ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Check for updates

Article submitted to journal

Subject Areas:

xxxxx, xxxxx, xxxx

Keywords:

computational modeling, data, provenance model, social simulation

Author for correspondence:

Adelinde M. Uhrmacher

e-mail: adelinde.uhrmacher@uni-rostock.de

Simulation Studies of Social Systems – Telling the Story Based on Provenance Patterns

Oliver Reinhardt[1], Toby Prike[2], Martin Hinsch[3], Jakub Bijak[4], Pia Wilsdorf[1] and Adelinde M. Uhrmacher[1]

[1]Institute for Visual and Analytic Computing, University of Rostock

[2]School of Psychological Science, University of Western Australia

[3]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

[4]Department of Social Statistics and Demography, University of Southampton

Social simulation studies are complex. They typically combine various sources of data and hypotheses, that are integrated by intertwined processes of model building, simulation experiment execution, and analysis. Various documentation approaches exist to support the transparency and traceability of complex social simulation studies. In particular, provenance patterns can be used to capture central activities and entities of a simulation study. Entities can include, simulation models, experiments, or research questions, and activities – model building, calibration, validation, and analysis. The exploitation of provenance standards enables information on sources and activities, which contribute to the generation of an entity, to be queryable and computationally accessible. In this study, we refine the provenance pattern-based approach to address specific challenges of social agent-based simulation studies. Specifically, we focus on the activities and entities involved in collecting and analyzing primary data about human decisions, and the collection and quality assessment of secondary data. We illustrate the potential of this approach by applying it to central activities and results of the Bayesian Agent-Based Population Studies project and by presenting its implementation in a web-based tool.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

Reproducible and interpretable simulation studies require thorough documentation of activities, sources, and products involved in this process [1,2]. Simulation studies involve complex modeling and analytical processes, in which activities such as model building and refinement, conducting simulation experiments, data processing, and interpretation are closely intertwined (Figure 1). These studies often span several years. Their documentation, therefore, requires significant effort and several reporting guidelines have been developed [1–3]. Computational support for recording crucial information about simulation studies, such as data, subsequent versions of simulation models, modeling assumptions, research questions, etc. has taken multiple forms. These include adopting documentation templates [4], archives [5], Wikis [6], electronic notebooks [7], and provenance standards with a graph-based database [8].

Generally speaking, all efforts in documenting simulation studies are—at least implicitly—concerned with provenance. That is, providing "information about entities, activities, and people involved in producing a piece of data or thing" [9]. The benefit of adopting a standard for provenance, such as W3C PROV [9], is that the various sources, activities, and products of a simulation study are put into well-defined relation(s) with each other. Its graph-based abstraction provides a historical and causal delineation of what contributed to a simulation model and how it did so in a simple and formal manner [10]. It can be mapped into a graph database which allows—in addition to storing the information—filtering and querying the stored information on demand [11]. Its graph-based visualization (e.g., in a web-based tool) makes it possible to easily access and assess dependency structures within and across simulation studies [8]. Provenance standards have already been applied to cell biological simulation studies [8,12] and to documenting a migration model in demography [13].

In demography, as is common for social science disciplines, the need to combine hypothesis-driven and data-driven modeling adds to the complexity of simulation studies [14]. Thus, there is also greater effort required for their thorough and systematic documentation. The situation becomes even more complicated whenever primary data about human behavior are collected, such as through interviews or psychological experiments conducted as part of a broader agent-based simulation study, or once evaluation schemes are used to account for uncertainty in secondary data. This is especially relevant in the context of the paradigmatic shift in demography towards more micro-level and multi-level studies [15] and the recognised need for greater use of simulation models to enhance the theoretical base of the discipline [16]. This is also in line with the general developments in social simulation more broadly, which explicitly recognises issues such as data quality and the necessity of collecting bespoke primary data for simulations [17]. So far, the diversity of sources, products, and processes has hampered the systematic and accessible documentation of entire demographic simulation studies.

To systematically and accessibly document entire social simulation studies, which include extensive data collection, evaluation, analysis, and adaptation, in this paper we present an approach based on provenance patterns specified in the W3C PROV standard. In [18], provenance patterns have been identified for the documentation of and reasoning about central activities of simulation studies, such as model building and refinement, or conducting simulation experiments for verification, calibration, and validation. These patterns shall now be extended to capture data evaluation schemes used to assess the quality and uncertainty of secondary data sources, and to the collection of primary data, exemplified by psychological experiments or interviews conducted to support the agent-based modeling of human decision processes. These patterns also take reporting guidelines in the respective areas into account. To do so, we use and adapt the tool WebProv, which combines a web-based visual interface and the graph database Neo4J to store and retrieve provenance information based on these patterns.

We demonstrate our approach by applying it to the research project BAPS (Bayesian Agent-Based Population Studies[1]), which aims at "transforming the study of migration – one of the most

[1] https://baps-project.eu/

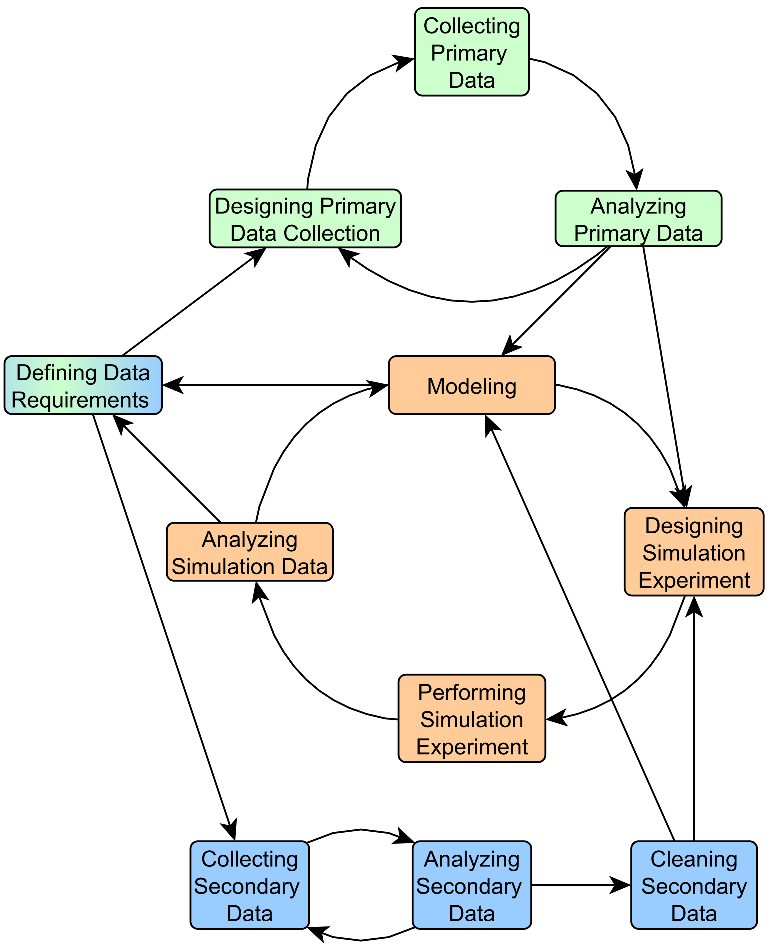


Figure 1. Central activities of the modeling and simulation lifecycle (orange) - including the procurement of primary data (green) and secondary data (blue).

uncertain population processes – in the way it can be understood, predicted, and managed". In this project, the development of a simulation model to analyze the formation of migrant routes from Syria to Europe is complemented by a framework for assessing existing secondary data, and their quality [19]. The project also involved the acquisition of primary data by conducting psychological experiments on human decision-making under uncertainty [20], and carrying out ethnographic interviews to provide richer contextual information [21].

The contributions of this study are the following:

(i) to identify crucial activities and entities for documenting data acquisition, quality assessment, and primary data collection (here, psychological experiments), and to encode them as patterns in a provenance standard,

(ii) to integrate this meta-information with previously identified patterns for conducting simulation studies,

(iii) to apply the patterns to the activities and results achieved within a major research project on migration to provide comprehensive documentation of the research done in this project, and

(iv) to implement the existing and newly developed patterns for data procurement in an openly published, web-based provenance tool.

With our provenance pattern-based approach, we structure the knowledge about model building, analysis, and the acquisition and use of primary and secondary data, as well as the methods of analysis. Documentations based on provenance graphs can thus now formally represent the entire story of a simulation study (in this case, of social systems). This is an important step forward since most insights come from the relationships between artifacts, particularly the relationships with data, but also theories or hypotheses embodied in the existing knowledge (literature). Modelers can benefit from the proposed approach by easily detecting gaps or inconsistencies in a simulation study thanks to the queryable nature of provenance graphs. The provenance information is also essential for many aspects of good scientific practice, including replicating the results, reusing simulation models, or interpreting the simulation results correctly in the light of available evidence. In addition, we expect the approach to be instrumental in assessing the robustness, reliability, and relevance of the data collected, as well as how the insights gained from the data and the uncertainty or errors of the data collection procedure may propagate to other artifacts and processes. This can, in turn, illuminate the data and knowledge gaps, and help direct further scientific enquiry.

2. Case Study: Developing an ABM of Migration Route Formation

Migration is a highly complex and uncertain population process, being driven by the decision-making of individuals and various institutions. Migration routes are highly volatile, with the flows responding to changes in various migration drivers, broader environments, and individual circumstances, which can sometimes change very rapidly [22]. In the study shortly presented here, agent-based simulation is applied to improve the theoretical understanding of human migration, with a specific focus on the question of how migration routes are established and sustained.

The core of the study [23] is the development of an agent-based model of migration route formation [24]. Therein, modeled migrant agents attempt to traverse an abstract landscape based on limited and uncertain information about locations on the way, potential paths, and the involved risks. As model development is an iterative process [25], multiple model versions were designed in succession, informed by knowledge from the scientific and non-scientific literature on the migration process, knowledge about decision-making, and lessons learned from previous iterations. For the latter, extensive simulation experiments with the model were necessary. To that end, at each step Gaussian Process emulators were fitted to the model outputs of each model version to assess sensitivity to the input parameters and the uncertainty of the results. While earlier model versions were very abstract and theoretical, later versions were designed and calibrated to capture the reality of migration routes in the Central Mediterranean. Thereby, a considerable amount of data was integrated into the model.

Data in general, especially migration data [26], tends to be difficult to compare and may sometimes be incomplete or of dubious quality. Hence, an important part of the project was assessing available data on asylum migration. To this end, an assessment framework was designed and applied to various potentially useful sources of migration data [19] so that the data were supplemented with the necessary meta-information about quality to enable the use of the data in the simulation study. This migration data from secondary sources was also complemented with information on the migrants' decision processes, elicited as primary data through psychological experiments and interviews which were designed to answer specific questions that arose during the modeling work. For example, the sensitivity analysis of earlier models highlighted information sharing and trust in information as key influences in forming migration routes. Subsequently, in a bespoke psychological experiment, data on migrants'

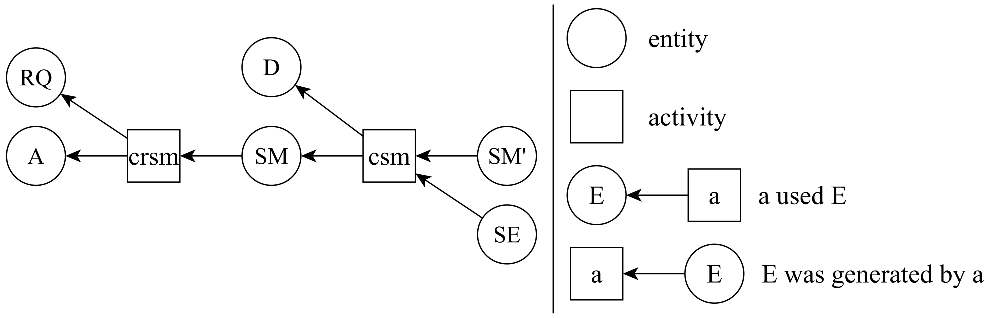


Figure 2. An example of a provenance graph as defined by PROV. The graph shows two typical activities in a simulation study. First (crsm) a simulation model is created, based on a research question (RQ) and an assumption about the modeled system (A). This produces a model (SM). The model is then calibrated (the activity csm) against some data (D), producing a calibrated model (SM') and a specification of the performed calibration experiment (SE).

subjective judgements based on different kinds of information and sources were collected. The results were then used to inform the parameterisation of the successive model versions [27].

The migration case study highlights that simulation studies of complex social systems are themselves complex and intertwined processes that include the modeling work itself, the execution of simulation experiments, the collection and assessment of secondary data sources, and the collection of new data to inform the model, as shown in Figure 1. Broader philosophical underpinnings of such a model-based approach, within which the iterative model development is situated, are discussed in more detail in [23, Chap. 2].

These diverse modeling activities, data, and all of the utilised information sources are dependent upon one another and contribute to the products of simulation studies. Each of these products can only be properly interpreted if their generation context is fully taken into account. Therefore, accessible and thorough documentation of simulation studies becomes of utmost importance.

3. Concept

3.1 Provenance Models and Provenance Patterns

The provenance of a simulation model documents the process of creating the model, including: what questions it was designed to answer, on which underlying theory and data it is based, how it was constructed, and how it was experimented with. This back-story of a model is crucial for interpreting and reusing a model, as well as for assessing the quality of the model and the results it generated. Following the W3C PROV standard [9], provenance information can be represented as a directed acyclic graph with two types of nodes: entities and activities. Edges between entities and activities relate the two (see Figure 2), specifying which entities were either *generated by* or *used by* which activities.

Applying PROV requires specializing the PROV Data Model by specifying types of entities and activities, and possible relations between them. For the modeling and model analysis, the central part of a simulation study (see Figure 1), important entities and processes have been identified in the literature [8,10]. Building on this, Wilsdorf et al. [18] identified *provenance patterns* for model building and simulation experiments arising in a simulation study: certain activities within a simulation study will always use and produce certain types of entities. A pattern consists of an activity at its center, and the types of entities that are used and produced by this activity. For example, creating a simulation model (crsm in Figure 2) will always produce a model, and calibrating a simulation model (csm) will always use a model and a calibration target, and will always produce a calibrated model and a specification of the calibration experiment. This example matches the patterns *Creating Simulation Model* and *Calibrating Simulation Model* (see Figure 3).

In a provenance graph, we can annotate entities with meta-information that contains the entities' documentation. We recommend this meta-information follows established reporting guidelines for these types of entities or refers to a document following such guidelines, such as to an ODD [Overview, Design concepts, and Details] document for an agent-based model [4]. Additionally, the meta-information should include references to all relevant artifacts, such as the implementation of the model or the data set for a data entity.

However, the modeling and model analysis itself, although central, is only part of a social simulation study, which also needs to grapple with the agency of the objects of the scientific enquiry (human beings) and the resulting high levels of uncertainty of the related social processes. In demography, migration is the one component of demographic change which – unlike fertility or mortality – does not have explicit biological underpinnings, and is thus much more challenging to analyze due to the high levels of agency of the various actors, and the highly complex underlying factors and drivers [23].

Another important ingredient is the data that grounds the model in reality - and the process of its collection. We distinguish primary and secondary data collection as follows:

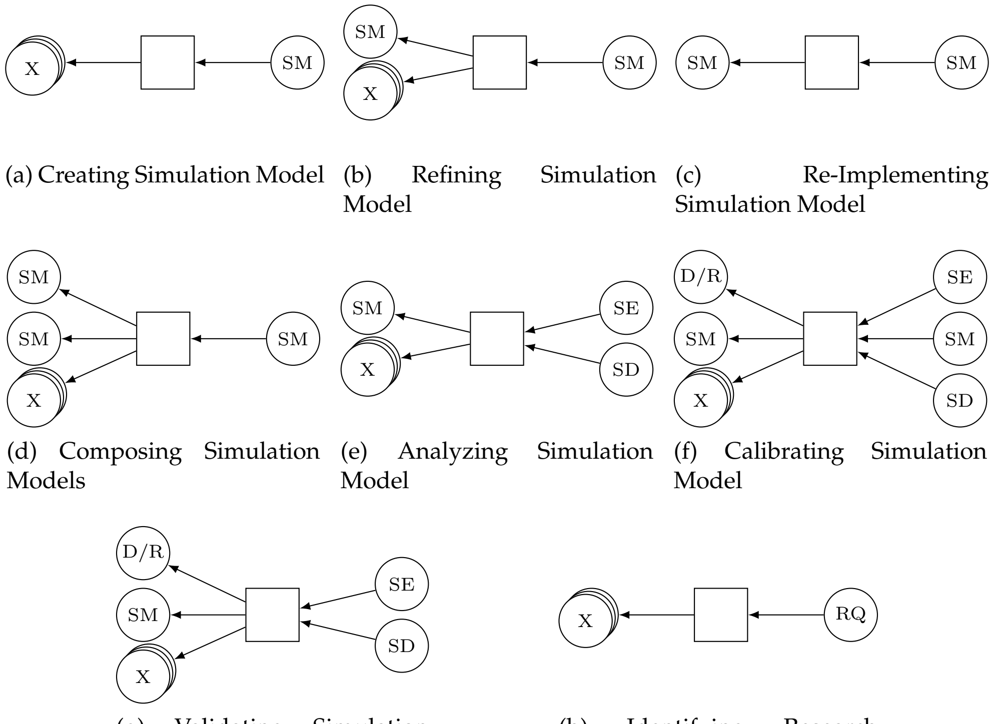
- Primary data was collected specifically by the conductors of the simulation study for the simulation study itself. This collection may take the form of surveys, interviews, psychological experiments, etc.
- Secondary data was collected for another purpose, typically by someone else. Hence, its suitability must be assessed, and the data may need to be cleaned to account for various sources of uncertainty and biases.

In this work, we extend the method of provenance patterns to consider the data-related processes. In the following, we summarise the entities and activities and the arising patterns in modeling and model analysis, as defined in [18]. Building on that, we then extend the scope of the approach by identifying entities, activities and patterns for primary and secondary data collection.

3.2 Modeling and Model Analysis

Wilsdorf et al. [18] mapped the considerations of existing reporting guidelines (e.g., [2,28]) into the provenance standard W3C PROV, by identifying the central activities of a simulation study. As in [18], we distinguish the following entities in the modeling and analysis part of the study; *Research Question* (RQ), *Simulation Model* (SM), *Simulation Experiment Specification* (SE), *Simulation Data* (SD), *Requirement* (R), *Assumption* (A) and *Other* (O). In the patterns, we also make use of an entity type *Data* (D), as a general placeholder for any kind of data, including simulation data produced in an experiment, but also primary and secondary data (which we otherwise distinguish due to different documentation requirements). Research questions may also appear in other parts of the simulation study, as they also provide the motivation for primary and secondary data collection.

Figure 3 shows the patterns graphically. The upper four patterns describe activities in the modeling process. When a new simulation model is created from scratch, the pattern *Creating Simulation Model* (a) applies. That activity uses various inputs, e.g., a research question, assumptions, theories or data, represented by the wildcard (X) in the pattern, and produces a simulation model (SM). When an existing model is refined, the pattern *Refining Simulation Model* (b) applies instead, which has an additional input in the form of the existing simulation model. For example, the model from the case study was refined when new data from psychological



(a) Creating Simulation Model (b) Refining Simulation Model (c) Re-Implementing Simulation Model

(d) Composing Simulation Models (e) Analyzing Simulation Model (f) Calibrating Simulation Model

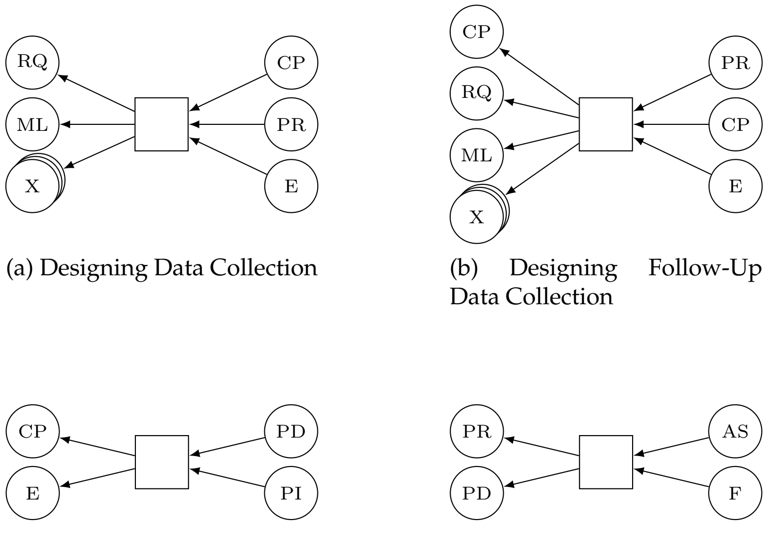
(g) Validating Simulation Model (h) Identifying Research Question

Figure 3. Provenance patterns for model development and model analysis activities (a to g from Wilsdorf et al. [18]).

experiments became available and could be used to improve the decision-making mechanisms. The pattern *Re-Implementing Simulation Model* (c) refers to an activity, where a simulation model is re-implemented in another language or tool, without refining or extending it. For example, in our case study, we re-implemented the model, originally implemented in Julia, in the modeling language ML3 to cross-check both models [29]. Finally, the pattern *Composing Simulation Models* (d) describes the composition of two simulation models. For example, when we combined the Julia and the ML3 implementation of models to gain a new Julia implementation that follows a similar rule-based approach as the ML3 implementation.

The patterns (e) to (g) describe activities during the analysis of and experimentation with a simulation model. When a simulation model is analyzed (the pattern *Analyzing Simulation Model* (e)), e.g., via a sensitivity or uncertainty analysis, a simulation model (SM) is used as well as potentially some other inputs (X). The result is some simulation data (SD), and a simulation experiment specification (SE), e.g., a script or description that allows the analysis to be repeated. For example, when one of the models developed in the case study was subject to uncertainty and sensitivity analysis (see [23], Chapter 8.3), that model served as the (SM) input. As this analysis was a repetition of an earlier analysis performed on an earlier model version, that earlier simulation experiment specification was used as an additional input (X). The uncertainty and sensitivity characteristics determined through the experiment are referred to in the simulation data entity (SD). The cited chapter may serve as a reference in the simulation experiment specification entity (SE). If there were scripts or other software artifacts to repeat the experiment, it would be preferable to reference these in the entity (SE). However, this analysis was performed with a GUI-based tool, so there were no scripts produced. The pattern *Calibrate Simulation Model*

royalsocietypublishing.org/journal/rsos Proc R Soc A 0000000



(a) Designing Data Collection

(b) Designing Follow-Up Data Collection

(c) Collecting Primary Data

(d) Analyzing Data

Figure 4. Provenance patterns for primary data collection.

(f) for model calibration is similar, but requires an additional input: some data or a requirement (D/R) that serves as the calibration target. A calibrated simulation model is produced as an additional output. The pattern (*Validating Simulation Model* (g)) is defined in a similar way to the calibration pattern. The only difference in the pattern is that validation does not produce a simulation model. For validation, the model behavior is compared with the data or requirement (D/R) and the results are stored as simulation data (SD).

Compared to Wilsdorf et al. [18] we added one additional pattern *Identifying Research Question* (h). Newly identified research questions are often a major driver of long-term simulation studies - as well as important results in and of themselves. For example, the modeling work may identify gaps in the data, that lead to research questions for data collection efforts, or the collected data may show interesting properties that pose new questions. In general, any entity (or combination of entities) might lead to new questions. Hence, the pattern for *Identifying Research Question* allows any input (X) to produce a research question (RQ). For example, the results of a sensitivity analysis of the model highlighted the importance of information sharing for model behavior. This led to a new research question for a psychological experiment, which was, in this case, how migrants judge information received from different sources.

3.3 Primary Data Collection

Primary data collection includes the design of a collection procedure, the execution of the collection procedure to gather data, and the analysis of the data to gain insights.

We begin by defining the relevant types of entities, using the case study experiment on migrant judgements of information from different sources as an exemplar, and then continue with defining patterns for the activities.

- **Methodology Literature (ML):** When designing a data collection procedure, researchers often rely on reusing or adapting methodologies from existing research. By including information about key papers that have informed the data collection procedure, other researchers are better able to understand, reproduce, and assess the data collection procedure, as well as the primary data and findings that are generated. For example, in the psychological experiment on migrant judgements and decisions, there were two key papers that informed the methodology. One was a review paper by [30] on how source impacts how people assess and make judgements and decisions based on information. The other was a paper by [31] that examined how people convert different verbal likelihood statements into numerical judgements.

- **Data Collection Procedure (CP):** The data collection procedure determines what data will be collected and how. Depending on the form of data collection, e.g., a survey, interviews, or psychological experiments, this entity may take different forms. For example, it might be a questionnaire, interview questions and instructions for the interviewer, or even a piece of interactive software that is presented to participants. In any case, when presented to the participants, the Data Collection Procedure allows data collection to be undertaken. The primary data collection conducted in the case study used a psychological experiment setup within the Qualtrics[2] survey software. This Data Collection procedure collected data on how people make risk-related judgements and decisions in response to migration-related safety information presented by a variety of sources. For an example of the data collection procedure used please see https://southampton.qualtrics.com/jfe/form/SV_20kQsSP0cyi6o06.
- **Participant Information (PI):** To allow for the assessment and reproduction of primary data collection, it is crucial to provide information about the participants included in the study. This includes information such as which populations they were recruited from, how they were recruited, and any specific requirements or exclusions that were used (e.g., language requirements, demographic characteristics, attention check questions etc.). Providing this information also allows other researchers to assess the primary data collection (e.g., whether the participants were appropriate to address the research question and support the findings) and to decide whether the data and/or findings are appropriate for other researchers to rely on or reuse (e.g., if they can be transferred to a new population of interest). Participant information about the exemplar experiment included in the current study can be accessed by looking at the demographic information included within the primary data[3] as well as by looking at the preregistration[4] for information about how many participants were included and what (if any) inclusion and exclusion criteria were used.
- **Preregistration (PR):** Preregistration is a document outlining several key aspects of a study methodology and analysis plan. Some of the key details included within preregistration are: the specific research questions and/or hypotheses, the methodology that will be used (e.g., dependent variables and independent variables/experimental conditions), the participant sample size to be collected along with exclusion or inclusion criteria, and the planned analyses that will be used to answer the research questions/test the hypotheses (for an example, see https://osf.io/3qrs8).
- **Ethical Approval (E):** Primary data collection from human participants, be it through interviews or psychological experiments, requires adherence to ethical standards that are set by the funders and institutions carrying out the data collection. Here, the Ethical Approval refers to the final version of the research ethics application, approved by the relevant body, which documents the interview/experiment schedules (Data Collection Procedure), and Participant Information and Consent forms, which sets out the conditions and standards of data collection, storage, use, and re-use. The experiment conducted as part of the case study underwent ethical review and approval by the University of Southampton Ethics Committee prior to data collection being conducted (ERGO Approval: 56865).
- **Primary Data (PD):** The data is the principal result of primary data collection. Depending on the kind and scope of data collection, it may take the form of a table or a set of tables,

[2] https://www.qualtrics.com
[3] https://osf.io/cqh6j
[4] https://osf.io/3qrs8

interview transcripts or summaries (excerpts, codes), or a database. In any case, this entity is a representation of the raw output of the data collection, potentially anonymised or pseudonymised if necessary, that may then be analyzed in further steps. The primary data for the psychological experiment in the case study is available in an OSF repository[3].

- **Findings (F):** The findings refer to the key conclusions or results generated by analyzing the data. These can take a variety of formats, such as a written results section, graphs, tables, or descriptive statistics (e.g., means, medians, correlations etc.). These findings can subsequently be used as inputs for modeling activities in a variety of ways, including to set or inform model parameters, to help specify the direction of relationships between model variables, or to test the broader implications of findings (e.g., how a masking intervention influences disease spread through a societal network). The results section that outlines the findings from the experiment on migration-related risk judgements and decisions can be found within the following paper [20].
- **Analysis Specification (AS):** To reproduce the findings, it must be possible to repeat the analysis of the data precisely, either with the same data or with comparable data, e.g., from a follow-up study. Hence, a specification of the conducted analysis is required. Often, the analysis will be conducted with some statistics programming language or library, e.g., in R or in Python. In this case, analysis scripts are a natural result of the analysis process and will allow for easy analysis repetition. For example, the analyses for the case study experiment were conducted in R, and the analysis scripts have been made publically available at the following link: `https://osf.io/ws63f/files`. If such scripts do not exist, e.g., if a GUI-based analysis software is used, or if the scripts are not sufficient on their own, the specification of the analysis may also be textual.

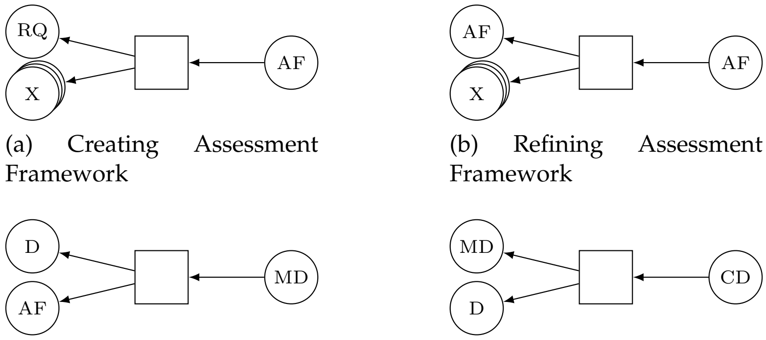
We identified the following patterns for the primary data collection activities (see Figure 4).

(a) **Designing Data Collection:** Before any data can be collected, the data collection process must be designed. This process requires a research question and methodology literature as inputs, but can also have other inputs (X). One example of a potential additional input (X) is the literature from the substantive area of relevance (in this case, cognitive studies of human decision-making). We identify three core products of the design phase: the data collection procedure, the preregistration, and the ethics document.

(b) **Designing Follow-Up Data Collection:** Some data collection efforts are designed to follow up on a previous one, e.g., to replicate the result, to refine the procedure, or to answer new questions raised by the findings. In this case, the data collection procedure (CP) of the original experiment is an additional input when designing the follow-up experiment. Including the original data collection procedure as an input connects the follow-up data collection to the original data collection and shows how the data collection procedure has been refined across multiple rounds.

(c) **Collecting Primary Data:** Once the data collection is designed, the data can then be collected. Participants are recruited, and the data collection is executed with them, e.g., they are given the survey or are interviewed. This process is based on the previously designed data collection procedure (CP) and must conform to the ethics document (E). Hence, both are inputs to this activity. The product is the collected data (PD), as well as information about the recruited participants (PI).

(d) **Analyzing Data:** When the data is collected, it must be analyzed. Apart from the data (PD), the preregistration (PR), containing the planned analyses, is an input for this activity. The activity produces two outputs: the findings (F), and the analysis specification (AS). While the preregistration contains plans for the analysis, the actual analysis may still differ from it, especially in the case of exploratory studies or if unexpected issues emerge (e.g., parametric analyses are not appropriate so non-parametric analyses are used instead). Researchers may also wish to explore additional research questions or test the robustness of their results by using additional unplanned analyses. Changes to



(a) Creating Assessment Framework (b) Refining Assessment Framework (c) Assessing Secondary Data (d) Cleaning Secondary Data

Figure 5. Provenance patterns for secondary data collection.

analysis are perfectly understandable and often recommended, but it is important there is a clear delineation between pre-planned confirmatory analyses and exploratory analyses. As it only becomes apparent what is actually analyzed - and how it is done precisely - during the activity, the analysis specification (AS) is produced as part of this activity.

Considerable variation exists in how these practices have been adopted by different research fields, subareas, lab groups, and even across different studies by the same researchers. For example, although there are many advantages to preregistration (PR) [32] it is not a mandatory practice and therefore may not always be present within a primary data collection process. Similarly, although the primary data (PD) and analysis specification (AS) entities are always generated from a primary data collection process, these are not always made publicly available or even shared with other researchers upon request (e.g., see [33] about the low response rates of authors to data requests). Nonetheless, the inclusion of as many of these entities as possible within a provenance model greatly increases the ability of researchers, including those who conducted the primary data collection, to assess the robustness, reliability, and relevance of the collected data as well as any findings that were generated. This also has important flow-on effects for subsequent modeling activities that incorporate and rely on the primary data. For example, further assessment, new data collection, and/or new information coming to light (e.g., failed replications [34]), may lead to the reliability and robustness of a primary data collection process being called into question. If this primary data collection has been incorporated within a provenance model(s) then researchers can quickly and easily discover which further processes relied on or built upon the questionable primary data collection. This makes it much easier to discover and reexamine or reassess whether subsequent pieces of work need to also be updated or adjusted in light of the questions raised about a primary data collection process or output.

3.4 Secondary Data Collection

Unlike primary data, secondary data is typically more generic - it does not need to be collected for a specific study. Still, such data can of course be useful (and used) for modeling. However, to make the quality of secondary data apparent, it must be assessed based on relevant criteria for the simulation study. Based on the results of the assessment, the data might then be used as is, or might require cleanup or transformations to address the shortcomings.

For dealing with secondary data, we identified have the following three entity types:

- **Assessment Framework (AF):** The assessment framework defines the criteria of the data assessment, dependent on the specific simulation study. For example, the criteria for our case study were specified in [19]. Therein, a set of criteria is defined (such as fitness for purpose, trustworthiness, level of disaggregation, timeliness, completeness, accuracy, and so on). There are five levels of evaluation for each criterion, ranging from "green" where a desirable criterion is met in full, through "amber" when it is met in part, to "red" where this criterion is not met (see e.g. [35]), in our case in-between ratings (green-amber and amber-red) were also included. Some criteria are general in nature, determining the extent a given source may be useful, whereas others are linked to the bias and variance inherent in the data source, which needs to be considered for the modeling process.

- **Metadata (MD):** Metadata are properties of the dataset in question, including the values of specific evaluation ratings from the Assessment Framework (AF) given to the data source. In the migration study presented above, this meta-information is available in the online Data Inventory on Syrian Migration to Europe, 2011-21[5]. As an example, the meta-information for UNHCR data on asylum registration includes source (UNHCR), a short description, a url, time detail (daily), source type (registrations), topic (destination population of interest), data types (quantitative, process-related and macro-level, i.e. aggregate numbers), as well as seven individual aspects of data assessment, ranging from "green" to "amber", with a "green-amber" rating overall.
- **Cleaned Data (CD):** A product of transforming the initial data (D) taking into account their properties (MD), aimed at creating new variables with desired properties, such as being devoid of explicit bias or having reduced variance. A migration-related example can be: if migrant registration data (D) is known to be under-reported (one of the properties of MD, completeness, is rated "amber", indicating a presence of bias), then CD can include daily rates of change in registrations rather than volume of registrations because the former would be less sensitive to the presence of systematic bias.

The following four patterns for secondary data have been also identified (Figure 5):

(a) **Creating Assessment Framework:** As the assessment framework is specific to the simulation study, its creation is the necessary first part of the assessment process. The connection to the rest of the study is realised by using the research question (RQ) as input. Other inputs (X) may include, but are not restricted to, earlier assessment framework(s) or knowledge about limitations of the data relevant in the field, e.g., about typical problems with migration data. The product is the assessment framework (AF).
(b) **Refining Assessment Framework:** At some point during the study, the existing assessment framework may need refinement, e.g., when the research question has shifted enough that the previously defined criteria no longer fit. This activity uses the previous assessment framework (AF), as well as potentially other sources (see Creating Assessment Framework). It produces a refined assessment framework (AF).
(c) **Assessing Secondary Data:** The assessment of some data is the application of the assessment framework to that data to determine the properties of the data. Hence, the assessment framework (AF) and the data (D) are used by the activity, while the metadata (MD) are produced.
(d) **Cleaning Secondary Data:** The transformation of the data (D) in the light of the data properties (MD) identified during the process of applying the assessment framework (AF), in order to produce cleaned data (CD). The process may involve steps such as removing the identified biases, smoothing data to reduce variance, applying a variable transformation to reduce other issues identified in the assessment process (such as log-transformation for strictly positive variables which exhibit exponential patterns of change), and so on.

There were existing frameworks for comprehensively assessing the different aspects of the quality of migration data according to different criteria, including those relying on the traffic-lights operationalisation (e.g. [35]). The inclusion of data assessment in a provenance model not

[5] https://www.baps-project.eu/inventory/data_inventory

only allows for quality checks and corrections to be formally embedded as a necessary element for secondary data need to undergo as part of the modeling process, but also enables identifying which parts of the model may be affected by potential problems with a particular data source. In the case of migration, where available data sources are notorious for their imperfections (e.g. [26]), this makes the ensuing modeling and analysis explicitly conditional on the information used and data cleaning activities undertaken. It also means that, where needed, uncertainty from the data can be propagated to the model results along the paths of the provenance graph, helping with analysis transparency and with honest reporting of the results and their limitations. Alternatively, the provenance sub-graphs related to data analysis and cleaning (Figure 5) may describe a piece of analysis in its own right, should data-related question be of specific interest to the analysts or the users of a particular data source.

4. Proof of Concept

To demonstrate the approach, we realised a provenance model of the case study project outlined in Section 2. In terms of software, we extended WebProv [8]. It allows for the creation and editing of a provenance graph of simulation studies with a web-based interface. The provenance graph is stored in a graph database (Neo4j), which not only allows for simple and efficient storage but also includes a powerful language for retrieving information from the database. Documents and artifacts referenced in the meta-information are stored online in appropriate repositories, e.g., on GitHub, OSF or Zenodo. Our extended version of WebProv and the provenance graph presented here are available in Zenodo archives at [36] and [37].

4.1 Overview

On the whole, the data for the presented model of migration came from a range of primary and secondary sources [23]. Figure 6 shows an overview of the provenance graph as a whole. Examples of primary data include psychological experiments on human decision making (three shown), later supplemented by ethnographic interviews [21]. Secondary data include administrative or survey-based statistics as well as qualitative information on the known numbers of arrivals, interceptions and fatalities, as well as potentially relevant aspects of migration journeys themselves, such as frequency and modes of communication with others and trust in information sources. A complete listing of secondary sources considered for this modeling exercise, together with their basic meta-information and quality assessment, are available in [38].

Figure 7 shows a part of the provenance graph in greater detail, in particular focusing on an instance of primary data collection: a psychological experiment to elicit subjective probability judgements migrants make based on information they gain from different sources. The full experiment and its results are documented in [20]. While the figure only shows the graph, the interactive user interface displays detailed information about each entity and activity when it is selected (see Figure 8) (often giving a concise description, some key properties, and referencing the document or piece of software represented by an entity). In the activity irq3 a research question is identified (see the pattern *Identifying Research Question* Figure 3h), based on some entities in the "Modeling and Model Analysis" area that are not displayed here. Starting from this question, a psychological experiment was designed (*Design Data Collection*, ddc2), using (ML2; referring to Briñol and Petty [30]) and ML3 (referring to Wintle et al. [31]) as methodology literature (ML) inputs - one satisfying the ML input of the pattern, and the other serving as an optional additional input (X). The results of this activity are a data collection procedure (CP2; referring to the survey[6]), the preregistration (PR2; linking to the preregistration stored on OSF[7]) and the ethical approval (E2; referring to the University of Southampton Ethics Committee, ERGO number 56865). Similarly, cpd2 and ad2 match the patterns *Collecting Primary Data* and *Analyzing Data*.

[6] https://southampton.qualtrics.com/jfe/form/SV_20kQsSP0cyi6o06
[7] https://osf.io/3qrs8

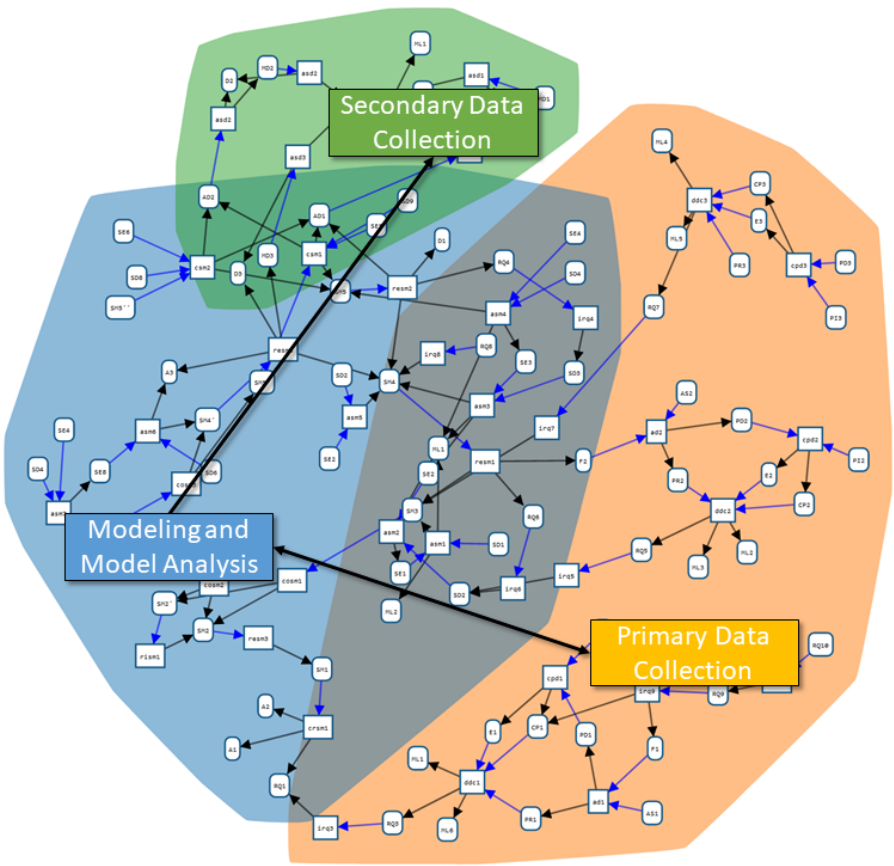


Figure 6. Overview of the case study provenance graph in WebProv. Each node is associated with one part of the project, as distinguished in this paper: the modeling and model analysis (blue), the primary data collection (orange) and the secondary data collection (green).

As demonstrated, the provenance graph can serve as a high-level overview of the various activities of a simulation study, connecting the various inputs and outputs. For large-scale studies with many interconnected parts, the graph will become increasingly large and complex, reflecting the complexity of the documented study. However, the semi-formal structured approach allows for computational processing of the provenance graph, as we demonstrate in the next section.

4.2 Retrieving Detailed Information: Querying

The provenance graph not only gives an overview of the conducted simulation study, but it is also rich in detailed information, linking various artifacts produced in the study. Querying allows for the retrieval of detailed information on demand. Using a dedicated graph database for storing the provenance graph, we can exploit the included querying language, in this case, Neo4J's Cypher, to formulate queries effectively and have them executed efficiently. In practice, retrieving some detail requires two steps: First, a Cypher query must be formulated and executed to retrieve the provenance nodes of interest. Second, the meta-information of the nodes of interest may be inspected, either for the information itself or to follow references to the relevant documents. In the following, we show some typical questions that may be asked of a simulation study about the research questions, the model building, and the relation to data, and demonstrate how they can be answered with queries on the provenance graph.

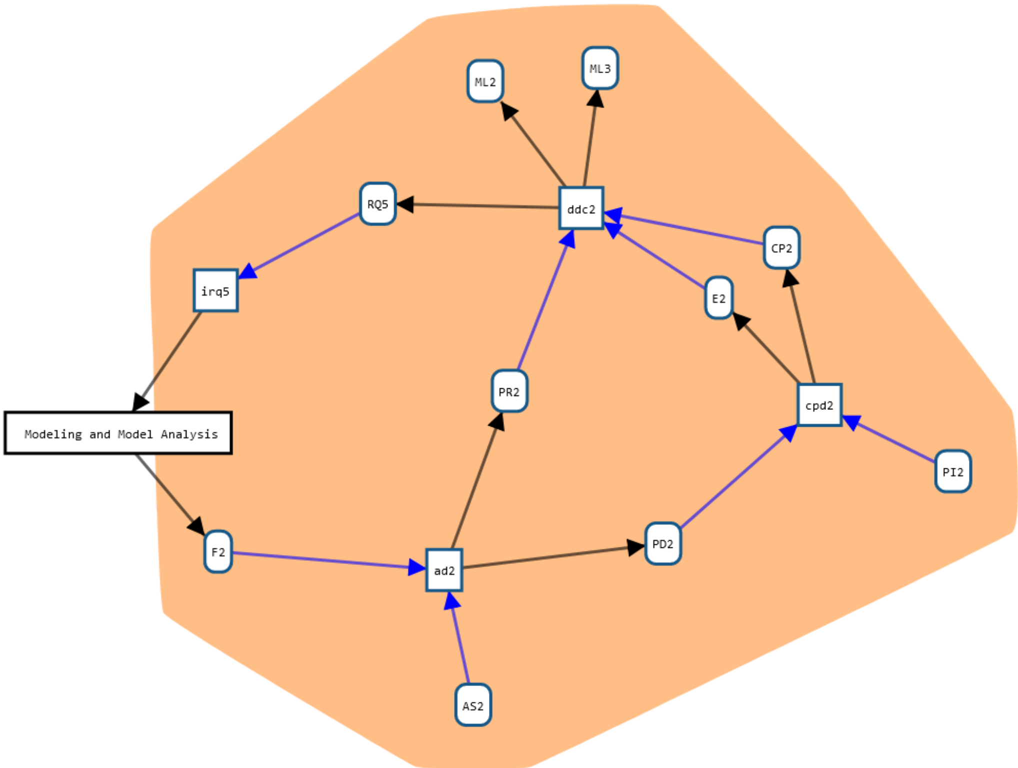


Figure 7. Detail of a part of the primary data collection: the experiment on subjective probability judgements. The box on the left labelled "Modeling and Model Analysis" refers to that part of the project (the blue area in Figure 6). Arrows pointing to or from it represent provenance relations ("used" or "was generated by") with nodes in the "Modeling and Model Analysis" part. In WebProv, this box may be "opened" to display the actual relevant nodes.

Often not only single entities, but their context within the study is of interest – after all, putting the entities into the context of their generation and use is the point of provenance models. For example, we might want to ask for research questions that were newly asked within the study – and what they are based upon. This context can be specified in the query as a graph pattern:

```
MATCH (n {definitionId: 'Research Question'})-[]->(m {definitionId:
   'Identifying Research Question'})-[]->(k:ProvenanceNode)
RETURN n,m,k
```

Here, we query for all Research Question entities `n`, the *Identifying Research Question* activities that generated them `m`, and any entities `k` that were used in these activities. The result is displayed in Figure 9. Please note that the initial research questions of the study are not displayed, as we specifically asked for research questions generated within the study. In this particular example, the query allows for identifying those experimental results (RQ5 "How do migrants make likelihood judgements?") that correspond to the mechanisms underpinning the model assumptions on decision making (RQ6 "How do risk perception and risk avoidance affect the formation of migration routes?"). This enables incorporating experimental findings into the model, along with identifying knowledge gaps that need filling through further data collection.

The same approach also allows for the querying of complex graph patterns, i.e., asking questions about relationships between entities and activities of the simulation study. One might want to know how a certain finding from a psychological experiment, e.g., the findings of a psychological experiment on the subjective judgement of migrants concerning different kinds

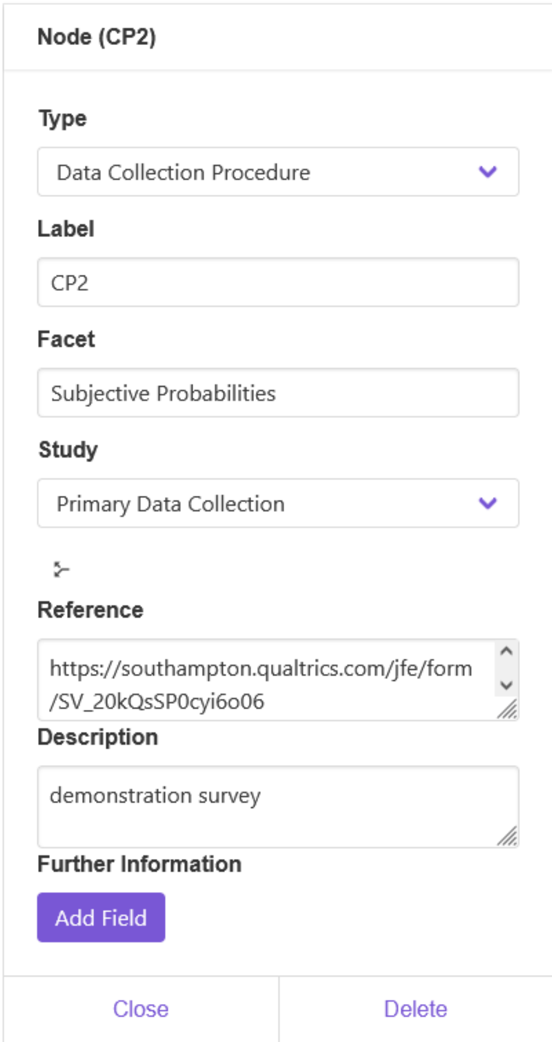


Figure 8. Information about the entity CP2, the data collection procedure of the psychological experiment described in the text, as displayed in WebProv. The field "Facet" allows a keyword to be entered, to aid in searching for related entities, e.g., all entities related to this experiment (see Figure 7) have the facet "Subjective Probabilities". The "Study" field refers to the different colored areas, which are called studies in WebProv. "Reference" contains a reference to the collection procedure itself, in this case in the form of a demonstration survey identical to the one given to the participants. The "Description" field allows additional information to be summarised.

of information and sources (the entity labeled F2), influenced the simulation models. In terms of the provenance graph this means asking for simulation models from which a path (possibly via multiple intermediary steps) leads to F2, as well as for the nodes on this path:

```
MATCH p=shortestPath(({definitionId: 'Simulation
    Model'})-[*]->({label: 'F2'})), (n)
WHERE n IN NODES(p)
RETURN n
```

Here, we use the shortest path, to only see the most direct path from any simulation model, hiding more indirect relations. The result of the query, a sub-graph of the provenance graph,

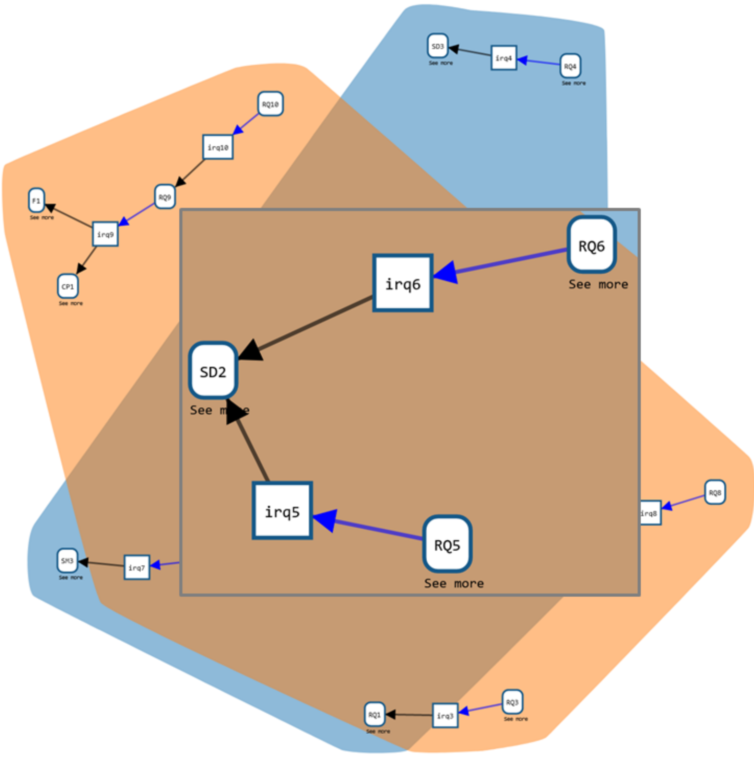


Figure 9. Result of a query for research questions identified in the project. The grey box shows a part of the result: RQ5 ("How do migrants make likelihood judgements?") and RQ6 ("How do risk perception and risk avoidance affect the formation of migration routes?") both follow from SD2, the result of an analysis of the model that shows that the parameters related to risk are most sensitive, requiring more research on this subject, and hence the two new research questions. The background shows the complete result of the query.

can be seen in Figure 10. This shows the empirical findings that informed the building of the simulation model SM4, as well as later model versions via SM4. In this way, the new version, incorporating the experimental results, transformed the model from being a mostly theoretical exercise to becoming grounded in the empirical evidence about decision making, in this case on the perceived risk of making a migration journey and how it varied depending on the information received from various sources [23].

As a final example, one might be interested in how the modeling work was grounded on the other work conducted in the study, e.g., on the collected data. In the query, we are looking for any links from nodes in the "Modeling and Model Analysis" area to other areas of the study. For the sake of clarity, we only want to display the first entity or activity outside of the "Modeling and Model Analysis" area:

```
MATCH (s:Study {label: "Modeling and Model Analysis"}),
    p=shortestPath((n {studyId: s.id})-[*]->(k)), (m)
WHERE k.studyId <> s.id AND exists(({studyId: s.id})-[]->(k)) AND m IN
    NODES(p)
RETURN m
```

These areas are called "study" in WebProv (which is a different use of the term study than elsewhere in this paper). We identify `s` as the WebProv study "Modeling and Model Analysis". Then we search for paths from a node `n` within the WebProv study `s` to other nodes `k` that

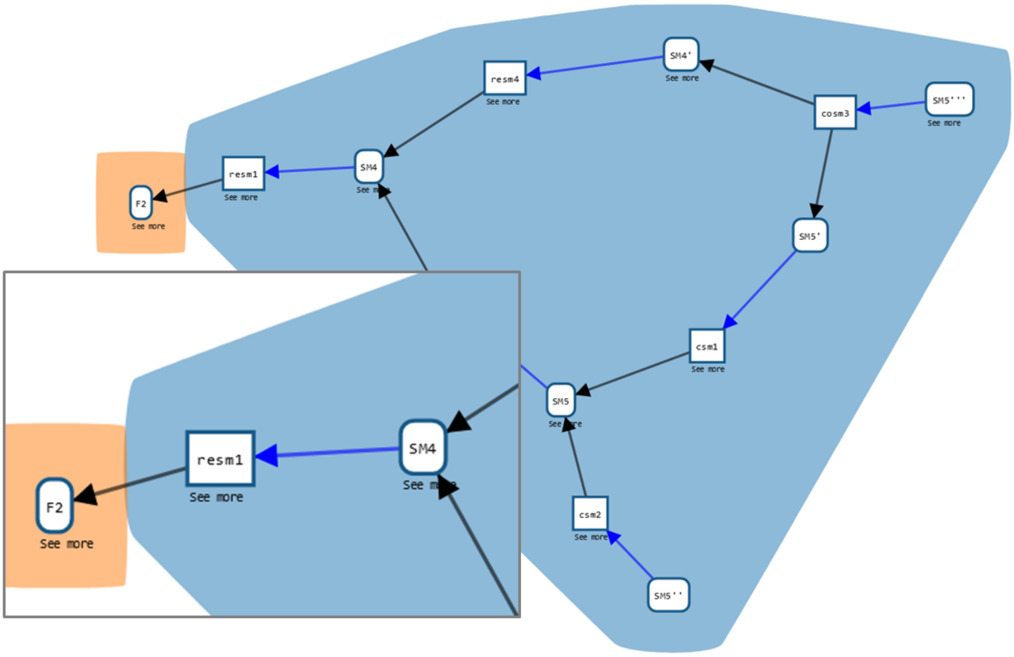


Figure 10. Result of the query for models that use the findings of an experiment on the subjective judgement of migrants on different kinds of information and sources (F2). The findings were used in creating the model version SM4 (see the enlarged box). Through SM4, they also influenced later model versions. If new conflicting findings emerged, the preceding model SM3 could be used to develop an alternative simulation model to SM4. Thus, model families might emerge that share a core but allow for exploring different scenarios based on different findings.

have a different WebProv study id, i.e., are in a different area. Further, these nodes shall have a predecessor with the same WebProv study id as `n`, i.e., which is in the same area as `n`. Figure 11 shows that the modeling work was grounded on the F2 findings from the primary data collection as well as several entities from the secondary data collection, leading to a more developed and fine-tuned version of the model. The provenance query may also be refined, e.g., to ask specifically about the secondary data.

5. Related Work

The wish to reproduce, interpret, and reuse the results of simulation studies has led to various reporting guidelines in different application fields [3,39]. Some focus on specific activities or products of a simulation study. For example, there are guidelines focused on the simulation model, e.g., MMRR (Minimum Model Reporting Requirements), PMRR (Preferred Model Reporting Requirements) for systems dynamics models [3], or ODD for agent-based models [4,28,40]). There are also guidelines for simulation experiments, e.g., MIASE (Minimum Information about a Simulation Experiment) [41], MSRR (Minimum Simulation Reporting Requirements), and PSRR (Preferred Simulation Reporting Requirements) [3]. Others, such as TRACE [1] or STRESS [2] aim at documenting the entire simulation study and thereby covering all of the essential steps, sources, and products of a modeling and simulation life cycle [42,43].

Provenance patterns can and should be combined with the aforementioned documentation guidelines for describing the meta-information or attributes of the entities. Semantic annotation of provenance using ontologies can add domain-specific meta-information, which is crucial when reasoning, e.g., about the cognitive paradigms used in psychological experiments [44]

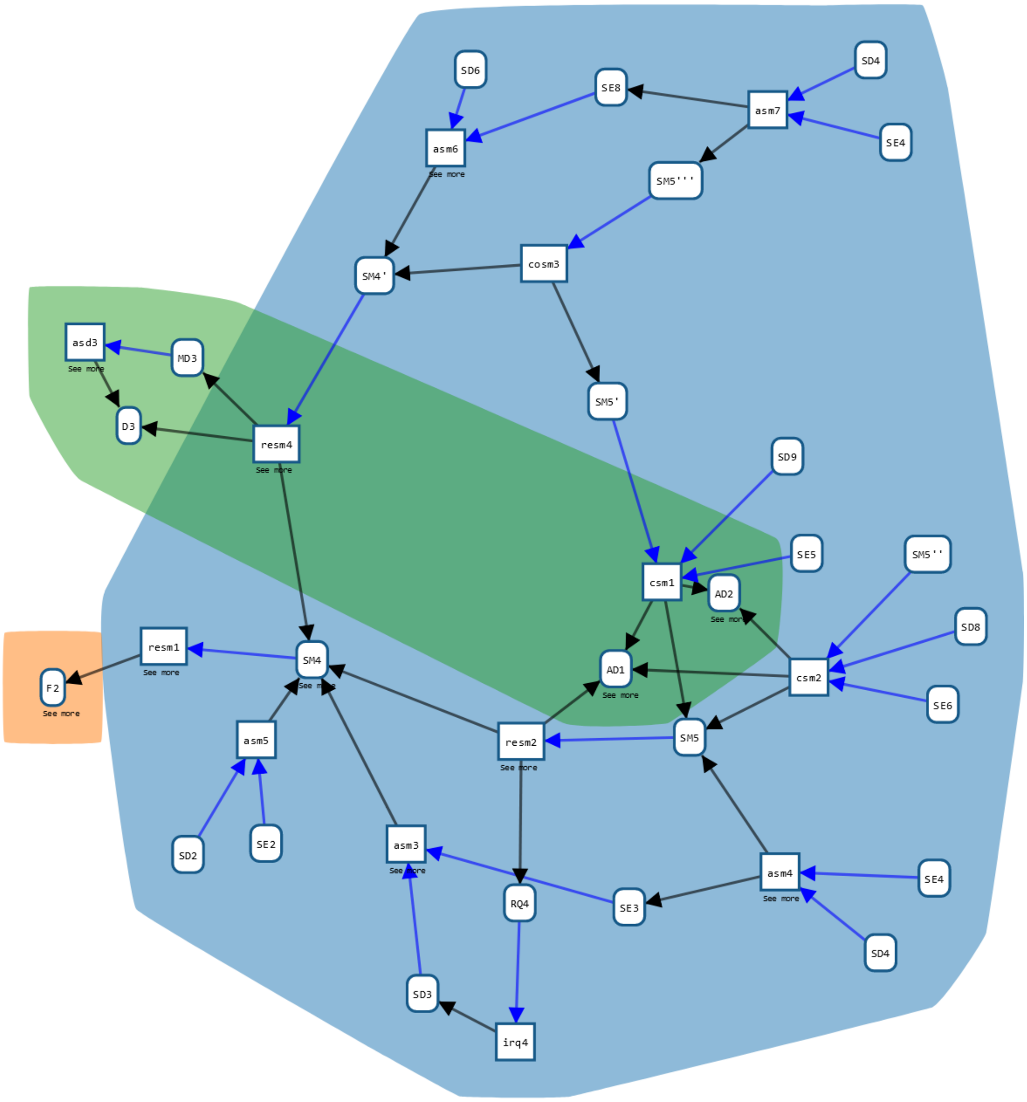


Figure 11. Result query for sources of simulation models outside the "Modeling and Model Analysis" area. The node on the left is F2 (as in Figure 10), the nodes in the green area are from the "Secondary Data Collection" area. The amount of external, different sources might indicate how many research results are brought together by the model and, thus, its integrative quality.

or the modeling methodologies applied [45]. Further ontologies for the social sciences can be inspired by the various ontologies of bioinformatics [46]. In addition, meta-information should be combined with accessible and executable versions of simulation models and simulation experiments in accordance with the FAIR (findable, accessible, interoperable, reusable) open data principles [47]. Thus, each entity of type simulation model or simulation experiment should refer to an openly accessible code repository. The accessibility and transparency of code can be enhanced by exploiting developments and standards in specifying executable simulation experiments, including domain-specific languages [48–50] or model-based experiment designs [51].

Whereas documentations based on the various reporting guidelines assume one purpose and one single simulation model, we assume that different versions of simulation models and even research questions can belong to the documentation of a simulation study. The provenance graph can be seen as a "meta-model" (a term used here in a different sense than in statistics and uncertainty quantification, see [23] for discussion), describing the generation process of a simulation model in terms of activities such as model creation, refinement, and composition, as well as the generation of research questions and their interrelations with sources and (intermediate) products of the simulation study. Our approach's unique perspective on the story of simulation study also becomes evident when we look at activities such as model analysis, calibration, and validation.

Another advantage of our approach concerns data. In all reporting guidelines of simulation studies, information about the used data is required, e.g., in its checklist STRESS asks for details and purpose of data sources, input parameters for base runs of the model and scenario experiments, assumptions, and data pre-processing. The latter refers to any manipulation of the data that occurred. In the TRACE documentation, the data evaluation section should provide insights into the quality and sources of numerical and qualitative data that have been used to parameterise the model.

While the use of data is generally included in these documentation standards and sometimes even put into focus e.g., by the Rigour And Transparency – Reporting Standard (RAT-RS [52]) and ODD+Decision+Data (ODD+2D [53]), the procurement of data, including assessment of data with explicit criteria and data transformation, is usually not. Often, the approaches for documenting and recording information can rely on reporting guidelines for data acquisition and generation in the respective application field. In the social sciences, this may include various types of quantitative and qualitative data, such as data from psychological experiments, interviews, surveys, or official sources. The replication crisis in psychology, and the subsequent focus on uncovering questionable research practices in psychology and empirical research more broadly, led to the development of several suggestions and guidelines for how to document and improve rigour, openness, and transparency in empirical research [34,54]. For primary data collection, these practices include: making collected data and analysis code publicly available, publicly posting the study materials and procedure, being transparent about the ethical aspects, and preregistering study protocols and analysis plans ahead of time [55–57]. Although some of these practices are not directly applicable to secondary data collection and analysis, practices such as sharing analysis code and clearly specifying analysis plans ahead of time are also strongly recommended for improving the transparency and rigour of research relying on secondary data [58,59]. However, there is still a long way to go before these practices become standard.

6. Conclusion

Based on provenance graphs and patterns, the documentation approach presented in this paper differs from existing documentation guidelines for simulation studies such as TRACE or STRESS in three key aspects: scope, subject, and degree of formalization.

Unlike the aforementioned guidelines, we treat primary and secondary data collection as an integral part of the simulation study. To judge the foundations of a simulation model, it is not enough to just know that data were used. The quality and suitability of data must also be assessable. Unlike TRACE [1] or STRESS [2], which are primarily concerned with what data was used and where it is used in the modeling process, the provenance approach presented in this paper makes the collection of primary data and the assessment, preparation, and cleaning of secondary data explicit. The detailed documentation of both collection procedures (for primary data) and assessment criteria also aims to greatly improve the visibility of the limitations of the data.

The provenance graph focuses on documenting a simulation study's processes, not its products. The provenance patterns we suggest do not describe a simulation model, simulation experiment, or piece of data. Instead, they describe how they were created, what steps were undertaken, and how they relate to the specific research questions they were designed to answer, to data they were based upon, and to the results they generated. Consequently, the provenance graph is not intended to replace other documentation but to complement it. The whole graph models the process of the study. Single entities document individual (intermediate) products, for which existing documentation standards such as ODD [28,40] (for an agent-based model) or MIASE [41] (for a simulation experiment) should be employed.

Unlike most documentation standards and protocols used or suggested for social simulation, which are textual, we propose a more structured semi-formal approach. This aims to make the documentation more accessible for computational processing. We demonstrate some of these benefits in Section 4 by using graph queries to retrieve information about the simulation study. For instance, using a single query we could identify what primary or secondary data the different modeling steps were based on. This benefit is not restricted to the typical use cases for documentation that generally only occur after the simulation study has been completed. Wilsdorf at al. [18] demonstrate how the provenance graph can automatically generate new simulation experiments for new model versions while conducting the simulation study. The provenance graph can even be automatically generated during the simulation study, e.g., by exploiting workflow systems to conduct the simulation study [11]. However, this requires the modeler to get acquainted with the respective workflow system.

An automatic and less intrusive solution for automatically capturing provenance information is desirable and as such is the focus of follow-up research. In addition, the insights gained from provenance graphs are constrained by the intelligibility of the query language. Consequently, we plan to equip the web tool with additional capabilities to reduce complex graphs and templates for recurring queries. Lastly, we aim to further integrate the various reporting guidelines and ontologies for explicitly and formally specifying the provenance metadata. This will be essential for deepening the understanding and support of simulation studies and for using formal reasoning to help remedy the various methodological challenges of computational social sciences [60]. Some of these challenges involve the accurate and appropriate use of statistics, data science and other modeling approaches across a range of applications and scientific disciplines. By developing a common, formal standard for documenting, visualising, querying, and analysing different stages of the modeling process, provenance models provide a promising approach to addressing many of these challenges, making simulation studies and processes intelligible to both model producers and users from diverse areas of science and practice.

Data Accessibility

Our extended version of WebProv and the provenance graph presented here are available at `https://doi.org/10.5281/zenodo.6786191` and `https://doi.org/10.5281/zenodo.6786226`.

Acknowledgment

This research was funded by the European Research Council (ERC) via the research project Bayesian Agent-based Population Studies (grant number 725232) and the German Research Foundation (DFG) via the research project Modeling and Simulation of Linked Lives in Demography (grant number 25856074, UH-66/15).

Author contributions[8]: Conceptualization: O.R., A.M.U.; Methodology: O.R., P.W., A.M.U.; Software: O.R.; Investigation: O.R., T.P., J.B., A.M.U.; Data Curation: O.R., T.P, M.H., J.B.; Writing original draft: O.R., T.P., J.B., A.M.U.; Writing review and editing: O.R., T.P., M.H., J.B., P.W., A.M.U.; Visualization: O.R.; Supervision: J.B., A.M.U.; Project administration: O.R., P.W.; Funding Acquisition: J.B., A.M.U

The authors would like to thank the original developers of the tool WebProv, Jacob Smith and Kai Budde.

[8]Contributor Roles Taxonomy (CRediT): `https://credit.niso.org/`

References

1. Grimm V, Augusiak J, Focks A, Frank BM, Gabsi F, Johnston ASA, Liu C, Martin BT, Meli M, Radchuk V, Thorbek P, Railsback SF. 2014 Towards Better Modelling and Decision Support: Documenting Model Development, Testing, and Analysis Using TRACE. *Ecological Modelling* **280**, 129–139.
2. Monks T, Currie CS, Onggo BS, Robinson S, Kunc M, Taylor SJ. 2019 Strengthening the reporting of empirical simulation studies: Introducing the STRESS guidelines. *Journal of Simulation* **13**, 55–67.
3. Rahmandad H, Sterman JD. 2012 Reporting guidelines for simulation-based research in social sciences. *Systems Dynamics Review* **28**, 396–411.
4. Grimm V, Berger U, DeAngelis DL, Polhill JG, Giske J, Railsback SF. 2010 The ODD protocol: a review and first update. *Ecological modelling* **221**, 2760–2768.
5. Bergmann FT, Adams R, Moodie S, Cooper J, Glont M, Golebiewski M, Hucka M, Laibe C, Miller AK, Nickerson DP et al.. 2014 COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC Bioinformatics* **15**, 1–9.
6. Wilsdorf P, Haack F, Uhrmacher AM. 2020 Conceptual Models in Simulation Studies: Making it Explicit. In *Proceedings of the 2020 Winter Simulation Conference* pp. 2353–2360 Piscataway, New Jersey. IEEE.
7. Ayllón D, Railsback SF, Gallagher C, Augusiak J, Baveco H, Berger U, Charles S, Martin R, Focks A, Galic N et al.. 2021 Keeping modelling notebooks with TRACE: Good for you and good for environmental research and management support. *Environmental Modelling & Software* **136**, 104932.
8. Budde K, Smith J, Wilsdorf P, Haack F, Uhrmacher AM. 2021 Relating Simulation Studies by Provenance—Developing a Family of Wnt Signaling Models. *PLOS Computational Biology* **17**, e1009227.
9. Groth P, Moreau L. 2013 PROV-Overview – An Overview of the PROV Family of Documents. Technical Report World Wide Web Consortium.
10. Ruscheinski A, Gjorgevikj D, Dombrowsky M, Budde K, Uhrmacher AM. 2018 Towards a PROV Ontology for Simulation Models. In *Provenance and Annotation of Data and Processes* pp. 192–195. Springer International Publishing.
11. Ruscheinski A, Wilsdorf P, Dombrowsky M, Uhrmacher AM. 2019 Capturing and Reporting Provenance Information of Simulation Studies Based on an Artifact-Based Workflow Approach. In *Proceedings of the 2019 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation* pp. 185–196 New York, NY, USA. Association for Computing Machinery.
12. Haack F, Budde K, Uhrmacher AM. 2020 Exploring the Mechanistic and Temporal Regulation of LRP6 Endocytosis in Canonical WNT Signaling. *Journal of Cell Science* **133**, jcs243675.
13. Reinhardt O, Ruscheinski A, Uhrmacher AM. 2018 ODD+P: Complementing the ODD Protocol With Provenance Information. In Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, editors, *Proceedings of the 2018 Winter Simulation Conference* pp. 727–738 Piscataway, New Jersey. IEEE.
14. Bijak J, Courgeau D, Franck R, Silverman E. 2018 Modelling in Demography: From Statistics to Simulations. In *Methodological Investigations in Agent-Based Modelling* , pp. 167–187. Springer.
15. Courgeau D, Franck R. 2007 Demography, a Fully Formed Science or a Science in the Making? An Outline Programme. *Population* **62**, 39–45.
16. Burch TK. 2018 *Model-Based Demography: Essays on Integrating Data, Technique and Theory*. Demographic Research Monographs. Cham: Springer Nature.
17. Conte R, Gilbert N, Bonelli G, Cioffi-Revilla C, Deffuant G, Kertesz J, Loreto V, Moat S, Nadal JP, Sanchez A, Nowak A, Flache A, San Miguel M, Helbing D. 2012 Manifesto of Computational Social Science. *The European Physical Journal Special Topics* **214**, 325–346.
18. Wilsdorf P, Wolpers A, Hilton J, Haack F, Uhrmacher AM. 2022 Automatic Reuse, Adaption, and Execution of Simulation Experiments via Provenance Patterns. *ACM Trans. Model. Comput. Simul.* Just Accepted.
19. Nurse S, Bijak J. 2019 Meta-Information on Data Sources on Syrian Migration into Europe. Technical Report University of Southampton. https://www.southampton.ac.uk/baps/inventory/data-sources.page.

20. Prike T, Bijak J, Higham PA, Hilton J. 2022 How Safe Is This Trip? Judging Personal Safety in a Pandemic Based on Information from Different Sources.. *Journal of Experimental Psychology: Applied* **28**, 509–524.
21. Belabbas S, Bijak J, Modirrousta-Galian A, Nurse S. 2022 From Conflict Zones to Europe: Syrian and Afghan Refugees' Journeys, Stories, and Strategies. *Social Inclusion* **10**, 211–221.
22. Bijak J, Czaika M. 2020 Assessing Uncertain Migration Futures – A Typology of the Unknown. Quantmig project deliverable d1.1 University of Southampton and Danube University Krems. Available: https://www.quantmig.eu.
23. Bijak J. 2021 *Towards Bayesian Model-Based Demography* vol. 17*Methodos Series*. Cham: Springer.
24. Hinsch M, Bijak J. 2019 Rumours Lead to Self-Organized Migration Routes. In *The 2019 Conference on Artificial Life: How Can Artificial Life Help Solve Societal Challenges? (29/07/19 - 02/08/19)*.
25. Cioffi-Revilla C. 2010 A Methodology for Complex Social Simulations. *Journal of Artificial Societies and Social Simulation* **13**, 7.
26. Poulain M, Perrin N, Singleton A, editors. 2006 *THESIM: Towards Harmonised European Statistics on International Migration*. Louvain-la-Neuve: Presses Universitaires de Louvain.
27. Prike T, Higham PA, Bijak J. 2021 The Boundaries of Cognition and Decision Making. In Bijak J, editor, *Towards Bayesian Model-Based Demography: Agency, Complexity and Uncertainty in Migration Studies* , Methodos Series pp. 93–112. Cham: Springer International Publishing.
28. Grimm V, Railsback SF, Vincenot CE, Berger U, Gallagher C, DeAngelis DL, Edmonds B, Ge J, Giske J, Groeneveld J et al.. 2020 The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation* **23**.
29. Reinhardt O, Uhrmacher AM, Hinsch M, Bijak J. 2019 Developing Agent-Based Migration Models in Pairs. In *Proceedings of the 2019 Winter Simulation Conference* pp. 2713–2724 Piscataway, New Jersey. IEEE.
30. Briñol P, Petty RE. 2009 Source Factors in Persuasion: A Self-Validation Approach. *European Review of Social Psychology* **20**, 49–96.
31. Wintle BC, Fraser H, Wills BC, Nicholson AE, Fidler F. 2019 Verbal Probabilities: Very Likely to Be Somewhat More Confusing than Numbers. *PLOS ONE* **14**, e0213522.
32. Sarafoglou A, Kovacs M, Bakos B, Wagenmakers EJ, Aczel B. 2022 A survey on how preregistration affects the research workflow: better science but more work. *Royal Society Open Science* **9**, 211997.
33. Gabelica M, Bojčić R, Puljak L. 2022 Many researchers were not compliant with their published data sharing statement: mixed-methods study. *Journal of Clinical Epidemiology*.
34. Open Science Collaboration. 2015 Estimating the Reproducibility of Psychological Science. *Science* **349**, aac4716.
35. Vogel D, Kovacheva V. 2008 Classification report: Quality assessment of estimates on stocks of irregular migrants. Technical report HWWI Hamburg Institute of International Economics.
36. Reinhardt O. 2022a WebProv for "Simulation Studies of Social Systems - Telling the Story Based on Provenance". https://doi.org/10.5281/zenodo.6786191.
37. Reinhardt O. 2022b Provenance Data for "Simulation Studies of Social Systems - Telling the Story Based on Provenance". https://doi.org/10.5281/zenodo.6786226.
38. Nurse S, Bijak J. 2021 Syrian Migration to Europe, 2011-21: Data Inventory. Online resource University of Southampton. https://baps-project.eu/inventory/data_inventory,.
39. Erdemir A, Guess TM, Halloran J, Tadepalli SC, Morrison TM. 2012 Considerations for reporting finite element analysis studies in biomechanics. *Journal of Biomechanics* **45**, 625–633.
40. Grimm V, Polhill G, Touza J. 2017 Documenting social simulation models: the ODD protocol as a standard. In *Simulating social complexity* , pp. 349–365. Springer.
41. Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R, Chelliah V, Cooling MT, Cooper J, Crampin EJ et al.. 2011 Minimum information about a simulation experiment (MIASE). *PLoS computational biology* **7**, e1001122.

42. Balci O. 2012 A life cycle for modeling and simulation. *Simulation* **88**, 870–883.
43. Robinson S. 2014 *Simulation: The Practice of Model Development and Use*. MacMillan 2nd edition.
44. Turner JA, Laird AR. 2012 The cognitive paradigm ontology: design and application. *Neuroinformatics* **10**, 57–66.
45. Silver GA, Miller JA, Hybinette M, Baramidze G, York WS. 2011 DeMO: An ontology for discrete-event modeling and simulation. *Simulation* **87**, 747–773.
46. Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, Dumontier M, Finney A, Golebiewski M, Hastings J et al.. 2011 Controlled vocabularies and semantics in systems biology. *Molecular systems biology* **7**, 543.
47. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE et al.. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3**, 1–9.
48. Warnke T, Uhrmacher AM. 2018 Complex Simulation Experiments Made Easy. In *Proceedings of the 2018 Winter Simulation Conference* pp. 410–424 Piscataway, New Jersey. IEEE.
49. Salecker J, Sciaini M, Meyer KM, Wiegand K. 2019 The nlrx r package: A next-generation framework for reproducible NetLogo model analyses. *Methods in Ecology and Evolution* **10**, 1854–1863.
50. Waltemath D, Adams R, Bergmann FT, Hucka M, Kolpakov F, Miller AK, Moraru II, Nickerson D, Sahle S, Snoep JL et al.. 2011 Reproducible computational biology experiments with SED-ML-the simulation experiment description markup language. *BMC systems biology* **5**, 1–10.
51. Wilsdorf P, Heller J, Budde K, Zimmermann J, Warnke T, Haubelt C, Timmermann D, van Rienen U, Uhrmacher AM. 2022 A Model-Driven Approach for Conducting Simulation Experiments. *Applied Sciences* **12**.
52. Achter S, Borit M, Chattoe-Brown E, Siebers PO. 2022 RAT-RS: A Reporting Standard for Improving the Documentation of Data Use in Agent-Based Modelling. *International Journal of Social Research Methodology* **0**, 1–24.
53. Laatabi A, Marilleau N, Nguyen-Huu T, Hbid H, Ait Babram M. 2018 ODD+2D: An ODD Based Protocol for Mapping Data to Empirical ABMs. *Journal of Artificial Societies and Social Simulation* **21**, 9.
54. Simmons JP, Nelson LD, Simonsohn U. 2011 False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* **22**, 1359–1366.
55. Christensen G, Wang Z, Paluck EL, Swanson N, Birke DJ, Miguel E, Littman R. 2019 Open Science Practices are on the Rise: The State of Social Science (3S) Survey. .
56. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA. 2017 A manifesto for reproducible science. *Nature Human Behaviour* **1**, 0021.
57. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon TA, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T. 2015 Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility.. *Science* **348**, 1422–1425.
58. Miłkowski M, Hensel WM, Hohol M. 2018 Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience* **45**, 163–172.
59. Stodden V, Guo P, Ma Z. 2013 Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLOS ONE* **8**, 1–8. Publisher: Public Library of Science.
60. Hox JJ. 2017 Computational Social Science Methodology, Anyone?. *Methodology* **13**, 3–12.