You are right to point out that the assessment is model-specific, and we now make this point explicit. The point about aggregate data is more subtle, but we attempt spelling it out as well: in brief, the more detailed the model, the greater the need for disaggregated data, but higher-level aggregates are still likely to be useful for benchmarking the project output.
- The decision not to include specific examples was driven by the initial constraints of the Research Brief format, but in this version, besides adding a comment on the process by which the data assessment enters into the modelling, we have also provided cross-references to two example sources of data (UNHCR and Flucht 2.0), to illustrate it better.
- Finally, we have added a comment on transferability to the Discussion and conclusion part of the paper, with thanks for the suggestion – we agree it adds value to the argument.

Competing Interests: No competing interests were disclosed.