Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

A COMPREHENSIVE FRAMEWORK FOR DATA-DRIVEN AGENT-BASED MODELING

Ruhollah Jamali[1], and Sanja Lazarova-Molnar[2,1]

[1]Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, DENMARK
[2]Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, GERMANY

ABSTRACT

Integrating data-driven methodologies with agent-based simulation presents an opportunity to automate modeling and enable Digital Twins for complex systems. This integration allows for utilization of real-world data to extract models that update with changes in the corresponding real systems and enhance our abilities to make informed decisions. We were unable to identify a systematic approach for developing data-driven agent-based models beyond isolated attempts focused on specific aspects. In response, we reviewed existing literature to develop a framework that systematically approaches data-driven agent-based modeling. We believe that our framework can assist in systematically evaluating which parts of agent-based models' development processes can be data-driven. Furthermore, we provide a comprehensive exploration of data-driven methods that can be applied to each stage of the model development process. Finally, we utilize our prior works in this area to demonstrate the application of data-driven methodologies in capturing patterns and insights for model development.

1 INTRODUCTION

Agent-based modeling (ABM) is a computational modeling approach that conceptualizes systems as compositions of autonomous, interacting agents (Macal and North 2005; Macal and North 2014). Each agent in an agent-based model possesses their own set of properties and behavior (Castiglione 2020), ranging from simple software components to entities with learning capabilities. Besides agents, environment and interaction rules are components that shape the dynamics of an agent-based model of a system. The environment provides the context in which agents exist and interact, and it can influence the behaviors of agents and their interactions. Interaction rules define how agents interact with each other and with the environment based on physical laws, social norms, or adaptive strategies (Epstein and Axtell 1996). ABM demonstrated a wide range of applications, especially in modeling individual decision-making in social and organizational behavior (Bonabeau 2002). The recent development of ABM tools, availability of collected data, and advances in computation facilitated the expansion of ABM applications in a variety of domains and disciplines (Macal and North 2014).

The development of agent-based models can be broadly split into three key processes: model design (extraction), calibration, and validation. Despite the wide range of applications of ABM, designing an agent-based model requires a comprehensive understanding of the subject system, which makes the design stage complex and biased toward the expert's view of the system. Incorporating data-driven methods can reduce bias in the model design process and lead to a more comprehensive and objective representation of the system (Yang and van Dam 2022). The calibration process of agent-based models can often be demanding due to the typically large number of parameters that need fine-tuning. Data-driven methods can improve the calibration process by using data to fine-tune the multitude of parameters, thereby enhancing the model's accuracy (Quera-Bofarull et al. 2023). Furthermore, due to the stochastic nature of agent-based models, validating them is a complex task. Data-driven methods can enhance the validation process of agent-based models by utilizing real-world data to compare and validate the model's predictions (Drchal et al. 2016). Considering the whole process of model development, data-driven modeling employs data

979-8-3315-3420-2/24/$31.00 ©2024

and data streams to inform and refine models continuously, leading to more robust and up-to-date models, as well as accurate and reliable predictions (Lazarova-Molnar and Li 2019; Kavak et al. 2018).

This paper is motivated by the increasing need for more accurate and efficient agent-based models across various fields (Steinbacher et al. 2021). However, applying data-driven methods to extract and refine agent-based models is not straightforward. This gap between the potential of data-driven modeling and its application in ABM is the primary motivation for this paper. Here, we present a framework that points to how and where we can integrate data-driven modeling with ABM, identifying the different elements in an agent-based model that can be impacted and enhanced by data. Our framework provides a structured approach to incorporating real-world data into agent-based models, enhancing models' accuracy and predictive power using previous efforts in data-driven ABM to motivate our findings and framework. In addition, we present our prior works in data-driven ABM to illustrate the practical application of data-driven modeling, serving as a proof of concept.

In this paper, we contribute to the ongoing efforts to improve the development of agent-based models and inspire further research in this direction. We believe that the successful integration of data-driven modeling and ABM can open up new possibilities for simulating and understanding complex systems. To accomplish our goal, we commence by reviewing the literature on data-driven agent-based modeling (DDABM) and simulation. This is followed by a discussion on the challenges and limitations inherent in the existing frameworks and the identification of essential concepts and requirements for a Framework for DDABM (Section 2). Subsequently, we explore the essential part of an agent-based model to explain how we designed the framework for DDABM (Section 3). Then, we present our previous efforts in DDABM as case studies that showcase the role of data-driven methodologies in the model development process (Section 4). Finally, we summarize the findings and discuss the potential of our proposed framework in facilitating the development of more realistic and practical agent-based models (Section 5).

2 RELATED WORK ON DATA-DRIVEN AGENT-BASED MODELING

Data-driven approaches for ABM are becoming increasingly popular in a variety of fields, which underscores their importance. In the following, we present examples of some of the related works and frameworks that have been introduced. Steinbacher et al. (2021) discussed the advances in ABM of economic and social behavior, highlighting the role of data-driven approaches in capturing the complexity and heterogeneity of economic systems. Similarly, Sajjad et al. (2016b) advocated for a data-driven approach in social simulation, emphasizing its benefits in capturing the complexity and heterogeneity of social systems. The integration of big data, agents, and machine learning in ABM has been promoted by Kavak et al. (2018), arguing that this integration can lead to more robust and accurate models. Jamali et al. (2024) demonstrated the importance of using data in ABM development through a case study of Schelling's model.

The application of data-driven approaches in ABM motivated the development of multiple frameworks for DDABM. Ravaioli et al. (2023), presents a data-driven agent-based model for agricultural land use. This model integrates machine learning (ML) algorithms to learn agents' behavioral rules from data, as opposed to relying on pre-defined theoretical or heuristic rules. CFBM (Combination Framework of Business Intelligence and Multi-agent based platform) framework manages and integrates empirical data collected from the target system and the data produced by the simulation model (Truong et al. 2016). Zhou et al. (2024) proposed a data-driven framework for ABM of vehicular travel. Their framework utilizes publicly available data to model the behavior of vehicles within an urban environment, resulting in a more accurate and realistic simulation. Lastly, Patsatzis et al. (2023) proposed an Equation/Variable free machine learning (EVFML) framework to control the collective dynamics of complex systems modeled via agent-based simulators.

The aforementioned frameworks are designed for specific applications or domains, such as agricultural land use, business intelligence, vehicular travel, and controlling the collective dynamics of complex systems. While these frameworks have proven effective in their respective domains, their specificity limits their applicability to other contexts or domains. The absence of a general framework also means that each new application requires the development of a new, custom framework, which can be time-consuming and resource-intensive. The first step towards improving the state of DDABM is to develop a general framework that is flexible and adaptable, capable of accommodating different types of data, and applicable across a wide range of domains. In this paper, we focus on developing a general framework to address the current challenges and limitations and provide a basis for further advancements in DDABM and digital twins.

3 FRAMEWORK FOR DATA-DRIVEN AGENT-BASED MODELING

To motivate our framework, we reviewed state-of-the-art applications of DDABM in different fields. Through this review, we observed how data can contribute to different aspects of model extraction (Zhang et al. 2016; Zhou et al. 2024; Zilske et al. 2011; Rosés et al. 2021; An et al. 2005), model refinement (Lamperti et al. 2018; Kim et al. 2021; Chen and Desiderio 2022a; Dyer et al. 2023; Clark et al. 2021; Niida et al. 2019), and model validation (Hua et al. 2022; Thaler and Siebers 2020; Dehkordi et al. 2023). Hence, our resulting framework targets these three key processes in ABM: model extraction, model refinement, and model validation, each represented by a specific component. Data-driven approaches can support these processes to further automate ABM. Additionally, we introduce Data Pipeline as the fourth component, which focuses on data collection. In the following, we elaborate on the specific element of each of these components and illustrate our framework for DDABM. Figure 1 illustrates our framework for data-driven agent-based model development. **Data Pipeline** is concerned with the collection, processing, and preparation of data before its utilization for the other three key processes in ABM. **Model Extraction** employs the processed data from the Data Pipeline to design and implement the model. The extracted model can help revise data sources and improve data quality requirements, resulting in Data Pipeline improvement to better support the model extraction process. The implemented model is subjected to **Model Refinement**, where the model's parameters are adjusted to ensure it accurately represents the real-world system as it is intended to simulate. Finally, **Model Validation** is the process of assessing how well the extracted model reflects the real system's behavior considering predefined simulation goals. Further, we elaborate on each component of the framework with a focus on the model extraction from data, encircled in green in Figure 1.

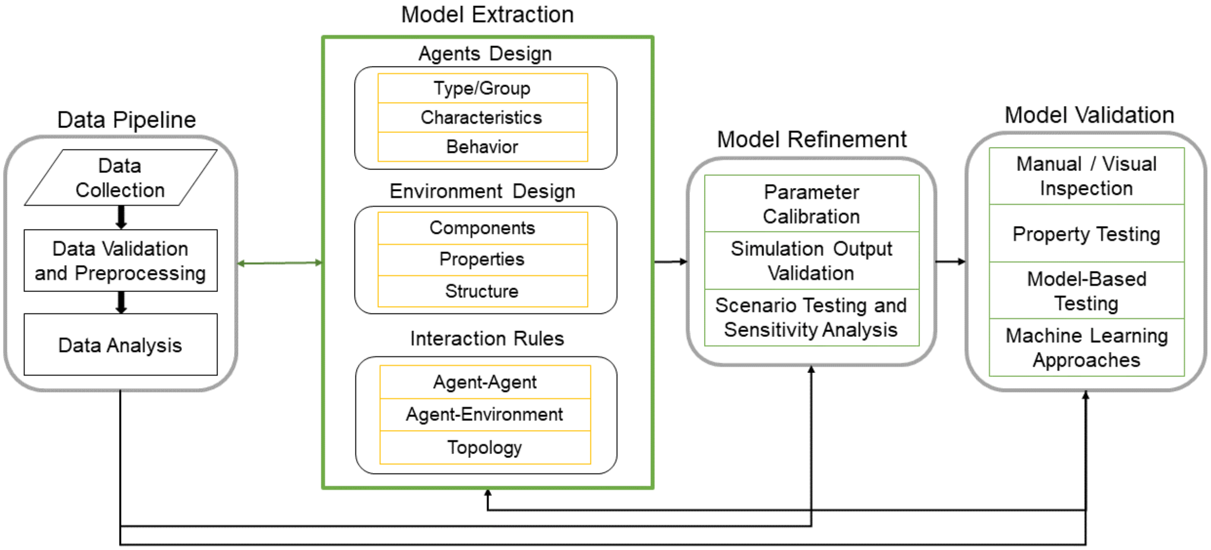

Figure 1: An overview of the framework for developing data-driven agent-based models.

3.1 Data Pipeline

The Data Pipeline serves as the foundation for collecting, validating, preprocessing, and analyzing relevant data before utilizing it in the other three components of our framework. The Data Pipeline consists of three main steps: data collection, data validation, and data analysis, elaborated as follows:

Data Collection: Understanding a system or process begins with data collection, which involves gathering numerical, categorical, text, or image data depending on the characteristics of the system or process under study. Data sources are diverse and can include databases, sensors, user inputs, or even systematic generation. Developing agent-based models from data can be challenging as the data needed for bottom-up behavior is not traditionally collected and may be less efficient to obtain than data for top-down techniques (McAllister et al. 2005). Additionally, collected data often lacks elements such as time scales, behavioral properties, and descriptions properly expressing behavioral actions. Subsequently, the information gathered often fails to clearly express the feedback relationships between agents and components of the environment. Therefore, defining and categorizing necessary data types is crucial for creating relevant models and identifying behavioral patterns. This approach streamlines data collection, ensures a comprehensive record of the data, saves project resources, and simplifies model creation (Altaweel et al. 2010). Altaweel et al. (2010) presented a framework to guide the determination of data requirements, structure data needs, and enable data collection for ABM.

Data Validation and Preprocessing: Collected data often requires validation and preprocessing. The quality of the derived model relies on the quality and validity of the data, which demonstrates the importance of data validation (Bokrantz et al. 2018). Besides data validation, suitable data preprocessing can improve the performance of data-driven models (Wu et al. 2009). Data preprocessing is the transformation of the validated data into a format or structure that is conducive to analysis and model extraction. This step may include handling missing values, removing duplicates, normalization, dimension reduction, or feature selection.

Data Analysis: The data analysis is applying methods and techniques to extract meaningful insights that will inform model extraction, model refinement, or model validation stages. The goal of this step is to understand the underlying patterns and relationships in the data. Depending on the nature of the data and the model's specific requirements, various statistical and machine-learning techniques could be employed for analysing data. According to Abu-Mostafa et al. (2012), there are two types of models for learning from data: Statistical learning models and Machine learning models. Machine learning models "make less restrictive assumptions and deal with more general models than in statistics" (Abu-Mostafa et al. 2012). Following this definition, Kavak (2019) proposed a categorization of model families for each type. Statistical learning models include Markov models, probability distributions, time series forecasting regression models, and probabilistic classifiers. Machine learning models include classifiers, clustering algorithms, and regressor model families. The choice of data analysis method depends on the specific element of the agent-based model being extracted from the data. In the following subsection, we highlight previous efforts in DDABM, focusing on elements of an agent-based model and suggesting data analysis methods applicable to each element's design and development.

3.2 Model Extraction

To identify the elements and sub-elements of an agent-based model, we conducted a literature review on applications of DDABM, summarized in Table 1. Based on our literature review, agent-based models have three core elements: **agents**, **environment**, and **interaction rules**. Subsequently, we describe relevant practical approaches to design and extract each of these elements and their sub-elements from data.

3.2.1 Agents Design

"An agent is identifiable, a discrete individual with a set of characteristics and rules governing its behaviors and decision-making capability" (Macal and North 2005). Kavak (2019) provided a data-driven approach for modeling agents where the author divided their approach into four main steps: data preparation, attribute

Table 1: Overview of DDABM approaches, viewed through our framework from Figure 1.

Model Element	Sub-element	Data-driven extraction approach	Field of Study	Summary	Reference
Agents	Group and characteristic	Descriptive statistics and classification	Transport	This paper introduces a DDABM framework for predicting electric vehicle market penetration and GHG emissions reduction under various incentive scenarios.	Zhan et al. (2024)
	Group and behavior	Classification	Transport	Presents a data-driven agent-based model to simulate the effects of policy interventions and social influence on consumers' transport mode preferences.	Wolf et al. (2015)
	Characteristic and behavior	Descriptive statistics	Disease outbreaks	Presents a data-driven agent-based model to simulate infectious disease in Irish towns.	Hunter et al. (2018)
	Behavior	Machine learning	Renewable energy	Presents DDABM framework, where agent's behavior model is learned through machine learning techniques.	Zhang et al. (2016)
	Behavior	Machine learning (Clustering)	Economics	Explores impact of reputation mechanisms on agents' explorative capabilities in uncertain environments.	Boero et al. (2010)
Environment	Component and characteristic	Geospatial Data Integration	Disease outbreaks	Presents a data-driven agent-based model to simulate infectious disease in Irish towns.	Hunter et al. (2018)
	Component, characteristic, and structure	Graph models and descriptive statistics	Transport	Presents a model that extracts road network information from available data sources for traffic simulation.	Zilske et al. (2011)
	Component, characteristic, and structure	Spatial analysis and descriptive statistic	Social science	Presents a data-driven agent-based model to predict crime patterns in an urban environment, highlighting the importance of employing available data.	Rosés et al. (2021)
	Characteristic and structure	Spatial analysis and classification	Environmental management	Develop a data-driven agent-based model of a forest to simulate the impact of the growing rural population on the forests and wildlife habitat.	An et al. (2005)
Interaction Rules	Agent-agent	Descriptive analysis	Transport	Presents a data-driven agent-based model to simulate the effects of policy interventions and social influence on consumers' transport mode preferences.	Wolf et al. (2015)
	Agent-agent and agent-environment	Statistical analysis	Transport	Introduces a DDABM framework for predicting electric vehicle market penetration and GHG emissions reduction under various incentive scenarios.	Zhan et al. (2024)
	Agent-agent and Topology	Spatial analysis and classification	Environmental management	Develops a data-driven agent-based model of a forest to simulate the impact of the growing rural population on the forests and wildlife habitat.	An et al. (2005)

model creation, behavior model creation, and integration. Building on this definition and our literature review, we divide agent design into three main parts: type/group, characteristics, and behaviors. Further, we explain how each of these parts can be designed using data-driven approaches.

Type/group: Classification of agents in types or groups depends on factors such as their roles, functions, behaviors, or other distinguishing attributes. Typically, model definition and target groups provide a clear indication of agents' types or groups. For example, in an agent-based model of a pandemic, agents are grouped into categories such as healthy, infected, and recovered. However, in some cases, agents' types are not obvious. For instance, Wolf et al. (2015) needed to group the inhabitants of Berlin based on their information for developing a model, where each group indicates the types of transport use. In this case, the authors applied classification methods on survey data to classify the inhabitants. In addition to classification methods, clustering is also practical, especially when there is no prior information to classify agents. For example, Saadat et al. (2018) employed clustering to find agents' types while developing an agent-based model for simulating large-scale usage trends of the GitHub collaborative development tool.

Characteristic: Agent characteristics are divided into static and dynamic characteristics. Static characteristics remain constant during the simulation, such as agents' roles or initial states. Descriptive statistics are useful to derive static characteristics from data. For example, Hunter et al. (2018) used data to determine age and gender breakdowns of populations in their infectious disease outbreak model. Dynamic characteristics, on the other hand, change over time based on the agent's interactions in the system. Descriptive statistics can also help to extract dynamic characteristics of agents from data. For instance, Rai and Robinson (2015) matched agents' attributes representing their beliefs to available survey data of population-wide household-level of Austin in their model. Sajjad et al. (2016a) also used descriptive statistics from Korean census data to calculate agent disposable income characteristics. Time-series analysis or state-transition models are also practical methods for analyzing dynamic characteristics from data. For example, Monti et al. (2023) employed time series analysis to improve agent income estimation in their agent-based model of the housing market.

Behavior: Agents' behaviors encompass their goals, actions, and decision-making processes. An agent's goal is what it aims to achieve. Depending on the model's subject, different types of data-driven methods can help to derive agents' goals from the data. Braubach et al. (2005) provided a generic representation of goal properties for agents representing humans. In some cases, identifying goals from available data is relatively straightforward. For example, Luo et al. (2018) developed an agent-based model of crowd movement using visual data, simply assigning agents' goals to their respective locations in the final frame of the data. When there is a common goal among agents, clustering algorithms can group agents based on their historical behaviors, and these clusters can then be interpreted as their goals. If the goals of agents are already known and labeled in the historical data, supervised learning algorithms can predict the goals of new agents. When the goals can be quantified, regression analysis can be used to predict these quantities based on other characteristics of the agents. However, in some cases, finding agents' goals from data might be challenging. If the goals of the agents involve a sequence of actions (like in a game or a plan), hidden Markov model (HMM), sequence analysis, or process mining techniques can be used to learn these sequences from historical data. For instance, Rai and Hu (2013) in their model of smart office feed sensor data in real-time to the HMM, and the output is the recognized behavior patterns that can be interpreted as the agent's goals in the model. Recently Kasumba et al. (2024) introduced a data-driven approach for goal recognition design that can account for agents with general behavioral models.

Agents' actions are activities they perform to achieve their goals. In an agent-based model, these actions represent the behaviors or steps agents take within the environment. Typically, actions can be derived by observing available data. For example, in a traffic agent-based model, an agent's (car) action could be moving forward, turning, or stopping. It is important to identify the possible actions and the conditions under which these actions occur, which can be derived by analyzing historical data. For instance, Hassan et al. (2010) employed data from the Spanish census sample of the 1980 survey to estimate action probabilities and distributions, such as age-related probabilities of having children, regression equations to determine whether an agent searches for a partner or not, and the fertility rate.

Agent's decision-making process is the logic or strategy it uses to decide what action to take at each step of the simulation. Decision-making is typically based on the agent's current state, the state of the environment, and the agent's goals. For example, in a traffic agent-based model, a car might decide which way to turn at an intersection based on its destination, the current traffic conditions, and the traffic rules. DeAngelis and Diaz (2019) provided an overview of different methods for integrating decisions into ABM. To derive decision-making processes, we need to capture how agents make decisions in the real world. Data-driven methodologies for replicating decision-making process of agents can be divided into interpretable methods and black-box methods. Interpretable methods, like decision trees, facilitate the extraction and understanding of agents' decision-making rules from data. In contrast, Black-box methods like support vector machines, can capture and replicate more complex decision-making rules but are less suitable for understanding the underlying rules. In our previous work (Jamali et al. 2024), we employed both categories to derive decision-making rules by experimenting with Schelling's model variations to demonstrate the advantages and disadvantages of each method. In some models, when authors can formulate a mathematical equation to represent the decision-making rule of agents, they derive the component of the equation from data. For instance, Rai and Robinson (2015) modeled agents' decisions on using solar solutions with an equation consisting of three main components: financial (covers payback period and net monthly electricity bill savings), environmental (overall environmental concern), and agent's social belief derived from survey data.

3.2.2 Environment Design

The environment in agent-based models is the context or space in which the agents operate. We can define the environment through its components, characteristics, and structure.

Component: The environment in an agent-based model can have physical or non-physical components, which can be identified using data-driven methods. Spatial analysis uncovers the physical layout and detects patterns in environmental features. For example, Rosés et al. (2021) based the spatial layer of their model's environment from location-based social networks data, taxi trip data, weather data, land use information, population density, and points of interest. To identify non-physical components, we can employ time-series analysis, machine learning, or network analysis. Network analysis reveals the relationships between different components, like connections in a transportation network. In the urban crime model by Rosés et al. (2021), the temporal layer of the environment (which is a proxy for the presence of victims on a given day with specific weather characteristics) was derived from data on the day of the week and various weather conditions.

Characteristics: Environment characteristics are divided into static and dynamic, similar to agent characteristics. Therefore, both agents and their environments can be modeled using similar data-driven techniques but applied to different types of data. For example, An et al. (2005) derived environment characteristics in their model, such as spatially varying forest volume and growth rate, from data of the environmental system.

Structure: The structure of the environment includes both spatial and time step structures, which can be derived from data. The spatial structure refers to the physical layout of the environment. Spatial analysis can be used to capture this layout. For example, Rosés et al. (2021) used data both to identify the components of the environment and determine their layout in the model.

3.2.3 Interaction Rules

In addition to agents' behaviors, which concerns agents' decision-making rules for selecting actions toward reaching their goals, agents can affect each other and the environment. The rules that define these interactions are known in ABM as interaction rules. Interaction rules are derived from theoretical principles, empirical data, or a combination of both. Here, we categorize interaction rules into three main sub-elements: how agents interact with each other (agent-agent interactions), with their environment (agent-environment interactions), and the structure or topology of these interactions.

Agent-agent: Agent-agent interaction rules govern how agents interact with each other in the model. Wolf et al. (2015) derived agent-agent interaction rules based on their survey respondents. Based on their descriptive analysis, they defined a sociodemographic coordinate system for agents. In their model, the likelihood that two agents communicate with each other is a function of sociodemographic coordinates and the distance of agents. In the model presented by Rai and Robinson (2015), the attitude evolution process is the result of agent-agent interactions where they used survey respondent data to define the basis of the interaction rules between neighbors that can affect their decision on adopting solar solutions. Statistical analysis can also be utilized to identify correlations and dependencies between different variables in the data that can help us capture how agents affect each other in the data. For instance, a strong correlation between the behaviors of two agents might suggest a rule where the behavior of one agent influences that of the other.

Agent-environment: Agent-environment interaction rules describe how agents interact with the environment. Sensitivity analysis can assist in understanding which environmental factors have the most significant impact on agents' behaviors, thereby helping to derive agent-environment interaction rules. The study by Zhan et al. (2024) is an example of capturing agent-agent and agent-environment rules from data by estimating the coefficients associated with the interaction rules using statistical data fitting. Spatial analysis techniques can analyze data in order to find the spatial distribution of agents and their interactions with the environment. For example, data indicating a higher density of agents in certain areas of the environment suggests an interaction rule where agents are attracted to these areas. Techniques such as autocorrelation, cross-correlation, and spectral analysis can identify patterns and derive interaction rules.

For instance, if there is a strong temporal correlation between an environmental variable and a variable describing agent's behavior, it may indicate that the agent's behavior is influenced by this variable.

Topology: Topology of interactions describes the structure of connections among agents in agent-based models. An et al. (2005) employed available data to find the spatial fuelwood model interaction topology. In their model, the interaction indicates how fuelwood collectors from the household interact with the environment in the model. In another study, Rai and Robinson (2015) used actual agent locations and distances from other agents to generate the networks in their agent-based model of technology adoption.

3.3 Model Refinement

Model Refinement refers to the processes of parameter calibration, simulation output validation, and scenario testing/sensitivity analysis. Parameter calibration aligns the model with empirical data to ensure its representativeness and applicability in real-world systems. Traditional methods for calibration often involve a manual process where parameters are adjusted iteratively based on the difference between the simulation output and the observed data. This process can be time-consuming and computationally intensive, especially for complex models with a large number of parameters. However, recent advancements in computational techniques have led to the development of more sophisticated and efficient calibration methods. E.g., Kim et al. (2021) developed a framework for the automatic calibration of dynamic and heterogeneous parameters in agent-based models. In another work, Lamperti et al. (2018) proposed an approach that utilizes machine learning surrogates for calibrating agent-based models. The process of simulation output validation is minimizing the difference between the model's output and the observed real-world data (Chen and Desiderio 2022a; Chen and Desiderio 2022b). Methods like regression-based calibration (Chen and Desiderio 2022b), meta-modeling nonparametric regression (Chen and Desiderio 2022a), and gradient-assisted calibration (Dyer et al. 2023) calibrate an agent-based model by employing simulation output validation technique. Scenario testing and sensitivity analysis are the last approaches used to calibrate agent-based models. Scenario Testing involves creating hypothetical situations or 'scenarios' to evaluate how well the agent-based model performs (Clark et al. 2021). These scenarios can be extreme or mild variations of input parameters to understand their effects on model outcomes. For example, Zhang et al. (2016) calibrated their model based on a random sample of available data where each data can be considered as a scenario. Finally, sensitivity analysis is a systematic method to determine how different values of an independent variable impact a particular dependent variable under a given set of assumptions (Niida et al. 2019).

3.4 Model Validation

Model validation and model refinement components are entangled as sometimes validating concludes by calibrating the model, which implies adjusting parameters within the current model structure. However, sometimes model validation requires partially changing the model, which means modifying the model's structure or adding a new component to it. Model validation component in the development of data-driven agent-based models ensures the reliability and accuracy of simulation outcomes. In our framework, we consider four main strategies for model validation as presented by Hua et al. (2022): manual/visual inspection, property testing, model-based testing, and machine learning approaches. Manual/visual inspection strategies refer to visual inspection of the agent-based model to verify its accuracy. Despite being a necessary strategy, a manual inspection can be tedious and prone to errors, and there will be a risk of overlooking certain aspects. However, manual inspection strategies can be practical, as Kumaresan et al. (2023) demonstrated. In their agent-based model for COVID-19 case forecasting, they employed manual and automated parameter fitting approaches to fit and validate their model with real-world data. In another study, Bemthuis and Lazarova-Molnar (2023) presented and approach for face validity assessment of agent-based models by extracting information from event logs generated by both real-life processes and simulation models. The second strategy is property-based testing, which can be used for testing various important properties of an agent-based model (Thaler and Siebers 2020). By developing software to run this strategy automatically, property-based testing can be faster than manual inspection. However, they can still be time-consuming if a large number of properties are being tested for an agent-based model. Thaler and Siebers (2019) categorize property-based testing into verification of models with real-world data and exploratory nature inspired by real-world phenomena while exploring both ways. Model-based testing is interacting with the agent-based model either online or offline to test certain properties and confirm their correctness. In the context of agent-based models, model-based testing can be seen as a method that automatically generates test cases from a model representing the system behavior and executes these cases to validate system requirements and consistency (Clark et al. 2021). For instance, Monti et al. (2023) employed time series analysis to calculate the R-squared coefficient that represents goodness of fit between learned time series and ground truth ones; to validate their model output. Machine learning algorithms can also learn patterns from large datasets and make predictions, which can be used to validate the outputs of agent-based models (Bonabeau 2002).

4 CASE STUDIES

In this section, we illustrate our proposed framework using two studies that considered different aspects of the ABM components. In both of our examples, the goal was to enable automation and continuous model updates as new data became available.

Our first study aimed to explore data-driven methods for reconstructing variations of Schelling's models (Jamali et al. 2024). The primary focus was on extracting the parameters and agent's decision-making logic from synthetic data generated by an original Schelling agent-based model. Our experiments confirmed that a decision tree could accurately determine agents' decision-making parameters in two variations of Schelling's model, even when high levels of irregular behavior were introduced among the agents. Irregular behavior refers to agents exhibiting random behaviors for a certain percentage of time. We specifically choose a decision tree as it is an interpretable model. Afterward, we introduced another variation of Schelling's model where the agent has a complicated decision-making policy. This time, we employed decision trees and support vector machines to compare their performance in capturing the decision-making logic of agents from data generated by the model. We choose these two ML methods as they are representative of interpretable and black-box models, respectively. We observed that their performance highly depends on how we are giving them the input features, which indicates the importance of the data pipeline in our framework. We observed that in the case that we used the most informative features, support vector machine (SVM) was able to perform better than the decision tree in capturing agents' decision-making logic, which demonstrates the ability of black-box models to capture complex patterns.

In two separate studies, we utilized data-driven methodologies to analyze decision-making patterns in the Danish pharmaceutical market, specifically focusing on parallel imported medicines (Jamali and Lazarova-Molnar 2023; Jamali and Lazarova-Molnar 2024). The first study concentrated on the top ten parallel-imported medicines in terms of sales quantity. We divided each company's pricing time series into subsequences and applied a time-series subsequence clustering method to identify competitors' pricing patterns. Focusing on two popular medicines in the market allowed us to observe market pricing patterns in a detailed case study. In the second study, we expanded our scope to include pricing information for over 400 parallel-imported medicines. We applied three distinct data analysis procedures to the historical pricing data. The first procedure involved using descriptive statistics to identify general characteristics of the competitors' pricing behavior. The second procedure was time series clustering, similar to the first study, but with an automated process for creating the subsequences. The third and final procedure was the visual representation of the competitors' pricing behavior. Employing these procedures, we were able to observe pricing patterns related to subgroups of competitors in the market. Moreover, the results demonstrated that competitors' pricing decisions are not only based on their current state in the model, and there are some latent factors that are affecting their decisions (e.g., the expiry date of medicines could be one of the factors that affect pricing). This complexity suggests that the decision-making process of the competitors is intricate, and developing an agent-based model that accurately represents their behavior may require additional information.

5 SUMMARY AND OUTLOOK

In this paper, we reviewed existing literature on data-driven agent-based modeling to develop a framework that systematically approaches data-driven agent-based modeling. This framework consists of four main components, summarized as follows. The Data Pipeline component processes and prepares data for model extraction, refinement, and validation. The core of our proposed framework is the Model Extraction component, where the agent-based model is extracted using processed data. The extracted models provide feedback to revise data sources and corresponding data quality, improving the Data Pipeline for improved extraction processes. After the data-driven extraction, models undergo Refinement, which involves fine-tuning models to accurately reflect corresponding real-world systems (reflected in the Model Refinement component). The fourth component of our framework is Model Validation, which evaluates to what extent the extracted model reflects the real system's behavior with respect to predefined simulation goals. We, furthermore, performed a comprehensive exploration of data-driven methods that have been or can be applied in each component of the model development framework.

Our proposed framework provides a foundation for future advancements in data-driven agent-based modeling. Its comprehensive design makes it adaptable to a wide range of scenarios and applications. The framework introduces a systematic approach to data-driven agent-based model development, functioning also as a checklist that can guide the process. This would ensure incorporation of the essential data-driven components, covering all aspects of the model. Given the wide range of applications of agent-based modeling, the framework can be further customized and developed.

In this work, we primarily focused on the contribution of data to the model extraction phase. However, in future, we aim to extend our focus to include data's role in model refinement and validation phases. Such an extension would make this work more comprehensive and holistic.

ACKNOWLEDGMENTS

This work is partly funded by the Innovation Fund Denmark (IFD) under File No. 9065-00207B.

REFERENCES

Abu-Mostafa, Y. S., M. Magdon-Ismail, and H.-T. Lin. 2012. *Learning from data*, Volume 4. AMLBook New York.

Altaweel, M. R., L. N. Alessa, A. Kliskey, and C. Bone. 2010. "A framework to structure agent-based modeling data for social-ecological systems". *Structure and Dynamics* 4(1).

An, L., M. Linderman, J. Qi, A. Shortridge and J. Liu. 2005. "Exploring complexity in a human-environment system: an agent-based spatial model for multidisciplinary and multiscale integration". *Annals of the association of American geographers* 95(1):54–79.

Bemthuis, R. and S. Lazarova-Molnar. 2023. "An approach for face validity assessment of agent-based simulation models through outlier detection with process mining". In *International Conference on Enterprise Design, Operations, and Computing*, 134–151. Springer.

Boero, R., G. Bravo, M. Castellani, F. Squazzoni *et al*. 2010. "Why bother with what others tell you? An experimental data-driven agent-based model". *Journal of Artificial Societies and Social Simulation* 13(3):6.

Bokrantz, J., A. Skoogh, D. Lämkull, A. Hanna and T. Perera. 2018. "Data quality problems in discrete event simulation of manufacturing operations". *Simulation* 94(11):1009–1025.

Bonabeau, E. 2002. "Agent-based modeling: Methods and techniques for simulating human systems". *Proceedings of the national academy of sciences* 99(suppl_3):7280–7287.

Braubach, L., A. Pokahr, D. Moldt, and W. Lamersdorf. 2005. "Goal representation for BDI agent systems". In *Programming Multi-Agent Systems: Second International Workshop ProMAS 2004, New York, NY, USA, July 20, 2004, Selected Revised and Invited Papers 2*, 44–65. Springer.

Castiglione, F. 2020. "Agent-based modeling and simulation, introduction to". *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*:661–665.

Chen, S. and S. Desiderio. 2022a. "Calibration of agent-based models by means of meta-modeling and nonparametric regression". *Computational Economics* 60(4):1457–1478.

Chen, S. and S. Desiderio. 2022b. "A regression-based calibration method for agent-based models". *Computational Economics* 59(2):687–700.

Clark, A. G., N. Walkinshaw, and R. M. Hierons. 2021. “Test case generation for agent-based models: A systematic literature review”. *Information and Software Technology* 135:106567.

DeAngelis, D. L. and S. G. Diaz. 2019. “Decision-making in agent-based modeling: A current review and future prospectus”. *Frontiers in Ecology and Evolution* 6:237.

Dehkordi, M. A. E., J. Lechner, A. Ghorbani, I. Nikolic, E. Chappin and P. Herder. 2023. “Using machine learning for agent specifications in agent-based models and simulations: A critical review and guidelines”. *Journal of Artificial Societies and Social Simulation* 26(1):9.

Drchal, J., M. Čertický, and M. Jakob. 2016. “Data driven validation framework for multi-agent activity-based models”. In *Multi-Agent Based Simulation XVI: International Workshop, MABS 2015, Istanbul, Turkey, May 5, 2015, Revised Selected Papers 16*, 55–67. Springer.

Dyer, J., A. Quera-Bofarull, A. Chopra, J. D. Farmer, A. Calinescu and M. Wooldridge. 2023. “Gradient-assisted calibration for financial agent-based models”. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, 288–296.

Epstein, J. M. and R. Axtell. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

Hassan, S., J. Pavón, L. Antunes, and N. Gilbert. 2010. “Injecting data into agent-based simulation”. In *Simulating Interacting Agents and Social Phenomena: The Second World Congress*, 177–191. Springer.

Hua, E. Y., S. Lazarova-Molnar, and D. P. Francis. 2022. “Validation of digital twins: challenges and opportunities”. In *2022 Winter Simulation Conference (WSC)*, 2900–2911 https://doi.org/10.1109/WSC57314.2022.10015420.

Hunter, E., B. Mac Namee, and J. Kelleher. 2018. “An open-data-driven agent-based model to simulate infectious disease outbreaks”. *PloS one* 13(12):e0208775.

Jamali, R. and S. Lazarova-Molnar. 2023. “Uncovering Competitor Pricing Patterns in the Danish Pharmaceutical Market via Subsequence Time Series Clustering: A Case Study”. In *2023 Winter Simulation Conference (WSC)*, 793–804 https://doi.org/10.1109/WSC60868.2023.10407784.

Jamali, R. and S. Lazarova-Molnar. 2024. “Data-Driven Insights for Agent-Based Modeling: Discovering Patterns to Model Price Competition in the Danish Pharmaceutical Market”. In *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: International Workshops of PAAMS 2024*. Springer.

Jamali, R., W. Vermeiren, and S. Lazarova-Molnar. 2024. “Data-driven Agent-based Modeling: Experimenting with the Schelling’s Model”. *Procedia Computer Science*.

Kasumba, R., G. Yu, C.-J. Ho, S. Keren and W. Yeoh. 2024. “Data-Driven Goal Recognition Design for General Behavioral Agents”. *arXiv preprint arXiv:2404.03054*.

Kavak, H. 2019. *A data-driven approach for modeling agents*. Ph.D. dissertation, Old Dominion University. https://digitalcommons.odu.edu/msve_etds/49/, accessed 15th April 2024.

Kavak, H., J. J. Padilla, C. J. Lynch, and S. Y. Diallo. 2018. “Big data, agents, and machine learning: towards a data-driven agent-based modeling approach”. In *Proceedings of the Annual Simulation Symposium*, 1–12.

Kim, D., T.-S. Yun, I.-C. Moon, and J. W. Bae. 2021. “Automatic calibration of dynamic and heterogeneous parameters in agent-based models”. *Autonomous Agents and Multi-Agent Systems* 35(2):46.

Kumaresan, V., N. Balachandar, S. F. Poole, L. J. Myers, P. Varghese, V. Washington, *et al*. 2023. “Fitting and validation of an agent-based model for COVID-19 case forecasting in workplaces and universities”. *Plos one* 18(3):e0283517.

Lamperti, F., A. Roventini, and A. Sani. 2018. “Agent-based model calibration using machine learning surrogates”. *Journal of Economic Dynamics and Control* 90:366–389.

Lazarova-Molnar, S. and X. Li. 2019. “Deriving simulation models from data: steps of simulation studies revisited”. In *2019 Winter Simulation Conference (WSC)*, 2771–2782 https://doi.org/10.1109/WSC40007.2019.9004697.

Luo, L., C. Chai, J. Ma, S. Zhou and W. Cai. 2018. “ProactiveCrowd: Modelling Proactive Steering Behaviours for Agent-Based Crowd Simulation”. In *Computer Graphics Forum*, Volume 37, 375–388. Wiley Online Library.

Macal, C. and M. North. 2014. “Introductory tutorial: Agent-based modeling and simulation”. In *Proceedings of the winter simulation conference 2014*, 6–20 https://doi.org/10.1109/WSC.2014.7019874.

Macal, C. M. and M. J. North. 2005. “Tutorial on agent-based modeling and simulation”. In *Proceedings of the Winter Simulation Conference, 2005.*, 14–pp https://doi.org/10.1109/WSC.2005.1574234.

McAllister, R., I. Gordon, and C. Stokes. 2005. “KinModel: An agent-based model of rangeland kinship networks”. In *International Congress on Modelling and Simulation (Modelling and Simulation Society of Australia and New Zealand, MODSIM)*, 170–176. Citeseer.

Monti, C., M. Pangallo, G. De Francisci Morales, and F. Bonchi. 2023. “On learning agent-based models from data”. *Scientific Reports* 13(1):9268.

Niida, A., T. Hasegawa, and S. Miyano. 2019. “Sensitivity analysis of agent-based simulation utilizing massively parallel computation and interactive data visualization”. *PloS one* 14(3):e0210678.

Patsatzis, D. G., L. Russo, I. G. Kevrekidis, and C. Siettos. 2023. “Data-driven control of agent-based models: An equation/variable-free machine learning approach”. *Journal of Computational Physics* 478:111953.

Quera-Bofarull, A., J. Dyer, A. Calinescu, and M. Wooldridge. 2023. "Some challenges of calibrating differentiable agent-based models". *arXiv preprint arXiv:2307.01085*.

Rai, S. and X. Hu. 2013. "Behavior pattern detection for data assimilation in agent-based simulation of smart environments". In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Volume 2, 171–178. IEEE.

Rai, V. and S. A. Robinson. 2015. "Agent-based modeling of energy technology adoption: Empirical integration of social, behavioral, economic, and environmental factors". *Environmental Modelling & Software* 70:163–177.

Ravaioli, G., T. Domingos, and R. F. Teixeira. 2023. "A framework for data-driven agent-based modelling of agricultural land use". *Land* 12(4):756.

Rosés, R., C. Kadar, and N. Malleson. 2021. "A data-driven agent-based simulation to predict crime patterns in an urban environment". *Computers, Environment and Urban Systems* 89:101660.

Saadat, S., C. Gunaratne, N. Baral, G. Sukthankar and I. Garibay. 2018. "Initializing agent-based models with clustering archetypes". In *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11*, 233–239. Springer.

Sajjad, M., K. Singh, E. Paik, and C.-W. Ahn. 2016a. "A data-driven approach for agent-based modeling: Simulating the dynamics of family formation". *Journal of Artificial Societies and Social Simulation* 19(1):9.

Sajjad, M., K. Singh, E. Paik, and C.-W. Ahn. 2016b. "Social simulation: The need of data-driven agent-based modelling approach". In *2016 18th International Conference on Advanced Communication Technology (ICACT)*, 818–821. IEEE.

Steinbacher, M., M. Raddant, F. Karimi, E. Camacho Cuena, S. Alfarano, G. Iori *et al*. 2021. "Advances in the agent-based modeling of economic and social behavior". *SN Business & Economics* 1(7):99.

Thaler, J. and P.-O. Siebers. 2019. "Show me your properties: the potential of property-based testing in agent-based simulation". In *Proceedings of the 2019 Summer Simulation Conference*, 1–12.

Thaler, J. and P.-O. Siebers. 2020. "Specification testing of agent-based simulation using property-based testing". *Autonomous Agents and Multi-Agent Systems* 34(2):47.

Truong, T. M., F. Amblard, B. Gaudou, and C. S. Blanc. 2016. "Cfbm-a framework for data driven approach in agent-based modeling and simulation". In *Nature of Computation and Communication: Second International Conference, ICTCC 2016, Rach Gia, Vietnam, March 17-18, 2016, Revised Selected Papers 2*, 264–275. Springer.

Wolf, I., T. Schroeder, J. Neumann, and G. de Haan. 2015. "Changing minds about electric cars: An empirically grounded agent-based modeling approach". *Technological forecasting and social change* 94:269–285.

Wu, C.-L., K.-W. Chau, and Y.-S. Li. 2009. "Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques". *Water Resources Research* 45(8).

Yang, L. and K. H. van Dam. 2022. "Data-Driven Agent-Based Model Development to Support Human-Centric Transit-Oriented Design". In *International Conference on Autonomous Agents and Multiagent Systems*, 60–66. Springer.

Zhan, W., Z. Wang, J. Deng, P. Liu and D. Cui. 2024. "Integrating System Dynamics and Agent-Based Modeling: A Data-Driven Framework for Predicting Electric Vehicle Market Penetration and GHG Emissions Reduction Under Various Incentives Scenarios". *Available at SSRN 4674454*.

Zhang, H., Y. Vorobeychik, J. Letchford, and K. Lakkaraju. 2016. "Data-driven agent-based modeling, with application to rooftop solar adoption". *Autonomous Agents and Multi-Agent Systems* 30:1023–1049.

Zhou, Y., X. C. Liu, B. Chen, T. Grubesic, R. Wei and D. Wallace. 2024. "A data-driven framework for agent-based modeling of vehicular travel using publicly available data". *Computers, Environment and Urban Systems* 110:102095.

Zilske, M., A. Neumann, and K. Nagel. 2011. "OpenStreetMap for traffic simulation". Retrieved from https://depositonce.tu-berlin.de/bitstream/11303/4976/2/zilske_neumann_nagel.pdf, accessed 15th March 2024.

AUTHOR BIOGRAPHIES

RUHOLLAH JAMALI is a PhD student in the SDU Software Engineering Section of the Faculty of Engineering at the University of Southern Denmark. His research interests concern Modeling, Data Analysis, and Simulation, especially as applied in developing decision support systems. His email address is ruja@mmmi.sdu.dk.

SANJA LAZAROVA-MOLNAR holds two full professorships at the Karlsruhe Institute of Technology and the University of Southern Denmark. Her research focuses on data-driven simulation, digital twins, and cyber-physical systems modeling for reliability and energy efficiency enhancement. Actively engaged in developing advanced methodologies, she leverages her expertise to optimize complex systems through data-driven simulation. She leads activities focused on digital twins and data-driven simulation in several European and national projects. Furthermore, Professor Lazarova-Molnar assumes leadership roles in IEEE and The Society for Modeling & Simulation International (SCS), contributing significantly to these professional organizations. She was also one of the Proceedings Editors for the Winter Simulation Conference in 2019 and 2020 and an associate editor of SIMULATION: Transactions of The Society for Modeling and Simulation International. Her email address is sanja.lazarova-molnar@kit.edu.