Machine Learning to Derive Complex Behaviour in Agent-Based Modelling

Ellen-Wien Augustijn
Dept. of Geo-Information Processing
Faculty of Geo-Info. Science & Earth Observation
University of Twente
Enschede, The Netherlands
p.w.m.augustijn@utwente.nl

Shaheen A. Abdulkareem
Department of Computer Science
College of Science
University of Duhok
Duhok, Kurdistan-region, Iraq
sheheen.abdulkareem@uod.ac

Mohammed Hikmat Sadiq
Department of Computer Science
College of Science
University of Duhok
Duhok, Kurdistan-region, Iraq
mohammed.sadiq@uod.ac

Ali A. Albabawat
Department of Computer Science
College of Science
University of Duhok
Duhok, Kurdistan-region, Iraq
mohammed.sadiq@uod.ac

***Abstract*—The use of machine learning algorithms to enrich agent-based models has increased over the past years. This integration adds value when combining the advantages of the data-driven approach and the possibilities to explore future situations and human interventions. However, this integrating is still in its infant stage. Full integration of learning algorithms and agent-based models is often technically challenging and can make the behavioural rules of the agents less transparent. Experiments are needed in which different integration strategies are compared using the same agent-based model to determine when each of these approaches is most effective. In this paper, we present a comparison of two versions of the same cholera model. In the initial version, agent behaviour was driven directly by a learning algorithm. In our experiments, we replace this strategy by applying a learning algorithm directly on the data and implement the outcomes as behaviour rules in the model. The results showed that when the integration aims to create agents that show characteristics that are data-driven, deriving rules based on these data is a good alternative. In addition, a key element in this strategy is the dataset. A large dataset representing the behaviour of different types of agents over the complete time period is needed.**

Keywords—Intelligent agent, disease simulation, agent-based simulation, risk perception, decision trees

I. Introduction

Global change, sustainability, and complexity in which the behaviour of humans plays a crucial role are targetable phenomena in academic research. We need empirical methodologies to identify the processes that explain these phenomena and provide managerial strategy tools. Agent-based models (ABMs) have proven to be one of these simulation tools that help policymakers to identify and assess different strategies [1]. The integration of machine learning (ML) algorithms and ABMs has increased over the past years, yet it is still in its infant stage. The added value is the combination of the advantages of the data-driven approach and the possibilities to explore future situations and human interventions [2]. In this paper, we present a comparison of two versions of the same ABM of cholera diffusion. In the initial version, agent behaviour was driven directly by an ML algorithm [3]. In our experiments in this paper, we replace this strategy by applying ML directly on the data using decision tree C4.5 and implement the outcomes as behaviour rules in the ABM.

II. Integration of ABM and ML

ABMs provide a framework that allows representing a spatial environment containing heterogeneous agents, that display micro/macro relationships, and have with adaptive behaviour [4]. Typically, ABMs are developed to identify the realism and level of details required to model and understand certain behavioural rules in real-life applications [5].

In order to make agents behave naturally, ABMs often apply ML algorithms governing agents' behaviour [6]. Behaviour of human agents in ABMs may employ various ML algorithms to form expectations and opinions about the environment and future trends of other variables of interests [7].

Augustijn *et al.* (2019) mentioned that the integration of implementing ML algorithms and ABM can be formulated in three phases [8]:

- Learning algorithms driven by data are used to derive information to generate agents, agent attributes and agent behaviour [9];
- ML and ABM are fully integrated to generate agents that learn during the ABM simulation [3];
- Two individual models (one ML and one ABM) are generated and outcomes are matched during the calibration/validation process [10], [11].

However, full integration of ML and ABM is often technically challenging and regarded by ABM designers as a Blackbox as behaviour rules are less transparent [5], [12], [13]. In addition, usually behavioural data that needed to simulate human behaviour in models are not available for most of real life applications [14]. Therefore, experiments are needed in which different integration strategies are compared using the same ABM model to determine when each of these approaches is most effective.

978-1-7281-5249-3/20$31.00 ©2020 IEEE

III. METHODS

A. Case Study: Cholera Diffusion ABM

To test the benefit of applying machine learning in the design of the ABM to determine the behaviour rules of the agents, we used the cholera agent-based model (CABM) as a testbed. CABM is a geographically explicit model that simulates cholera transmission in the urban area of Kumasi, Ghana [15]. The goal of CABM is to examine the role of runoff water from open dumpsites as a trail for the diffusion of cholera. CABM incorporates environmental and human behavioural elements. The model has been used to explore the impact of implementing ML algorithms to steer the behaviour of agents [3], to compare the individual and collective learning [16] and to integrate the spatial intelligence for risk perception of agents in the model [17]. CABM contains three agent types: households, individuals, and rain particles. Household agents are heterogeneous in terms of their attributes such as income level, hygiene level, water source, and house location. Individual agents have heterogeneous attributes including age, gender, blood type, and health status (susceptible, infected, and recovered). The agent population (households and individuals) is generated using a synthetic population generator that provides the model with its largest stochastic element.

Household agents use an ML algorithm to perceive cholera risk and, in case of risk, to adapt themselves and making a coping decision. Protection motivation theory [18] was adapted to simulate the processes of risk perception (threat appraisal) and coping appraisal in CABM Fig. 1. Several factors impact the risk perception and coping appraisal of household agents. These factors vary from environmental to demographic characteristics. The factors for threat appraisal include: visual pollution (VP) of the river water (spatial environment), news of cholera diffusion via media, communication with other agents (social interaction with neighbours) and previous experience with cholera (memory).

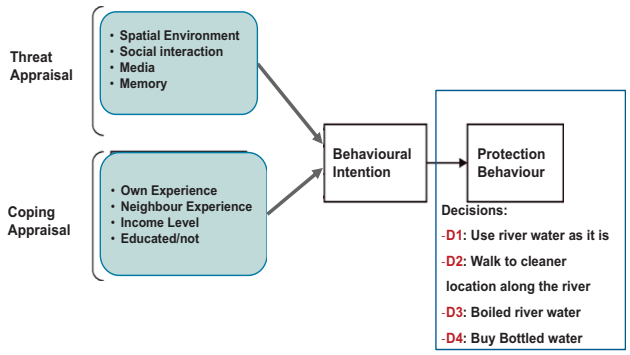

Fig. 1. Protection Motivation Theory Adapted to CABM

While for coping appraisal the factors: household' own experience of using river water, neighbours' experience with using the river water, income level, and education level are concerned.

In the study area, 14% of low-and middle- income level household agents do not have access to tap water. This percentage increases when it rains heavily in Kumasi [15]. Therefore, households need to go to the river and fetch water. Based on their risk perception, they evaluate if the water is infected with cholera or not. Then accordingly, they need to make one of the decisions shown in Fig. 1.

B. Behavioural Dataset

One of the challenges in the field of behavioural science is the lack of available behavioural datasets [19]. For the cholera outbreak of 2005 in Kumasi, Ghana, no risk perception data is available that shows how the people of Kumasi changed their health-seeking behaviour during the course of the disease outbreak. The only available dataset for the 2005 epidemics is the disease surveillance data of cholera cases reported to the Disease Control Unit (DCU). onto collect individual behaviour data, we ran a survey through a Massive Open Online Course (MOOC) – Geohealth in two rounds 2016 and 2017. MOOC participants were introduced to the cholera disease problem by showing them pictures of water and asking what their coping decision would be under different conditions (factors we included in the process of risk perception). This dataset was used in [20] to construct and train Bayesian Networks (BNs) that were used to steer the risk perception and coping appraisal behaviours of agents. In this research, we use the same dataset to derive the behavioural rules of household agents for their both risk perception and coping appraisal.

C. Decision Trees to Derive Agents Behavioural Rules

In machine learning, there are three types of learning: supervised, unsupervised and reinforcement learning. With the existence of data, supervised learning is implemented. Decision trees are one of the techniques that use data for training and creating predictive models. This technique is mapping observation in the dataset to a decision. Decision trees have been implemented with agent-based modelling to derive a certain behaviour such as in [21]. Nodes in a decision tree represent attributes (factors) that impact the final decision (leaves). The root node represents the factor that has the highest impact on the decision. The next connected nodes are those that correlated to the previous node, they have less impact on the decision and their impact depends on the value of the previous node. There is a number of methods to create decision trees: Chi-squared automated interaction detector (CHID), classification and regression trees (CART), iterative dichotomizer (ID3) and C4.5 (successor of ID3). For the purpose of this paper, we use the C4.5 decision tree algorithm. C4.5 has been selected to be used to derive the risk perception and coping appraisal rules of the agent-based model. This is because C4.5 has high modularity and can easily interpret data to rules. A rule can be easily understood and be read without connecting it to other rules in the tree [21], [22].

D. Software

The CABM is developed using NetLogo (version 5.2.0). NetLogo is a multiagent modelling software that is developed

at the CCL and authored by Uri Wilnesky [23]. The rules of both risk perception and coping appraisal were derived by C4.5 using Weka. Weka is a Java-based machine learning toolkit that is developed at the University of Waikato [24]. Both software packages are open-source and easy to code and use even for non-expert programmers.

E. Implementation Steps

To achieve the purpose of this paper, Fig. 2 shows the steps of setup of this research:

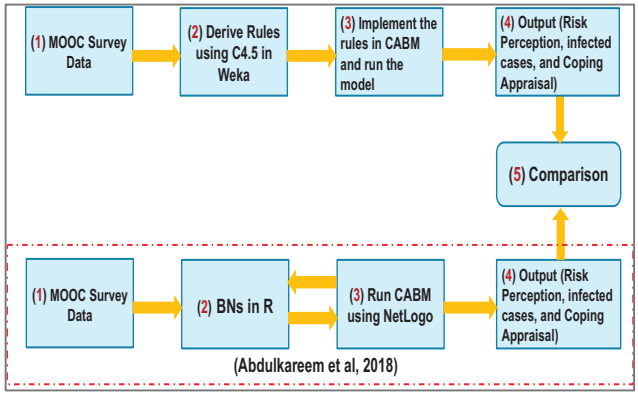

Fig. 2. Flow Diagram of the Setup of this Research

The implementation of ML algorithms in CABM consists of four stages: preparation of MOOC data (1), using ML to derive the behaviour of agents (top row) and/or steering the agent behaviour (bottom) (2), implementing the behaviour rules in CABM (top row) and running the CABM (3), and evaluating the output of the simulations (4). To evaluate the difference between the two implementations of ML, we compare the results (5).

IV. Results and Discussion

A. Risk Perception Decision Tree

Using Weka, we use the MOOC dataset as input to derive the rules for the stages of both risk perception and coping appraisal. For risk perception (threat appraisal), the resulted decision tree is shown in Fig. 3.

According to the dataset, media has the highest impact on the perception of cholera risk. Whenever media starts to broadcast news about cholera, people pay attention to it even though they may not have previous experience with cholera (memory). The level of visual pollution of the river water does not impact their risk perception when media is activated. Communication with neighbours also does not impact their risk perception. They fully depend on what the media broadcasts. The sequence of importance of information sources is as follows: Media > Memory > Visual Pollution > Neighbours.

Under absence of media, people depend on other channel of information, represented by communication with neighbours. When there are infected cases in their neighbourhood, then, no matter whether they had a previous cholera experience, they perceive cholera risk.

Although, normally, people depend on their visual observations and memory to make decisions of feeling at risk. With the rules derived from MOOC dataset, the level of pollution (no, low, high) or the type of memory (positive, negative) with cholera has less impact on the risk perception of participants comparing to communication with neighbours as can be seen in Fig. 3. The sequence of importance of information sources is as follows: Neighbours > Memory > Visual Pollution.

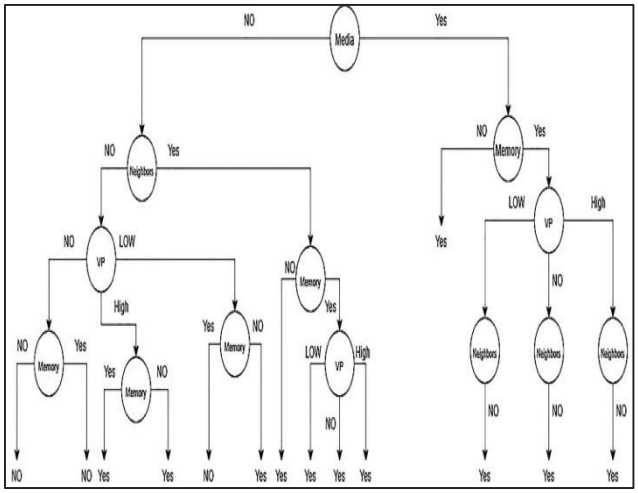

Fig. 3. Decision Tree of Risk Perception (Threat Appraisal)

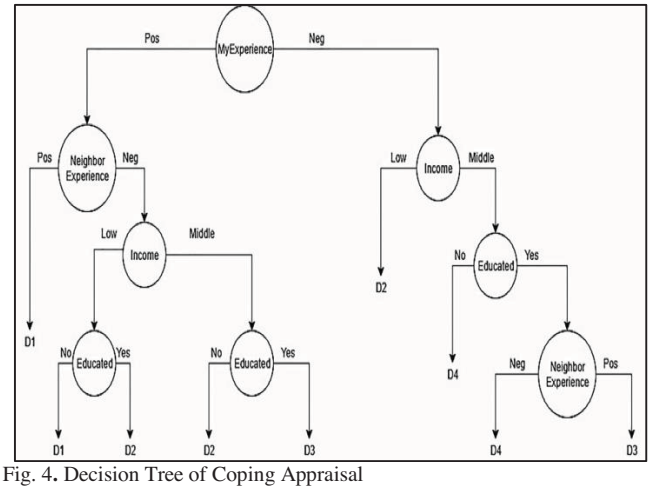

Fig. 4. Decision Tree of Coping Appraisal

B. Coping Appraisal Decision Tree

Figure 4 shows the rules derived for coping appraisal. The own experience of the individual with cholera is the main attribute towards which decision is made. When individuals do not have a negative experience with cholera (infection in their household), then infections in their neighbourhood will impact their coping decision. The selection of a coping decision takes the income level and education of the individual into account. Educated individuals with a middle-income level boil the water fetched from the river (D3) while the non-educated ones prefer to walk to a cleaner water point and fetch river water (D2). However, individuals with a low-

income level either use the river water as is (D1) if they are uneducated or walk to a cleaner water point and fetch the water (D2) if they are educated.

Whenever individuals experience cholera and have a negative experience, then depending on their demographic attributes, they make one of the coping decisions except using the river water as it is (D1). If they are middle-income level and educated but there are no infected cases in their neighbourhood then boiling water (D3) is a preferable decision. While a negative experience of their neighbours (infected cases in their neighbourhood) leads to buying bottled water (D4). Also, uneducated-middle income level individuals prefer D4, even when infected cases in their neighbourhood exist. Finally, for individuals with a low-income level, walking to a cleaner water point is a preferable decision since it does not cost money.

C. *Running CABM*

The epidemic of cholera in Kumasi, Ghana in 2005 lasted for 90 days. Therefore, the simulation in CABM run for 90 days as well. Every tick in the model represents one hour in real life. We ran CABM 100 times using rules of risk perception and coping appraisal that were derived from MOOC data using C4.5 to get stable results. The number of infected cases and the number of household agents who perceived risk is recorded daily using their rule-based threat appraisal function. In addition, the number of households that follow a specific coping decision is recorded. In order to evaluate the effectiveness of implementing C4.5 to derive agents behaviour, we compared the outcomes of this version of CABM with the one in [3]. The results in [3] had been statistically evaluated using sensitivity analysis and normality test.

Comparing the risk perception curve of the current CABM with the one in [3] shows that household agents behave smartly in the current version Fig. 5.

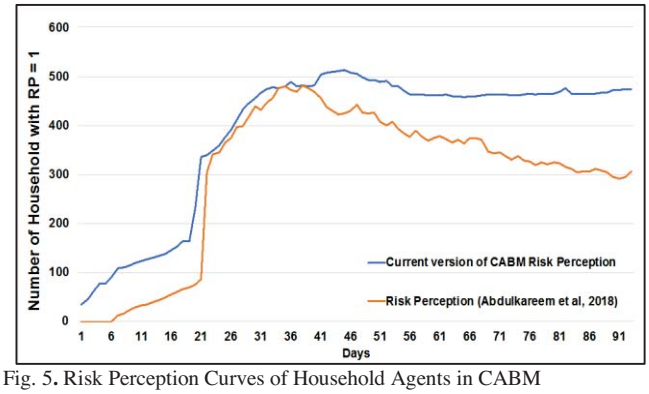

Fig. 5. Risk Perception Curves of Household Agents in CABM

From day one onward, household agents perceiving cholera risk, although no infected cases occurred yet.

However, when the media starts to broadcast news about cholera after three weeks, more household agents perceive risk. Their number keeps increasing until the middle of the epidemic period where it keeps high and stable to the end of the simulation time. While in [3], once the number of infected cases decreases Fig. 6, the number of households with risk perception decreases as well Fig. 5.

Comparing the number of infected cases per model Fig. 6, we notice that fewer cases were recorded in the current version of CABM. This is due to the high-risk perception of the households in the model. They take protective measures to prevent cholera infection.

According to the coping appraisal rule of the agents Fig. 4, household agents avoid using river water as it is (D1) and look for alternative water sources. These effective decisions explain the smaller number of infected cases in the model. In [3], household agents avoid buying bottled water. Their learning process taught them that boiling water can be safe and costs less than buying bottled water. Within the rule-based coping appraisal, household agents follow the most effective decision rather than learning which decision could be cheaper and also protective.

Figures 7 and 8 show the decisions that had been made by household agents in both versions of CABM.

Using ML to steer the coping appraisal behaviour in [3] helped household agents to learn during the simulation and improve their decision making process. They select the decision that fits their risk perception and their demographic attributes (income level and being educated/not). Buying bottled water (D4) is an expensive decision comparing to other decisions (D1 – D3). Household agents learned that boiling river water (D3) is cheap and effective Fig. 7.

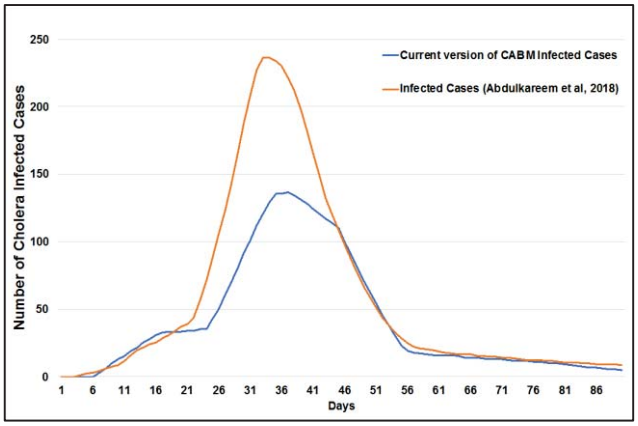

Fig. 6. Epidemic Curves of CABM Indicating Infected Cases per Model Version

All MOOC participants were well – educated people that, most probably, come from a middle or high-income level. Furthermore, they were from different countries around the world with different cultural backgrounds. This impacts the decision making of how/which water they should use/drink. Buying bottled water (D4) reflects that they are concerned about protecting themselves, even if the decision costs more money

. In addition, the education level of the household agents helps them to also think about boiling water fetched from river especially after they knew from media that there is cholera Fig. 8.

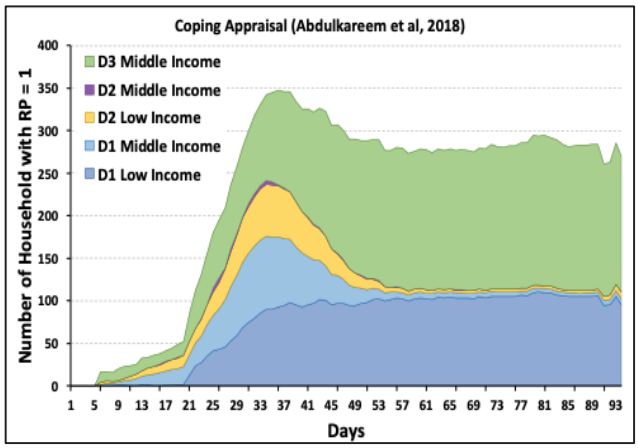

Fig. 7. Coping Appraisal Decisions of Household Agents in [3]

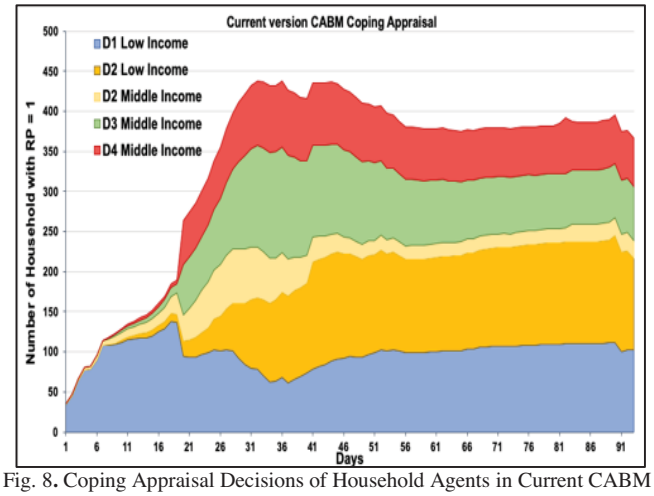

Fig. 8. Coping Appraisal Decisions of Household Agents in Current CABM

D. Models Performance

Running one simulation of CABM in [3] requires 95 minutes to be accomplished. In [3], CABM implemented BNs to steer risk perception and coping appraisal behaviour. BNs were implemented in the R statistical language. Exchanging data between NetLogo and R requires extra simulation.

The current version of CABM runs the simulation only via NetLogo as rules of risk perception and coping appraisal were derived in Weka before running the simulation. CABM does not connect to any other software to steer the behaviour of the agents. Therefore, running one simulation requires 30 minutes. This saves one hour per simulation run.

V. CONCLUSIONS

The machine learning algorithm in this version of CABM was used to derive behaviour rules outside the ABM. The rules were derived from MOOC data using the C4.5 machine learning algorithm. The MOOC data were collected from well-educated people from countries all around the world that belonged to different cultural backgrounds. Therefore, the rules derived from these data reflect rational behaviour. The MOOC participants education level, and possibly their higher income level, was reflected in their responses to the survey. They are aware of the impact cholera can have and know how to effectively protect themselves against cholera. By implementing risk perception and coping appraisal rules derived from this data, agents behave like the participants of the MOOC survey.

In the previous version of CABM, the machine learning algorithm (represented by Bayesian Networks) was used to steer the behaviour of agents inside the ABM. In these versions, agents learnt gradually and simultaneously with the learning process of the machine learning algorithm itself. Therefore, they started to show rational behaviour later, and had a stronger preference for a cheaper yet effective coping strategy (boiling the water).

There is a clear difference between deriving rules for an ABM via ML and using ML to steer agent-behaviour as an integrated part of the ABM, reflected in the differences found in the two models. In case that learning during the simulation is needed, integration of the ABM and ML is required. This does not mean that deriving behavioural rules from data cannot be a good alternative in some cases. Technically, full integration of ML and ABM is difficult as ABM platforms do often not contain ML algorithms. When the integration aims to create agents that show characteristics that are data-driven, deriving rules based on these data is a good alternative. As the ML algorithms are often not integrated in the ABM software, this approach makes it easier to experiment with different ML algorithms and compare their results.

A key element in this strategy is the dataset. We cannot depend on a limited group of people to derive the behaviour that should be used for others. A large dataset representing the behaviour of different types of agents over the complete time period is needed.

Decision trees are good at ordering variables in a sequence of importance (media is more important than memory). However, they do not indicate numeric information that can be used to derive agent behaviour rules unless regression models of decision trees are used which gives probabilities of each attribute that impact the decision.

The current implementation of the model derives rules that remain fixed throughout the simulation. For the future work, an ML algorithm is required that is able to derive rules from datasets but also update these rules later during the simulation. Also, integrating ML algorithms in ABM software can recover the related technical problems of connecting ABM software with another software that ML is implemented with.

REFERENCES

[1] C. M. Macal, "Everything you need to know about agent-based modelling and simulation," J. Simul., vol. 10, no. 2, pp. 144–156, 2016.

[2] H. Kavak et al., "Big Data, Agents, And Machine Learning: Towards A Data-Driven Agent-Based Modeling Approach," in Proceedings of the Annual Simulation Symposium, 2018, no. Gilbert 2008, p. 12.

[3] S. A. Abdulkareem, E.-W. Augustijn, Y. T. Mustafa, and T. Filatova, "Intelligent judgements over health risks in a spatial agent-based model," Int. J. Health Geogr., vol. 17, no. 1, p. 8, Dec. 2018.

[4] T. Filatova, P. H. Verburg, D. C. Parker, and C. A. Stannard, "Spatial agent-based models for socio-ecological systems: Challenges and prospects," Environ. Model. Softw., vol. 45, pp. 1–7, Jul. 2013.

[5] E. Bruch and J. Atwell, "Agent-Based Models in Empirical Social Research," Sociol. Methods Res., vol. 44, no. 2, pp. 186–221, May 2015.

[6] P. Grim et al., "Polarization and Belief Dynamics in the Black and White Communities : An Agent-Based Network Model from the Data," Artif Life, pp. 186–193, 2012.

[7] A. Chakraborti, I. M. Toke, M. Patriarca, and F. Abergel, "Econophysics review: II. Agent-based models," Quant. Financ., vol. 11, no. 7, pp. 1013–1041, Jul. 2011.

[8] E.-W. Augustijn, R. Kounadi, T. Kuznecova, and R. Zurita-Milla, "Teaching Agent-Based Modelling and Machine Learning in an integrated way," in 15th Geocomputation 2019: Adventures in GeoComputation, 2019.

[9] C. M. Buchmann, K. Grossmann, and N. Schwarz, "How agent heterogeneity, model structure and input data determine the performance of an empirical ABM - A real-world case study on residential mobility," Environ. Model. Softw., vol. 75, pp. 77–93, 2016.

[10] F. Lamperti, A. Roventini, and A. Sani, "Agent-based model calibration using machine learning surrogates," J. Econ. Dyn. Control, vol. 90, pp. 366–389, 2018.

[11] O. Matsumoto, M. Miyazaki, Y. Ishino, and S. Takahashi, "Method for Getting Parameters of Agent-Based Modeling Using Bayesian Network: A Case of Medical Insurance Market," Agent-Based Approaches Econ. Soc. Complex Syst., pp. 45–57, 2018.

[12] I. Lorscheid, H. Bernd-Oliver, and M. Meyer, "Opening the 'Black Box' of Simulations: Transparency of Simulation Models and Effective Results Reports Through the Systematic Design of Experiments," SSRN Electron. J., 2012.

[13] T. Van Der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints," BMC Med. Res. Methodol., vol. 14, no. 1, p. 137, Dec. 2014.

[14] A. Vinciarelli et al., "Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions," Cognit. Comput., vol. 7, no. 4, pp. 397–413, 2015.

[15] E. W. Augustijn, T. Doldersum, J. Useya, and D. Augustijn, "Agent-based modelling of cholera," Stoch. Environ. Res. Risk Assess., vol. 30, no. 8, pp. 2079–2095, 2016.

[16] S. A. Abdulkareem, E. W. Augustijn, T. Filatova, K. Musial, and Y. T. Mustafa, "Risk perception and behavioral change during epidemics: Comparing models of individual and collective learning," PLoS One, vol. 15, no. 1, p. e0226483, 2020.

[17] S. A. Abdulkareem, E.-W. Augustijn, Y. T. Mustafa, and T. Filatova, "Integrating Spatial Intelligence for risk perception in an Agent Based Disease Model," GeoComputation, no. September 2017, pp. 1–7, 2017.

[18] L. Bui, B. Mullan, and K. McCaffery, "Protection motivation theory and physical activity in the general Population: A systematic literature review," Psychol. Health Med., vol. 18, no. 5, pp. 522–542, Oct. 2013.

[19] S. Funk et al., "Nine challenges in incorporating the dynamics of behaviour in infectious diseases models," Epidemics, vol. 10, pp. 21–25, Mar. 2015.

[20] S. A. Abdulkareem, Y. T. Mustafa, E.-W. Augustijn, and T. Filatova, "Bayesian networks for spatial learning: a workflow on using limited survey data for intelligent learning in spatial agent-based models," Geoinformatica, vol. 23, no. 2, pp. 243–268, Apr. 2019.

[21] N. . Sánchez-Maroño, A. . Alonso-Betanzos, O. . Fontenla-Romero, J. G. Polhill, and T. Craig, "Empirically-derived behavioral rules in agent-based models using decision trees learned from questionnaire data," Underst. Complex Syst., no. 9783319463308, pp. 53–76, 2017.

[22] Lin Tan, "Decision Trees - an overview | ScienceDirect Topics." 2015.

[23] S. Abar et al., Agent Based Modelling and Simulation tools: A review of the state-of-art software, vol. 24. Elsevier, 2017, pp. 13–33.

[24] A. Jović, K. Brkić, and N. Bogunović, "An overview of free software tools for general data mining," in 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings, 2014, pp. 1112–1117.