Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

MODELLING MIGRATION: DECISIONS, PROCESSES AND OUTCOMES

Jakub Bijak
Philip A. Higham
Jason Hilton
Martin Hinsch
Sarah Nurse
Toby Prike
Peter W.F. Smith

Social Statistics and Demography, Psychology
University of Southampton
Highfield campus, University Road
Southampton, SO17 1BJ, UK

Oliver Reinhardt
Adelinde M. Uhrmacher

Institute of Visual and Analytic Computing
University of Rostock
Albert-Einstein-Straße 22
D-18059 Rostock, GERMANY

ABSTRACT

Human migration is uncertain and complex, and some of its distinct features, such as migration routes, can emerge and change very rapidly. Agency of various actors is one key reason for why migration eludes attempts at its theoretical description, explanation and prediction. To address the complexity challenges through simulation models, which would coherently link micro-level decisions with macro-level processes, a coherent model design and construction process is needed. Here, we present such a process alongside its five building blocks: an agent-based simulation of migration route formation, resembling the recent asylum migration to Europe; an evaluation framework for migration data; psychological experiments eliciting decisions under uncertainty; the choice of a programming language and modelling formalisms; and statistical analysis with Bayesian meta-modelling based on Gaussian Process assumptions and experimental design principles. This process allows to identify knowledge advancements that can be achieved through modelling, and to elucidate the remaining knowledge gaps.

1 INTRODUCTION

Uncertainty and complexity are some of the key defining features of human migration - one of the main global challenges of today's world. In particular, this holds true for the formation or change of migration routes, or responses of flows to the underlying drivers, which are characterised by very high volatility, as witnessed during the 2015–16 Syrian asylum crisis (Kingsley 2016). One reason behind this complexity, and behind the inefficiency of attempts to control migration, is the agency of various actors involved – migrants, institutions, intermediaries, and so on (Castles 2004). This agency also remains one fundamental reason why migration typically eludes attempts at its theoretical description, explanation and prediction, the efforts undertaken having historically remained scattered across various disciplines (Arango 2000).

Simulation models can help understand and address some of these challenges posed by migration, by connecting the micro-level individual decisions with the processes observed at the macro-level. Still, for simulations to fulfill their promise, a unified and interdisciplinary model-building process is needed that would formally reflect this underlying agency, and describe the different aspects of migration in a coherent way. In this paper, we present five building blocks of such a process, grounded in the methodology of agent-based modelling. These building blocks include: construction of an agent-based simulation model of migration, in this case illustrating the formation of migrant routes; a unified framework for assessment of the existing data and their quality; results of psychological experiments on human decision making under uncertainty; the choice of an appropriate programming language and modelling formalisms; and statistical (Bayesian) meta-modelling, using Gaussian Processes, which enables the analysis of the uncertainty of the model outputs and their sensitivity to various parameters. With this approach, we build upon and extend the existing literature on iterative simulation model-building *process* (see, e.g. Kleijnen and van Beers 2004; Sargent 2013; Cottineau et al. 2015), by applying it to an agent-based model involving human decisions, and formally incorporating insights from several relevant scientific disciplines.

In this paper, the five building blocks of the proposed approach are presented first, in Section 2, including a preliminary analysis of the model-building process for an agent-based model of migrant route formation. The process itself, the aim of which is ultimately to bring these building blocks together, is outlined in Section 3, where we discuss and evaluate the construction of increasingly sophisticated models. A formalisation of process description by using a provenance model is offered in Section 4. The conclusions and lessons learned from the iterative model-building exercise, with a focus on knowledge advancements that can be achieved through modelling, and the remaining knowledge gaps are discussed in Section 5.

2 BUILDING BLOCKS OF THE MODELLING APPROACH

2.1 Agent-Based Model of Migration Routes

At the core of the proposed approach is an agent-based model. We start from the question of how the migrant routes, clearly observed for real migration processes, are formed and sustained. Migrants attempting to reach a safe destination often have to make their navigation decisions based on very limited information which is to a large degree sourced from other migrants that have made the journey before (see, e.g. Kingsley 2016). Therefore, communication between migrants could be a key factor in determining the dynamics of flows, especially in rapidly-changing processes such as Syrian asylum migration (Dekker et al. 2018).

In this example, we study the effect of information transfer on variability and optimality of migration routes by using an agent-based model with explicit representation of geography, resources and agents' knowledge, where the agents need to navigate a partially unknown landscape (see Figure 1). The first instantiation of the model is described in more detail in Hinsch and Bijak (2019) and the key innovations towards the state of the art (see Klabunde and Willekens 2016 for a review) include an explicit modelling of agent knowledge, their social networks and information exchange.

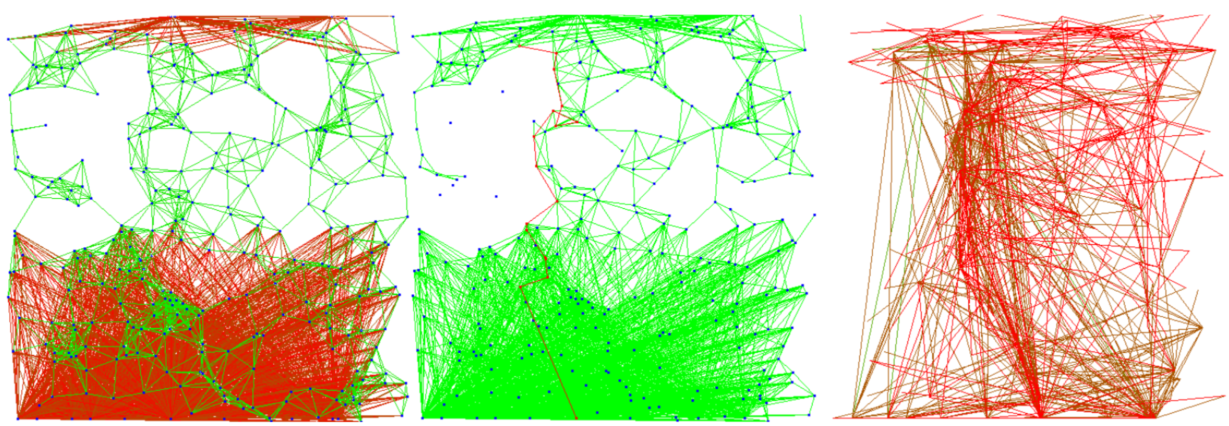

Figure 1: Sample screenshot from the model monitoring: the simulated world with cities and routes (left), a random agent's perception of the world and their route plan (centre) and their social network (right).

From the first version of the model we found that unless agents very quickly acquire objective information from the environment, a higher degree of social information exchange leads to less predictable and less optimal migration routes. This indicates that if a high proportion of information is socially received, routes result from self-organization rather than optimization. We suggest that similar effects should occur in all situations where individuals have to make complex decisions with limited information in a social context.

2.2 Data Sources and Their Quality Evaluation

Migration data are notorious for their problems with completeness, quality and comparability (Poulain et al. 2006), problems which are only exacerbated in the case of asylum migration (Singleton 2016). To make the various aspects of data quality explicit and useful for inclusion in modelling, a common and transparent framework for assessment of different data sources is required (for an example, see Vogel and Kovacheva 2008), which would enable data to be supplemented with meta-information on quality.

Specifically, we have proposed such a unified framework of analysis for recent Syrian asylum migration to Europe (Nurse and Bijak 2019). The framework is comprised of eight dimensions: purpose of collection; timeliness of data; trustworthiness; detailed disaggregation; definition of population under study and associated definitions; transparency of the sources; their completeness; as well as sample design (for surveys). Figure 2 presents an example of the application of this framework to the main process data coming from the daily UNHCR registrations of Syrian asylum seekers in five key destination countries.

01	**UNHCR operational portal**			**Topic:** Destination population	
Source type: Registration	Quantitative Qualitative	Process Context	Macro-level Micro-level	**Time detail:** Daily	**Geography:** 5 countries
Content description: Total cumulative daily numbers of registered Syrian refugees and asylum seekers					
Link: https://data2.unhcr.org/en/situations/syria			**Access:** Data series publicly available for download		
Purpose	Timeliness	Trustworthiness	Disaggregation	**Summary rating:** Green/amber	
Population and definitions	Transparency	Completeness	Sample design N/A		

Figure 2: Summary quality indicators for an example source of data on Syrian migration – UNHCR registrations. The coloured shading corresponds to a traffic-lights assessment of individual quality criteria, ranging from green (good quality), to amber (acceptable), to red (problematic) (Nurse and Bijak 2019).

Each of the quality dimensions, as well as a global quality score for a given data source, is classified into one of three categories: green, amber or red (with two in-between classes, green-amber and amber-red), reflecting the various aspects of variation and bias inherent in the source, which need to be adequately described in any modelling endeavour using the data, ideally by using probability distributions. As a general rule of thumb, we posit that in the agent-based modelling exercise, contextual data, as well as data on micro-level drivers of migration processes should be used as model input, whereas macro-level information on the features of these processes can be used to benchmark model output.

2.3 Eliciting Decisions under Uncertainty

Utility functions for financial decisions have been widely studied. However, it is less clear how these findings generalise to decisions in other domains, such as migration. Developing a better understanding of decision-making is crucial for improving our knowledge of migration processes, and yet previous work has often assumed that migrants simply maximise their utility, which is reflected in how decisions are typically described in existing agent-based models of migration (Klabunde and Willekens 2016). This assumption is called into question by psychological research on decision-making, such as prospect theory (Kahneman and Tversky 1979), and has not previously been empirically tested (although see Czaika 2015).

In our work, we conducted two pre-registered experiments, with 130 lab-based and 403 online participants respectively, who were tasked to choose between gambles presented in either a migration or financial context. A financial context is commonly used for studying risky decision-making and allowed our findings to be compared with previous research. We elicited non-parametric utility functions following Abdellaoui et al. (2016) and tested whether they differed depending on the context. Loss aversion was calculated based on the inflection point of the utility function at the reference value, as well as by regressing the points of the utility function elicited for gains on those elicited for losses. Figure 3 presents the median points of the utility functions elicited for each context in the two experiments.

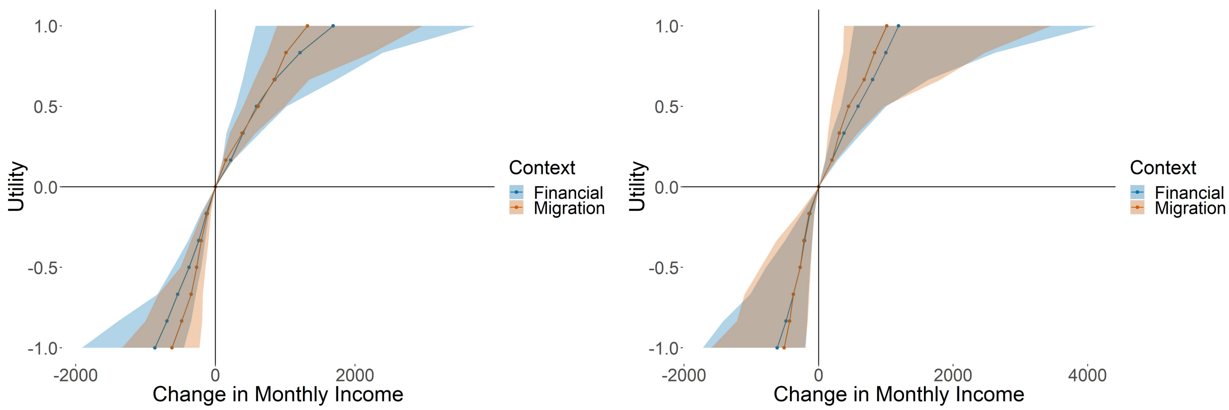

Figure 3: Medians and inter-quartile ranges for the estimated utility curves in the lab-based experiments (left) and online experiments, carried out on Amazon Mechanical Turk (right).

Overall, the results showed many similarities in risky decision-making across the financial and migration contexts. There was evidence of loss aversion in both contexts, and in both contexts the shape of the utility function was generally more consistent with prospect theory than with the commonly assumed utility maximisation. We also found some differences between the contexts, particularly in relation to losses, although this result proved somewhat sensitive to the analytical strategy. The results highlight the need to examine migration decision-making specifically, rather than simply relying on assumptions from other areas without testing them in a migration context.

2.4 Programming Languages and Formalisms

A wide variety of approaches for agent-based modelling exist. Many models are implemented from scratch in a general purpose programming language. A range of agent-based modelling frameworks, either stand-alone, such as NetLogo (Wilensky 1999), or designed as libraries in general-purpose languages, such as Mesa (Masad and Kazil 2015), specifically offer support for the usual difficulties presented by the construction and analysis of agent-based models. In addition, dedicated formal modelling languages such as Carma (Bortolussi et al. 2015) or Chromar (Honorato-Zimmer et al. 2019), offer modellers efficient solutions for implementing the models. In these languages, the model implementation does not just involve the program code, but also includes a higher-level description of a precisely defined stochastic process.

As any translation of an informal idea or description of a model into a concrete implementation – be it in the program code or any other formalism – is always ambiguous, particular care is necessary. Seemingly innocuous decisions made by the modellers may have important downstream effects for the simulation results. This reinforces the need for the highest levels of transparency of model description at all stages of the model development process. The use of a formal modelling language with clear and unambiguous semantics can be very helpful in that respect.

On the other hand, the choice of an implementation language or formalism is far from unambiguous. Such decisions, usually made early in the modelling process, can shape the model significantly by setting up a framework of possibilities and restrictions. In particular, the granularity of time and formalisation of the stochastic processes governing events in the model – continuous or discrete – is usually deeply engrained in the formalism, and might influence the results obtained. Moreover, many dedicated agent-based modelling languages tend to be restricted in their expressiveness, which might hinder the implementation of complex processes involving human decisions and social interactions, such as those presented in our example.

For these reasons, in the work presented in this paper, we elucidated the differences between alternative implementation approaches through which the model could be realised by implementing it in parallel, first as a time-stepped model in a general-purpose programming language (Julia), and second in a domain-specific language specifically developed for demographic agent-based modelling applications with continuous-time semantics (ML3; see Warnke et al. 2017). This approach of implementing our model twice in two fundamentally different languages has led us to questioning many details of the model's implementation. It has also provided us with valuable insights regarding the robustness of the the results in the face of changes to these details and changes to the fundamental implementation paradigm enforced by the language. In addition, parallel model realisations helped us identify crucial features and trade-offs for the future implementation of the model (see Section 3), and also provided important guidelines for the further development of the domain-specific language (Reinhardt et al. 2019).

2.5 Statistical Analysis of Model Properties

As argued in Reinhardt et al. (2018a), the existing methods for analysing the results of agent-based models have so far been largely descriptive and did not use the possibilities offered by contemporary statistical theory and principles of experimental design. To fill this gap, we apply the methods of Bayesian meta-modelling based on Gaussian Process assumptions (Oakley and O'Hagan 2004). This permits us to analyse the uncertainty of model outputs and their sensitivity to a range of input parameters, and to calibrate the model to the extent allowed by the available data.

Formally, let $f(\mathbf{x})$ denote the outputs of the computational agent-based model of interest, based on parameters (inputs) $\mathbf{x}$. The Gaussian Process emulator then assumes a multivariate Normal distribution for realisations of model outputs f (Oakley and O'Hagan 2004):

$$f(\cdot)|\theta, \sigma^2, \mathbf{R} \sim \mathscr{MVN}(\mathbf{m}(\cdot), \sigma^2\mathbf{c}(\cdot,\cdot)). \tag{1}$$

In equation (1), the term $\mathbf{m}(\cdot)$ represents the mean of a linear regression of $f(\cdot)$ on $\mathbf{x}$, with regression (hyper)parameters θ, for example $\mathbf{m}(\mathbf{x}) = h(\mathbf{x})'\theta$ for some known function $h(\cdot)$ of the inputs. Further, σ^2 is the joint variance parameter across the inputs, whereas $\mathbf{c}(\cdot,\cdot)$ is their correlation matrix, the parameterisation of which involves a *roughness matrix*, $\mathbf{R} = diag(r_1, \ldots, r_n)$, which indicates how strongly the emulator responds to particular inputs, and which can include a separate *nugget* term to represent code uncertainty.

In the first phase of the analysis, we have considered four model outputs. One was related to agent behaviour: the proportion of time the agents were following their route plan (*mean_freq_plan*); one to route concentration: the standard deviation of the number of visits over all links (*stdd_link_c*); one to route optimality: the correlation of the number of passages over links with the optimal scenario (*corr_opt_links*); and one to replicability: standard deviation of traffic between replicate runs (*prop_stdd*). For these four outputs, a global, variance-based sensitivity analysis was performed (Oakley and O'Hagan 2004; Saltelli et al. 2008; ten Broeke et al. 2016), to identify relative contributions of different parameters to the output variation. The initial set of 17 parameters was reduced to seven by the means of the preliminary analysis utilising the Definitive Screening Design (Jones and Nachtsheim 2011): five related to information exchange, and two to exploration of the environment by the agents.

Global sensitivity analysis can help identify the input parameters that matter for a given output – in our case, seven parameters related to information exchange, information errors, social networks and exploration (see Table 1). The analysis was performed by using a Latin Hypercube Sample design with a Gaussian

Process emulator. Figure 4 shows an example analysis of an estimated response surface for one selected output, *mean_freq_plan*. Both the estimation and sensitivity analysis have been carried out by using the GEM-SA package (Kennedy and Petropoulos 2016). The substantive conclusion from the preliminary stage of the analysis is clear: in the model of route formation, *information exchange* matters the most.

Table 1: Global sensitivity analysis: percentage of output variance attributed to the key individual inputs and their combinations, under uniform prior distributions for the parameters.

Input: Probability of	*mean_freq_plan*	*stdd_link_c*	*corr_opt_links*	*prop_stdd*
Losing contacts	0.48	5.78	1.07	5.29
Communication with local agents	9.68	5.86	8.52	11.34
Communication with contacts	9.56	0.31	3.75	2.82
Information exchange	66.78	16.11	39.06	21.63
Information error	0.11	24.95	17.62	7.61
Exploration	0.59	3.64	2.99	4.14
Interactions	9.55	28.14	19.80	39.60
Residual	3.25	15.20	7.19	7.58
Total explained	96.75	84.80	92.81	92.42

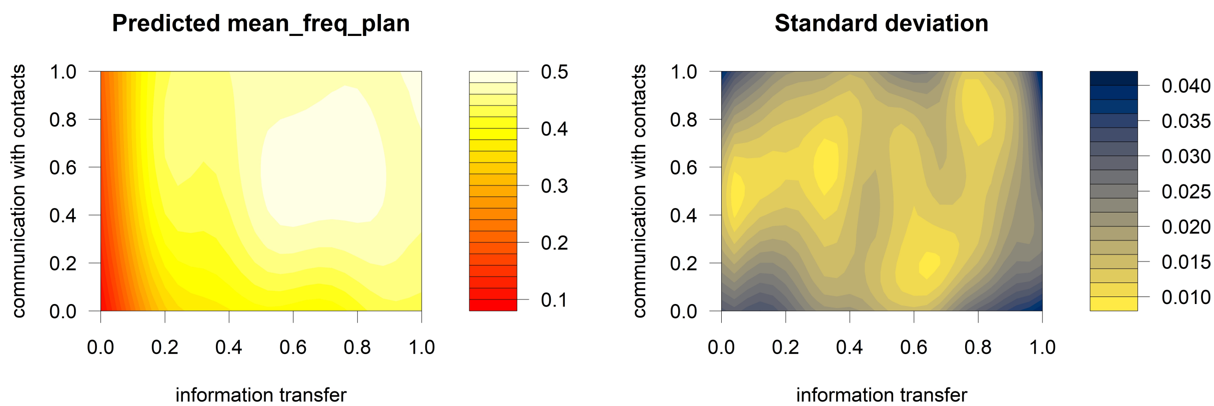

Figure 4: Estimated response surface of the proportion of time the agents follow a plan vs two input parameters, probabilities of information transfer (horizontal axis) and of communication with contacts (vertical axis): Gaussian Process estimates of the mean proportion (left) and its standard deviation (right).

3 MODEL DEVELOPMENT PROCESS

This section outlines the iterative modelling process, aimed at building an increasingly more sophisticated, useful, scientifically justifiable and valid (cf. Sargent 2013) model of migrant route formation. The aim of the process is to ultimately include the five building blocks discussed above, and to do it in such a way, that all the interim decisions, additions and simplifications and other changes, could be formally described and reflected on, aiding further development as well as substantive analysis.

In the case of our migrant route formation model, the development has been carried out in several stages. The initial, preparatory work was based on the scientific literature as well as non-scientific reports such as books, newspaper articles and other accounts, which could shed light on the features of the migration processes (Kingsley 2016). In particular, we were interested in data and other information on and description of the microscopic processes making up migration, such as individual decisions and behaviour, migration conditions and drivers. Based on this literature review we were able to draw a number of conclusions:

- Migrants tend to possess relatively little information about their destination as well as the areas they are travelling through (Kingsley 2016). Moreover, migrants from a number of more authoritarian countries tend to distrust official sources of information (Borkert et al. 2018).
- In general, it appears that peers are the main source of information on destination as well as travel for many migrants. This information is transferred either directly, for example by phone, or indirectly via social media (Borkert et al. 2018).
- Nevertheless, clear migration routes emerge as a result of an interplay between the micro-level decisions of migrants and macro-level factors, such as policies. These routes can adjust to the changing political circumstances, such as border closures (Kingsley 2016; Crawley et al. 2016).

As a first testable hypothesis, these observations suggest that migration routes emerge as a consequence of social interactions between migrants, who make decisions constrained by a certain socio-political environment. Furthermore, we could conclude that if this was indeed the case then there was a possibility that due to the little amount of external information entering the system the resulting routes occur by chance and are not necessarily the best routes possible. In this case migration routes would not be easily predictable, which would have profound political and humanitarian consequences. To verify this, we set out to test whether the behaviour and interactions of migrants could lead to sub-optimal, self-organised routes, by using a detailed agent-based model. The model construction process was iterative, starting from a simple version, which has been subsequently expanded to include increasingly more detail, based on the framework suggested by Brenner and Werker (2009) and Courgeau et al. (2017).

3.1 First Iteration

Conceptually, the first version of the model (unpublished) was based on an initial discussion and evaluation of potential factors and drivers that could plausibly affect mobility decisions of migrants. In the resulting model, the entire area migrants had to travel through was mapped as a finely grained grid, where each cell was characterised by a number of properties such as availability of different resources (including shelter and information) and ease of travel. Settlements and transport links were only distinguished from other grid cells by specific values of these properties, such as high ease of travel for transport or high availability of resources for cities. The agents' information was modelled as a collection of items, each representing information on a single grid cell. An agent had an estimate for the value of each property of each item it knew about together with a confidence value. Information exchange between agents worked as taking the weighted average of value and confidence (for similar models see Lorenz 2007).

We used this version of the model to run a number of exploratory simulation experiments. These showed clear indications of the emergence of self-organized migration routes. However, it also became apparent very quickly that our approach had some technical and conceptual flaws that called for an overhaul of the design, before any meaningful analysis could be carried out on the results.

We were able to identify three main problem areas – the grid-based landscape model, the detailed location properties, and the way information exchange was implemented. Modelling the landscape as a grid structure provided for a very generic way to map geographical topologies into the modelled world and allowed for fine-grained migration routes. At the same time, however, it lead to prohibitively long simulation times and excessive memory consumption, as agents potentially could keep track of many thousands of information items that had to be compared during each instance of information exchange.

Furthermore, empirical information suggested that information exchange between individuals as well as their orientation happens on a more macroscopic scale, involving locations and connections between them (Madl et al. 2015). Implementing that type of information structure on top of a grid would have been difficult to do efficiently. Additionally, we found that it was not useful to model a large number of different landscape properties as they would mostly have similar effects and very little empirical information on their values would be available. Finally, while the belief dynamics in the first iteration of the model have been used in a similar form in previous studies (Lorenz 2007), their behaviour exhibits a number of undesirable properties. For example, a confident agent interacting with a less-confident one will always lose confidence irrespective of the similarity in their beliefs. Similarly, two equally confident agents will experience no change in confidence after interacting, even if their beliefs are diametrically opposed.

3.2 Second Iteration

In the second version of the model (Hinsch and Bijak 2019) we adressed all three issues listed above. We replaced the grid-based landscape structure with a graph-based model, with settlements as vertices and transport links as edges. At the same time we removed most of the properties of grid cells, keeping only generic *quality* and *resource availability* parameters for settlements, as well as *friction* for transport links. Together, these changes reduced run-times to the point that systematic simulation studies became possible. Furthermore, the graph-based landscape structure made it straighforward to add global pathfinding as well as more structured information uptake and exchange to the model.

To address the final point, we designed a new model of belief dynamics based on mass-action-like interactions between doubt and certainty components (space constraints preclude a full model description; for more details see Hinsch and Bijak 2019). In this version of the model, change of belief is affected by similarity of beliefs and differences in confidence. In particular, contrary to the previous model, doubt increases in interacting agents with strong, but very different beliefs, but decreases for very similar beliefs.

Using the second, updated version of the model we ran systematic simulation experiments and carried out the statistical sensitivity analysis of selected outputs, examples of which are listed in Section 2.5. We identified different features of information exchange as the main drivers of the characteristics of the emerging migration routes. We also found clear indications for self-organised, sub-optimal routes if the agents' main source of information was other agents (Hinsch and Bijak 2019).

3.3 Third Iteration

The second version of the model clearly contained some features that could have resulted from some of the implementation choices, rather than being related to the characteristics of the model as such. To test this, a re-implemention of a simplified version of the model in ML3 (Reinhardt et al. 2019), a domain-specific language for agent-based demographic models, was carried out. When comparing these two implementations of the same conceptual model we found differences in results for a number of scenarios. We suspected that this was mainly due to ML3 requiring event-based scheduling in continuous time, whereas scheduling in the original Julia-based version of the model was based on discrete time steps.

We therefore modified the agent-based model further to use event-based scheduling, creating a third version of the model. This also enabled us to remove a number of awkward or implausible assumptions that were mostly dictated by the step-wise scheduling. In particular, we separated movement, information exchange, exploration and path planning into four independent "processes" that happen non-deterministically at specific rates, and let the duration of travel depend on the distance and friction of a transport link.

While the transition to an event-based paradigm did not in itself have significant effects on the model behaviour, the additional changes we made exposed a number of unexpected effects. Preliminary results indicated that large parts of the system dynamics are determined by feedback effects in the early stages of the modelled process. For example, a greater error rate in the agents' communication can lead to less uniform movement of the agents and therefore to more spread out exploration in the initial stages, before a contiguous route is known. This change leads to an increase in knowledge available in the population as a whole. Similarly, if agents have more prior knowledge a route will be found faster and therefore more areas remain unexplored. This reinforced the findings from the exploratory sensitivity analysis, presented in Section 2.5, that information is crucial for the route formation process.

These insights into the model behaviour enabled us to prepare for the inclusion of additional information from other components of the proposed approach, which are the focus of the current work. In particular, the fourth version of the model, presently under development, includes additional empirical data on selected aspects of information exchange between migrants (Section 2.2), as well as results of two rounds of psychological experiments on different features of human decisions – one on utility functions, presented in Section 2.3, and another on subjective probabilities, which is currently in progress.

4 PROVENANCE DESCRIPTION OF THE MODELLING PROCESS

Assessing the the reliability and robustness of the results of a simulation study requires knowledge of how they were attained. Hence, a thorough documentation of the model, its assumptions, and its foundation in theory and data is essential. Equally important is meta-information about the process of the model's creation, an example of which is described in Section 3. While this can be, and usually is, described in text, e.g., as ODD (Grimm et al. 2020), a more structured approach is desirable, especially for complex studies, such as this one. *Provenance models* have recently been proposed to capture this information (Ruscheinski and Uhrmacher 2017), and to supplement text-based documentation (Reinhardt et al. 2018b). We follow this approch, and represent the provenance as a directed acyclic graph, applying the W3C PROV standard (Groth and Moreau 2013). Simulation models, data sets, or other artifacts of a simulation study are modelled as *entities*. In addition, the provenance model contains *activities*, e.g., creating a model or performing an experiment. Edges in the provenance model represent relationships between entities and activities; an entity might be *used in*, or *generated by* an activity. Figure 5 gives an overview of a provenance model for our simulation study, including connections between the different elements and processes.

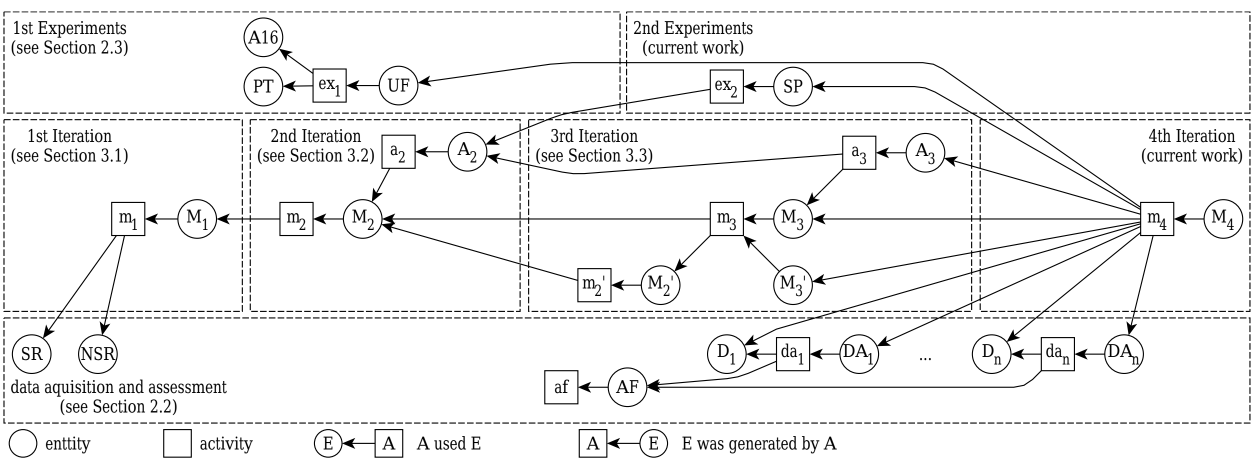

Figure 5: Provenance model of the agent-based migration model and the modelling process. Arrows point towards dependencies, and time flows roughly left to right. See Table 2 for a short description of the items.

Table 2 provides short descriptions of the entities and activities included in the provenance model. The boxes refer to parts of the study described in this paper, e.g. the box labelled *3rd iteration* depicts the process described in Section 3.3: Re-implementation of the model in ML3 (activity m_2' and its product M_{2-ML3}), comparison and adaptation of the two implementations (m_3), resulting in two adapted versions (M_3 and M_{3-ML3}), as well as a repeated analysis of the model (a_3 and the results A_3). The models, analysis, results, and other entities will feed into the (currently ongoing) fourth iteration of the model (m_4 and M_4).

5 CONCLUSIONS AND LESSONS LEARNED

The proposed scientific approach, coupling agent-based modelling with statistical analysis, is a natural way to address the dual challenges of complexity and uncertainty of contemporary migration processes. It does

Table 2: Short description of the entities and activities included in the provenance model in Figure 5.

Item	Description
$A16$	Methodology of Abdellaoui et al. (2016)
A_2, A_3	Analysis of the results, e.g., Table 1 and Figure 4
AF	Data assessment framework
D_1 to D_n	Datasets
DA_1 to DA_n	Assessment of the datasets (see Figure 2)
$M_1, \ldots, M_4$	Model versions
M_{2-ML3}, M_{3-ML3}	Model versions in the ML3 implementation
SR, NSR	Reports about migration route formation
PT	Prospect theory (Kahneman and Tversky 1979) as the theoretical foundation of ex_1
UF, SP	Utility functions (see Figure 3) and subjective probabilities elicited in the experiments
a_2, a_3	Model analysis
af	Creation of the data assessment framework
$da_1, \ldots, da_n$	Conduction of the data assessment
ex_1, ex_2	Conduction of experiments
$m_1, \ldots, m_4, m_2'$	Modelling activities

so by shedding light on the micro-level mechanisms underpinning the observed macro-level processes, and at the same time describing the probabilistic properties of the migration system under study. Each of the five building blocks provides a contribution in its own right, but combining them together by means of a statistical analysis (cf. Kleijnen and van Beers 2004) allows for formal exploration of the behaviour of complex migration systems in a more rigorous and transparent way than has been the case before.

By following an iterative process shown in Figure 5, we were able to construct models of increasing analytical potential, while maintaning their computational tractability. In the process, the trade-offs between some level of detail (e.g., topology of the modelled world) and practical considerations, such as run-times, became clear. Additionally, parallel implementation in two languages using different underlying formalisms illuminated specific issues with some of the details, and helped us better understand the model dynamics (Reinhardt et al. 2019). Formal documentation of the process by using provenance modelling helps keep track of the building blocks used in the successive versions of the model, which is very much in line with the inductive principles of model-based science advocated for by Courgeau et al. (2017) and others.

More generally, by framing the analysis in a multi-perspective way, we also hope to strengthen the case for model-based demography and migration research, which is intended to fill the widely acknowledged theoretical gaps in wider population studies (Courgeau et al. 2017; Burch 2018). At the same time, migration – despite being one of the key social challenges of the contemporary world – is particularly lagging behind in terms of utilising the potential for model-based enquiries, and especially lacking a formal discussion of causal mechanisms, which would enable testable predictions to be made (Willekens 2018). We hope that the approach proposed in this paper will go some way towards achieving this goal.

ACKNOWLEDGMENTS

This research received funding from the European Research Council (ERC) via research grant Bayesian Agent-based Population Studies (CoG-2016-725232), which is gratefully acknowledged.

REFERENCES

Abdellaoui, M., H. Bleichrodt, O. L'Haridon, and D. van Dolder. 2016. "Measuring Loss Aversion under Ambiguity: A Method to Make Prospect Theory Completely Observable". *Journal of Risk and Uncertainty* 52(1):1–20.

Arango, J. 2000. "Explaining Migration: A Critical View". *International Social Science Journal* 52(165):283–296.

Borkert, M., K. E. Fisher, and E. Yafi. 2018. "The Best, the Worst, and the Hardest to Find: How People, Mobiles, and Social Media Connect Migrants in(to) Europe". *Social Media + Society* 4(1):1–11.

Bortolussi, L., R. De Nicola, V. Galpin, S. Gilmore, J. Hillston, D. Latella, M. Loreti, and M. Massink. 2015. "CARMA: Collective Adaptive Resource-Sharing Markovian Agents". *Electronic Proceedings in Theoretical Computer Science* 194:16–31.

Brenner, T., and C. Werker. 2009. "Policy Advice Derived from Simulation Models". *Journal of Artificial Societies and Social Simulation* 12(4):2.

Burch, T. 2018. *Model-Based Demography: Essays on Integrating Data, Technique and Theory*. Cham: Springer.

Castles, S. 2004. "Why Migration Policies Fail". *Ethnic and Racial Studies* 27(2):205–227.

Cottineau, C., P. Chapron, and R. Reuillon. 2015. "Growing Models from the Bottom Up. an Evaluation-Based Incremental Modelling Method (EBIMM) Applied to the Simulation of Systems of Cities". *Journal of Artificial Societies and Social Simulation* 18(4):9.

Courgeau, D., J. Bijak, R. Franck, and E. Silverman. 2017. "Model-Based Demography: Towards a Research Agenda". In *Agent-Based Modelling in Population Studies*, edited by J. Van Bavel and A. Grow, 29–51. Cham: Springer.

Crawley, H., F. Düvell, K. Jones, S. McMahon, and N. Sigona. 2016. "Destination Europe? Understanding the dynamics and drivers of Mediterranean migration in 2015". Technical Report, Centre for Trust, Peace and Social Relations at Coventry University, Coventry.

Czaika, M. 2015. "Migration and Economic Prospects". *Journal of Ethnic and Migration Studies* 41(1):58–82.

Dekker, R., G. Engbersen, J. Klaver, and H. Vonk. 2018. "Smart Refugees: How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making". *Social Media + Society* 4(1):1–11.

Grimm, V., S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld, A. S. A. Johnston, A. Milles, J. Nabe-Nielsen, J. G. Polhill, V. Radchuk, M.-S. Rohwäder, R. A. Stillman, J. C. Thiele, and D. Ayllón. 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(2):7.

Groth, Paul and Moreau, Luc 2013. "PROV-Overview – An Overview of the PROV Family of Documents". https://www.w3.org/TR/prov-overview/, accessed August 3rd 2020.

Hinsch, M., and J. Bijak. 2019. "Rumours lead to self-organized migration routes". Technical Report, University of Southampton. https://www.southampton.ac.uk/baps/publications.page, accessed August 3rd 2020.

Honorato-Zimmer, R., A. J. Millar, G. D. Plotkin, and A. Zardilis. 2019. "Chromar, a Language of Parameterised Agents". *Theoretical Computer Science* 765:97–119.

Jones, B., and C. J. Nachtsheim. 2011. "A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects". *Journal of Quality Technology* 43(1):1–15.

Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk". *Econometrica* 47(2):263–291.

Kennedy, M. C., and G. P. Petropoulos. 2016. "The Gaussian Emulation Machine for Sensitivity Analysis". In *Sensitivity Analysis in Earth Observation Modelling*, edited by G. P. Petropoulos and P. K. Srivastava, 341–361. Amsterdam: Elsevier.

Kingsley, P. 2016. *The New Odyssey: The Story of Europe's Refugee Crisis*. London: Faber & Faber.

Klabunde, A., and F. Willekens. 2016. "Decision-Making in Agent-Based Models of Migration: State of the Art and Challenges". *European Journal of Population* 32(1):73–97.

Kleijnen, J. P. C., and W. C. M. van Beers. 2004. "Application-Driven Sequential Designs for Simulation Experiments: Kriging Metamodelling". *Journal of the Operational Research Society* 55(8):876–883.

Lorenz, J. 2007. "Continuous Opinion Dynamics under Bounded Confidence: A Survey". *International Journal of Modern Physics C* 18(12):1819–1838.

Madl, T., K. Chen, D. Montaldi, and R. Trappl. 2015. "Computational Cognitive Models of Spatial Memory in Navigation Space: A Review". *Neural Networks* 65:18–43.

Masad, D., and J. L. Kazil. 2015. "Mesa: An Agent-Based Modeling Framework". In *Proceedings of the 14th Python in Science Conference*. July 6th–12th, Austin, Texas, 51–58.

Nurse, S., and J. Bijak. 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, accessed August 3rd 2020.

Oakley, J. E., and A. O'Hagan. 2004. "Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(3):751–769.

Poulain, M., N. Perrin, and A. Singleton. (Eds.) 2006. *THESIM: Towards Harmonised European Statistics on International Migration*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Reinhardt, O., J. Hilton, T. Warnke, J. Bijak, and A. M. Uhrmacher. 2018a. "Streamlining Simulation Experiments with Agent-Based Models in Demography". *Journal of Artificial Societies and Social Simulation* 21(3):9.

Reinhardt, O., A. Ruscheinski, and A. M. Uhrmacher. 2018b. "ODD+P: Complementing the ODD Protocol With Provenance Information". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 727–738. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Reinhardt, O., A. M. Uhrmacher, M. Hinsch, and J. Bijak. 2019. "Developing Agent-Based Migration Models in Pairs". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H. G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 2713–2724. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ruscheinski, A., and A. Uhrmacher. 2017. "Provenance in Modeling and Simulation Studies – Bridging Gaps". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 872–883. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. *Global Sensitivity Analysis: The Primer*. Chichester: John Wiley.

Sargent, R. G. 2013. "Verification and Validation of Simulation Models". *Journal of Simulation* 7(1):12–24.

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

ten Broeke, G., G. van Voorn, and A. Ligtenberg. 2016. "Which Sensitivity Analysis Method Should I Use for My Agent-Based Model?". *Journal of Artificial Societies and Social Simulation* 19(1):5.

Vogel, D., and V. Kovacheva. 2008. "Classification Report: Quality Assessment of Estimates on Stocks of Irregular Migrants". Technical Report, Hamburg Institute of International Economics, Hamburg.

Warnke, T., O. Reinhardt, A. Klabunde, F. Willekens, and A. M. Uhrmacher. 2017. "Modelling and Simulating Decision Processes of Linked Lives: An Approach Based on Concurrent Processes and Stochastic Race". *Population Studies* 71(sup1):69–83.

Wilensky, U. 1999. "NetLogo". https://ccl.northwestern.edu/netlogo/, accessed August 3rd 2020.

Willekens, F. 2018. "Towards Causal Forecasting of International Migration". *Vienna Yearbook of Population Research* 16:199–218.

AUTHOR BIOGRAPHIES

JAKUB BIJAK is Professor of Social Demography and currently a Joint Head of Department of Social Statistics and Demography at the University of Southampton, UK. At present, he leads an ERC funded project on Bayesian Agent-Based Population Studies and a H2020 project QuantMig: Quantifying Migration Scenarios for Better Policy. His email is j.bijak@soton.ac.uk.

PHILIP A. HIGHAM is Reader in Cognitive Psychology within Psychology at the University of Southampton. His research focuses on long-term human memory and metacognition (cognition about cognition). One strand of his research focuses on the basic decision processes underlying metacognitive judgments. His email address is p.a.higham@soton.ac.uk.

JASON HILTON is Lecturer in Social Statistics and Data Science in the Department of Social Statistics and Demography at the University of Southampton. His research focuses on probabilistic projection of demographic processes, and on the management of uncertainty in agent-based demographic models. His email address is j.d.hilton@soton.ac.uk.

MARTIN HINSCH is a Research Fellow in the Department of Social Statistics and Demography at the University of Southampton. He is interested in emergent structures and complex systems and has done research in theoretical biology, bioinformatics, machine learning, epidemiology and swarm robotics. His email address is hinsch.martin@gmail.com.

SARAH NURSE is a Research Fellow in the Department of Social Statistics and Demography at the University of Southampton. Her main research interests include migration, with a particular focus on forced migration and asylum, evaluation of migration data quality, and the demography of conflict and violence. Her email address is s.l.nurse@soton.ac.uk.

TOBY PRIKE is a Research Fellow in the Department of Social Statistics and Demography at the University of Southampton. He holds a PhD in Psychology from Flinders University, Australia. His research looks at non-evidence based beliefs, probabilistic reasoning, cognitive biases, and decision making under uncertainty. His email address is t.m.prike@soton.ac.uk.

OLIVER REINHARDT is a Ph.D. student in the Modeling and Simulation Group at the University of Rostock. He holds an MSc in Computer Science from the University of Rostock. In his research, he is concerned with domain-specific modeling languages and the methodology of agent-based simulation. His email address is oliver.reinhardt@uni-rostock.de.

PETER W. F. SMITH is Professor of Social Statistics in the Department of Social Statistics and Demography, University of Southampton, and a Fellow of the British Academy. His research includes developing statistical methods for handling non-response, longitudinal data, and applications in demography, medicine and health sciences. His email address is p.w.smith@soton.ac.uk.

ADELINDE M. UHRMACHER is Professor at the Institute of Visual and Analytic Computing of the University of Rostock and head of the Modeling and Simulation Group. She holds a PhD in Computer Science from the University of Koblenz and a Habilitation in Computer Science from the University of Ulm. Her email address is adelinde.uhrmacher@uni-rostock.de.