Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

EXPLORING DATA REQUIREMENTS FOR DATA-DRIVEN AGENT-BASED MODELING

Hui Min Lee,[1] and Sanja Lazarova-Molnar[1,2]

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

ABSTRACT

Extracting Agent-Based Models (ABMs) from data, also known as Data-Driven Agent-Based Modeling (DDABM), requires a clear understanding of data requirements and their mappings to the corresponding ABM components. DDABM is a relatively new and emerging topic, and as such, there are only highly customized and problem-specific solutions and approaches. In our previous work, we presented a framework for DDABM, identifying the different components of ABMs that can be extracted from data. Building on this, the present study provides a comprehensive analysis of existing DDABM approaches, examining prevailing trends and methodologies, focusing on the mappings between data and ABM components. By synthesizing and comparing different DDABM approaches, we establish explicit mappings that clarify data requirements and their role in enabling DDABM. Our findings enhance the understanding of DDABM and highlight the role of data in automating model extraction, highlighting its potential for advancing data-driven agent-based simulations.

1 INTRODUCTION

Agent-Based Modeling and Simulation (ABMS) has emerged as one of the most effective modeling approaches for simulating Multi-Agent Systems (MASs), which are systems composed of multiple autonomous agents that interact with one another and their environment (Macal and North 2014). Examples of MASs include supply chains (Chaib-draa and Müller 2006), where agents represent suppliers and consumers that make autonomous decisions to optimize logistics; and transportation management systems (Aloui et al. 2024), where agents represent vehicles, infrastructure and traffic management system, each making decisions that collectively influence system-wide behavior. More broadly, MASs are found across multiple domains in real life (Lü et al. 2011).

One approach to building simulation models of MASs is through Data-Driven Agent-Based Modeling (DDABM), which implies deriving agent-based models from real-world data, enabling more informed and realistic simulations (Lee et al. 2025). One of the important applications of data-driven ABMs could be in Digital Twins (DTs), which are virtual representations of physical systems that enable real-time simulation, optimization, and monitoring (Mariani et al. 2022). DTs typically rely on simulation models to replicate behaviors of real-world systems. Compared to traditional Discrete-Event Simulation (DES), ABMS can be a more suitable modeling and simulation approach for developing underlying core models of DTs for certain types of systems, e.g., production systems (Moyaux et al. 2023). This is because ABMs allow for autonomous decision-making by agents, enabling simulations that can capture micro-level behaviors and provide various levels of system abstraction. In contrast, DES relies on predefined events to drive the simulation, limiting its ability to model complex, agent-driven behaviors (Moyaux et al. 2023).

Despite the research by dos Santos et al. (2022) showing that in the past DTs relied more on models built using DES, there has been a growing trend in recent years towards using ABMs as core models of DT simulations (Marah and Challenger 2024; Ambra and Macharis 2020; He and Shen 2025; Stary 2021). This shift is driven by the recent advancements of ABMS tools, increased data availability, and improved computational capabilities that have expanded ABMs' applications across various domains and disciplines (Macal and North 2014; Jamali and Lazarova-Molnar 2024).

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

Consequently, DDABM has emerged as a promising methodology for improving modeling and simulation accuracy, especially for complex systems and social phenomena, by utilizing real data (Sajjad et al. 2016; Monti et al. 2023). Data-driven approaches enable the extraction of ABMs directly from data by identifying ABMs' components (agents, environment, interaction rules), rather than relying solely on modelers' judgement. DDABM has been applied across various disciplines (Lounis and Bagal 2020; Venkatramanan et al. 2018; Yang and van Dam 2022; Lee et al. 2021; Bae et al. 2024). The growing adoption of DDABM highlights its potential beyond general simulation, especially for DTs requiring dynamic updates (Niloofar et al. 2023).

One of the key factors to consider when developing a DT is ensuring it facilitates the automatic generation of simulation models through continuous data collection (Lazarova-Molnar 2024). In our earlier work (Jamali and Lazarova-Molnar 2024), we developed a DDABM framework, where we identified the key components of ABMs that can be extracted from data. However, as demonstrated across different fields and studies, existing approaches in DDABM appear to be problem-specific, lacking a more generalized approach for extracting ABMs from data.

Building on this, our previous study introduced the Multi-Agent Systems' Digital Twins (MASDT) conceptual framework (Lee et al. 2025) for constructing DTs for MASs. In this framework, data-driven ABMs serve as core simulation models of DTs. MASDT conceptualized the dynamic extraction and continuous updating of the ABM-based DT models using real-world data. The framework also establishes a bidirectional feedback loop between the digital and physical systems: data from the real world informs the ABM-based DT, while insights and decisions generated by the DT are fed back into the physical system. This feedback loop supports adaptive and data-informed decision-making.

To enable effective data-driven ABMs, it is essential to have a clear understanding of data requirements and their mapping to the corresponding components in ABMs. This understanding supports the overarching goal of advancing DDABM (Jamali and Lazarova-Molnar 2024) and MASDT (Lee et al. 2025) frameworks, and is essential for operationalizing them in real-world applications. In this study, we conduct a comprehensive review of existing DDABM approaches, systematically identifying their data requirements and mapping them to the corresponding ABM components, including agents, states/attributes, interactions, environment, and decision-making rules. Our goal is to identify prevailing trends and approaches in DDABM, with a particular focus on clarifying how and which data are mapped to the different ABM components. Based on this analysis, we provide recommendations for enhancing DDABM practices of extracting ABMs from data, thereby enabling the development of more accurate ABMs.

The main contribution of this paper is a structured analysis of existing DDABM works, synthesizing data requirements based on ABMs' components, and offering practical guidelines for implementing DDABMs. By analyzing what types of data can be used for which ABM components, our study provides a useful reference for researchers and modelers to integrate data effectively when designing DDABMs.

Our paper is structured as follows: Section 2 provides the background and state-of-the-art developments in DDABM, along with the foundational requirements for ABMs extraction. Section 3 presents the results from our literature review on the data requirements for existing DDABM approaches, summarizing key insights. Section 4 offers our analysis, insights and findings. Lastly, Section 5 concludes the paper and provides recommendations for future research and practice.

2 BACKGROUND & RELATED WORK

2.1 Data-driven Agent-based Modeling

Data-driven simulation is a modeling paradigm that integrates empirical data into simulations to improve model accuracy, relevance, and adaptability. Understanding this broader paradigm is important as DDABM is a specific instance of it, combining data-driven approaches and ABMs. A useful distinction within this paradigm is between Static Data-Driven Simulation (SDDS) and Dynamic Data-Driven Simulation (DDDS) (Niloofar et al. 2023) that helps position current DDABM practices and highlights their limitations.

SDDS typically relies on classical statistical models and utilizes static datasets as input. In contrast, DDDS employs automated systems that continuously update simulation models using real-time data, making it suitable for building DT models. DDABM, as an emerging approach, extends beyond the direct utilization of static datasets to run simulations. It aims to extract key ABM components from data, and ideally, facilitate continuous model updates to reflect real-world dynamics (Jamali and Lazarova-Molnar 2024), which aligns with DDDS paradigm.

However, most existing DDABM approaches rely on static data, aligning more closely with the SDDS paradigm rather than DDDS. For example, most studies in the field of DDABM (Zhang and Tan 2024; Lee et al. 2021; Ravaioli et al. 2023; Sajjad et al. 2016; Bae et al. 2024) are using pre-collected datasets instead of real-time data streams. As a result, these models can only run simulations with the data already available and cannot adapt dynamically to new information.

Additionally, the current landscape of DDABM is characterized by solutions and approaches that are often highly customized and problem specific. For example, the DDABM proposed by Bae et al. (2024) is tailored specifically for simulating the public bicycle-sharing system in Sejong city. Thus, the general DDABM literature remains limited, with no existing studies systematically addressing the data requirements for extracting ABMs from real-world data. Given these limitations, it is essential to examine what types of data are necessary to extract accurate ABMs from real-world datasets. Section 2.2 describes the common data types and practices used in the development of data-driven ABMs.

2.2 Common Data Types and Practices for Data-driven ABMs Development

Different types of data serve different roles in ABM extraction and development, influencing how agents are characterized, environments are represented, and interaction rules are derived. These data types can be categorized based on their properties (qualitative vs. quantitative), level (micro vs. macro), and spatiotemporal dimension (capturing the complexity of location and time). Understanding these differences is key to selecting the appropriate data to extract ABMs' components.

In general, all data can be categorized as either *empirical qualitative* or *empirical quantitative*, based on how it is measured and represented. These data types are common in many studies (Ghorbani et al. 2015; Lu Yang and Gilbert 2008; Paudel and Ligmann-Zielinska 2023) that develop data-driven ABMs. *Empirical quantitative data* are measurable and enable capturing of trends through data; while *empirical qualitative data* are not directly measurable, and they provide valuable insights into behaviors and decision making (Ghorbani et al. 2015). Qualitative data are typically collected through surveys, interviews, observation and experiments (Ghorbani et al. 2015), particularly common in studies of social and behavioral science (Squazzoni and Boero 2005).

Beyond this broad classification, data can also be distinguished by their scales of analysis, particularly in ABMs. Data can be categorized into micro-level and macro-level based on their *scale* and *granularity*, which refers to the level of detail at which data are collected, analyzed, and applied within a model. These two types of data complement each other by providing insights into both detailed individual behavior and aggregated system-wide patterns, making their distinction vital for understanding complex systems.

Micro-level data focuses on individual agent level, such as demographic attributes (age, gender, etc.), behavioral patterns, and decision-making processes. This type of data is essential for simulating individual agents and understanding how their behaviors or decisions contribute to system-wide phenomena (Bruch and Atwell 2015). For example, Walsh et al. (2013) demonstrate the integration of household-level data into ABMs to model land use changes effectively. Micro-level data can be sourced from various sources, including government census data and user databases.

Macro-level data, on the other hand, represents aggregate information at a larger scale, such as system-wide phenomena or emergent patterns resulting from the collective behavior of agents. Examples include national census or population statistic typically obtained from government reports and databases. Often, macro-level data can be derived from the same source as micro-level data by aggregating the individual records. For instance, individual income data (micro) can be aggregated into the average income of a population in a town (macro). The same variable, such as income data, can therefore be used at both levels, depending on the goal of the simulation. The interplay between these two levels of data is fundamental to ABMs, as it allows researchers to explore how individual behaviors contribute to higher-level outcomes while examining how system-wide patterns influence individual actions (Monti et al. 2023).

Temporal data captures time-dependent information, typically including timestamps to indicate when events occur. It is often represented as time-series data, which is essential for capturing the dynamic nature of the system being modeled, as it reflects changes over time (Monti et al. 2023). In DDABM, temporal data enables accurate tracking of trends and patterns in agent behaviors and system states, providing insight into system dynamics. This data type can be obtained from various sources, such as sensors, system logs, transaction records, and other time-stamped datasets. For instance, log data is a common source of temporal data, it records system events with precise timestamps, allowing for tracking changes and behavior over time (Rahman 2014). Temporal data has been effectively utilized in ABM to simulate dynamic human behavior and enhance predictive accuracy, as demonstrated by Flamino et al. (2019).

Spatial data is critical for ABMs that rely on location data or geographical information. This data type is typically represented and stored using Geographic Information Systems (GIS) (Brown et al. 2005). Spatial data can be obtained through various methods, including GPS, satellites, and other technologies. This data type is essential for ABMs that require location as agents' parameter as relevant for the goals of a given simulation study. This can be exemplified by the use of ABMS in epidemiology (Hunter et al. 2018), where human agents' locations influence the spread of disease.

Furthermore, spatial data is often associated with and used together with temporal data, as space and time can be relative (Brown et al. 2005). For instance, a study of urban planning by Tian and Qiao (2014) utilize spatiotemporal data to simulate urban expansion of Guangzhou city over time.

Data preparation is an essential initial step in ABM extraction (Kavak et al. 2018) to ensure the accuracy and reliability of the extracted model. It involves several key processes, including data cleaning, data integration, data transformation, data reduction, and data discretization (Bell and Mgbemena 2018). In our previous study, we proposed a component called the "Data Pipeline" within our DDABM framework to collect, validate, preprocess, and analyze data before it is used for model extraction (Jamali and Lazarova-Molnar 2024). By performing steps such as handling missing values, removing duplicates, and validating the collected data during preprocessing, the robustness and validity of the extracted ABMs can be enhanced.

2.3 Extractable Agent-Based Models Components from Data

ABMs have three primary components extractable in a data-driven manner: agents, environment, and interaction rules, as identified in our previous work (Jamali and Lazarova-Molnar 2024). Each component can be further broken down into sub-components that can be also derived using data-driven approaches. Table 1 summarizes these components and their sub-components, along with brief explanations.

Table 1: Components and sub-components of ABMs.

Component	**Sub-Component**	**Explanation**
Agents	Type/Group	Categories or classifications of agents based on their characteristics in the model
	Characteristics	Static (e.g., age, gender) and dynamic (e.g., income, health status) attributes of agents
	Behaviors	Actions, decision-making processes, and goals of agents within the model
Environment	Components	Physical and non-physical elements that make up the model's setting
	Characteristics	Properties of the environment that can influence agent behavior
	Structure	Spatial and temporal organization of the environment
Interaction Rules	Agent-Agent	Rules governing how agents interact with each other
	Agent-Environment	Rules defining how agents interact with their environment
	Topology	Structure of connections or networks among agents and environmental elements

In most existing studies, only some ABM components are extracted in a data-driven manner, while others are predefined or manually specified. Rather than deriving the entire model from data, researchers typically focus on extracting specific components. For instance, in a crime pattern simulation, Rosés et al. (2021) used machine learning (decision trees) to derive agent-environment interaction rules from spatial and temporal data. Their model incorporated diverse data sources, including location-based social networks, taxi trip data, weather conditions, land use information, population density, and points of interest, to define the environment's spatial layer. This illustrates how real-world data can inform key ABM components.

Building on the background in Section 2, which covers common data types in DDABM and extractable ABM components, Section 3 offers a detailed analysis of the data requirements and their mapping to various components of ABMs.

3 DATA REQUIREMENTS FOR AGENT-BASED MODEL EXTRACTION

Understanding the data requirements for ABM extraction is essential for enabling and advancing data-driven ABMs as underlying models of DTs. To this end, we conducted a comprehensive literature review to compare and analyze various approaches for extracting different ABM components from data. Our study focused on several key properties: application domain, extracted ABM components, extraction approaches, data sources, data types, data collection infrastructure and simulation goal. We also indicated the simulation approach — static (SDDS) or dynamic (DDDS), in the same column as the simulation goal.

We followed a systematic three-step search strategy to identify studies on DDABM. First, we conducted a keyword search on Google Scholar using terms such as "data-driven agent-based model, "agent-based model extraction", and "data-requirement of agent-based model" to identify relevant studies published in the past ten years. Second, we screened titles, keywords, abstracts and conclusions to select relevant studies. Third, the selected papers were analyzed and categorized based on predefined properties.

By analyzing these properties across multiple studies in different application domains, we aim to provide a comprehensive overview of the current state of data-driven ABM extraction. This analysis helps us better understand how different types of data are used to inform the development of ABMs and identify potential areas of improvements in DDABM approaches. We summarized the results of our study in the Table 2 to provide an overview of our findings.

Table 2: Comparative summary of empirical studies using data-driven agent-based model extraction.

Study	**Application Domain**	**Extracted ABM Components**	**Approach of Extraction**	**Data Source**	**Data Type/ Collection Method**	**Simulation Goal / Approach**

Hunter et al. (2018)	Epidemiology	Agents (Characteristics, Behavior)	Not explicitly stated	Census data;
				GIS data;
				School and workplace locations;
				Vaccination data
					Quantitative; Spatial;
					Data Collection method: Databases (Central Statistics Office (CSO) of Ireland)
						Reducing outbreak impact
						-
						Simulating the spread of airborne infectious diseases in Irish towns
						(SDDS)

		Environment (Components, Properties)	GIS analysis	GIS data		
		Interaction Rules (Agent-Env, Agent-Agent)	Inferred as Probabilistic Inference	School and workplace locations		
Bell and Mgbemena (2018)	Business	Agents (Type, Characteristic, Behavior)	Decision Tree (CART)	Dataset provided by a mobile network operator	*Quantitative*; Data Collection method: Not explicitly stated, likely from databases	Retain customers **-** Simulating customer behavior to stay or churn with Telco **(SDDS)**
		Interactions (Agent-Agent)	Social Network Analysis			

Table 2 (Continued)

Study	**Application Area**	**Extracted ABM Components**	**Approach of Extraction**	**Data Source**	**Data Type/ Collection Method**	**Simulation Goal/ Approach**

Bae et al. (2024)	Transportation	Agents (Characteristics, Behavior)	Not explicitly stated, likely Probabilistic Inference; Spatial and Temporal Analysis;
			Agent Behavior Learning from Data
				Interview Data;
				Rental-return Information;
				Member information;
				Station Information;
				Station monitoring History;
				Population Information;
				Road network Information
					Qualitative; Quantitative; Temporal; Spatial;
					Data Collection method: Interview, and others are Not explicitly stated. For demographic/ operational data, it is likely from System logs/ Databases
						Improve user convenience and system utilization
						-
						Simulating user behavior in a public bicycle sharing system
						(SDDS)

		Environment (Components, Properties, Structure)	Not explicitly stated, likely Spatial and Temporal Analysis	GIS data;
				Station locations;
				Road network data		

		Interaction Rules (Agent-Env, Agent-Agent)	Not explicitly stated	Rental-return data; Station monitoring history;Traffic data		
Zhang and Tan (2024)	Transportation	Agents (Characteristics, Behavior)	Not explicitly stated, likely Probabilistic Inference, Spatial and Temporal Analysis, and Agent Behavior Learning from Data	GPS trajectory data of ETs; Charging event logs	*Spatial; Temporal; Quantitative;* Data Collection method: Not explicitly stated, likely from Databases, sensors, System logs	Enhance charging station layout - Simulating traveling and charging behaviors of Electric Taxis (ETs) **(DDDS)**

		Environment (Components, Properties, Structure)	Not explicitly stated, likely Spatial and Temporal Analysis and Traffic analysis	GIS data;
				Traffic data;
				Road network data		

		Interaction Rules (Agent-Env, Agent-Agent)	Probabilistic inference-based rules for charging decisions	GPS trajectory data;
				Charging event logs;
				Traffic data		

Lee et al. (2021)	Disaster Management	Agents (Characteristics, Behavior)	Not explicitly stated	Demographic data;
				Evacuation data; Real-time gas station data;
				Evacuation order data;
				Meteorological data
					Quantitative; Spatial;
					Data Collection method: Not explicitly stated, likely includes government databases, surveys, and real-time data sources
						Improve disaster management strategy
						-
						Simulating evacuation behavior during hurricanes
						(DDDS)

		Environment (Components, Properties, Structure)	Not explicitly stated, likely Spatial and Temporal Analysis for spatial data	Meteorological data; GIS data		

		Interaction Rules (Agent-Env, Agent-Agent)	Decision Field Theory (DFT) and Extended DFT (EDFT)	Survey data;
				Meteorological data;
				Traffic data;
				Evacuation order data;
				Logistic regression results		

Table 2 (Continued)

Study	**Application Area**	**Extracted ABM Components**	**Approach of Extraction**	**Data Source**	**Data Type/ Collection Method**	**Simulation Goal/ Approach**
Ravaioli et al. (2023)	Agriculture	Agents (Characteristics, Behavior)	Machine Learning (ML)	Census; Surveys; Interviews; GIS Data; Remotely Sensed Data	*Qualitative; Quantitative; Spatial;* Data Collection method: depends on implementation	Improve policy assessment **-** Simulating farmers' decisions for land use **(SDDS)**
		Environment (Components, Properties)	GIS and Remote Sensing analysis	GIS Data; Satellite Images; Land use maps		
		Interaction Rules (Agent-Env, Agent-Agent)	Hybrid: ML and spatial analysis	Agent-environment feedback data; Social/spatial proxies		
Sajjad et al. (2016)	Sociology	Agents (Characteristics, Behavior)	Probabilistic modeling	Census Data	*Quantitative; Spatial;* Data Collection method: Not explicitly stated, likely from government databases	Understand family formation dynamics **-** Simulating family formation **(SDDS)**
		Environment (Components, Properties)	Not explicitly stated but likely involves Spatial analysis			
		Interaction Rules (Agent-Env, Agent-Agent)	Probabilistic equation			

Yang and van Dam (2022)	Urban Planning/ Sociology	Agents (Characteristics, Behavior)	Not explicitly stated, but likely involves Agent Behavior Learning	GIS files;
				Population statistics;
				Activity patterns;
				Mode choice parameters
					Quantitative; Spatial;
					Data Collection method: Behavioral surveys, Sensors, and others are not explicitly stated, likely from government databases
						Support urban transport planning
						-
						Simulating activities and behaviors of individuals in an urban environment
						(SDDS)

		Environment (Components, Properties, Structure)	Not explicitly stated, but likely involves Spatial and Temporal Analysis	Road network data; Public space design parameters; Air pollution data		

		Interaction Rules (Agent-Environment)	Not explicitly stated	Walking/driving speed;
				Pedestrian parameters;
				Car-related pollution coefficients		

Rosés et al. (2021)	Sociology	Agents (Characteristics, Behavior)	Machine learning (Decision Tree for decision-making)	NYPD complaint data;
				Census data and street segment data;
				Location-Based Social Network data;
				Taxi trip data;
				Weather data;
				Population density;
				Calls for service
					Quantitative; Spatial; Temporal;
					Not explicitly stated, likely from GIS maintained by government agencies and other organizations
						Support crime reduction strategies
						-
						Simulating crime patterns in an urban environment
						(SDDS)

		Environment (Components, Properties, Structure)	Not explicitly stated, but likely involves Spatial and Temporal Analysis of GIS data	GIS data;
				Land use information;
				Points of interest;
				Public transportation;
				Tree census data;
				School locations		

		Interaction Rules (Agent-Env)	Machine learning (decision tree for spatial probabilities)	Combination of agent and environment data; Spatial and temporal data from GIS layers		

Table 2 provides a detailed overview of the nine data-driven ABM approaches we analyzed. These approaches were selected to represent some of the few available contributions in the emerging field of DDABM. After careful screening, we considered them offer valuable insights as they describe the data sources used for ABM extraction and the data extraction approaches. In the following section, we offer a narrative summary of the key findings, and our insights derived from this analysis.

4 MAIN FINDINGS OF THE STUDY

Our review of approaches for extracting data-driven ABM components and their data requirements revealed several patterns and insights. These findings reflect the current state of DDABM and have important implications for developing DTs for MASs using data-driven ABMs. In this section, we present our key observations and discuss the potential of data-driven ABMs as core models for DTs of MASs.

4.1 General Observations

Our analysis revealed that while the interest in DDABM is growing, the field remains relatively new with relatively few studies compared to other established modeling approaches. We observed considerable variability in the reviewed literature, with some studies clearly describing data sources and methods, while others lack transparency in their methodologies. A few studies proposed frameworks to develop data-driven ABMs for specific domains. For instance, Ravaioli et al. (2023) introduced a framework for simulating farmers' land-use decisions. As noted in the background, existing frameworks and studies are highly domain- or problem-specific, so data types and methodologies cannot be directly applied to simulations in other areas. This emphasizes the need for generalizable, flexible and adaptable methodologies that can support DDABM development across diverse domains, presenting both a significant challenge and an opportunity for future research in the field of DDABM.

4.2 Data-Driven Approaches for Derivation of Agent-based Models

Our review highlights a current predominance of SDDS approaches over DDDS in existing studies, which is not unexpected given that DDDS remains a relatively emerging research area. Traditionally, most data-driven ABM studies using SDDS rely on static datasets (Li et al. 2016), often sourced from databases to construct ABMs. However, recent advancements in real-time data availability enable more accurate and promising analyses and predictions through simulations (Li et al. 2016).

Although often categorized as data-driven, these existing SDDS approaches primarily only map and use the data to extract certain ABM components (such as agents' state), calibrate parameters (Sajjad et al. 2016), and run simulations. SDDS approaches demonstrate the feasibility of data-informed ABM development but remain limited in their ability to calibrate and update the models dynamically. As a result, SDDS-produced models are less accurate than DDDS, as they rely on static data mapping without continuous adaptation and refinement.

In contrast, DDDS approaches integrate real-time data streams, enabling ongoing model refinement. This enables the identification of data changes and continuous adjustment of both simulation outcomes and model parameters (Niloofar et al. 2023). For instance, in a simulation study of traveling and charging behavior of electronic taxis (ETs), Zhang and Tan (2024) extracted ETs agents from data. These agents make decisions based on the real-time environment (which is the real-time state of charging stations nearby) and interact with other agents. This approach produces a more accurate simulation results, better aligning with actual driver habits and real-world behaviors, and represent an innovation over existing static data-driven ABMs' study.

While DDDS remains underutilized in data-driven ABMs, its emerging applications signify a critical shift in modeling ABMs. This approach represents greater potential for advancing modeling and simulation techniques, particularly in supporting the growing need for the adaptive and responsive DTs of multi-agent systems across multiple domains. The opportunity associated with incorporating DDDS for ABMs in the context of supporting DTs for MASs will be further explored in Section 4.4.

4.3 Data Requirements and Corresponding Components in Data-driven Agent-based Models

Data-driven ABMs require extensive, high-quality data to extract accurate models of agents' behaviors, environmental conditions and interaction patterns. Our review reveals distinct patterns in data requirements across different types of ABMs. These requirements can be categorized based on the nature of the system being modeled and the agents' behaviors. Although reviewed studies do not specifically mention distinct data types, we classify them based on their data sources and extraction methods. We identify patterns in how data is obtained and used. For example, we classify behavioral survey as micro-quantitative because it is quantitative data at the individual-level. This approach allows us to better categorize and analyze various data requirements and map them to the corresponding ABM components.

In human-related data-driven ABMs, both *micro-* and *macro*-level data are critical for extracting and modeling human agents. Micro-level data includes detailed individual data, such as *demographic data* or *government census* (e.g., gender, age, income), essential for simulating individual agents' behavior. In contrast, macro-level data describes broader system trends, such as overall *population statistics*, which are essential for understanding group behaviors and the emergence of collective dynamics within a population.

These data types can be both *empirical quantitative* and *empirical qualitative*. Empirical quantitative data are collected through surveys (Yang and van Dam 2022), whereas empirical qualitative data are typically gathered via interviews, as demonstrated by Bae et al. (2024) and Ravaioli et al. (2023). These data types are crucial for extracting *agents' characteristics, properties, and behaviors.*

In studies related to transportation (considering human behavior) (Bae et al. 2024), disaster management (Lee et al. 2021), sociology (Sajjad et al. 2016; Rosés et al. 2021), epidemiology (Hunter et al. 2018), urban planning (Yang and van Dam 2022), and business (customer behavior) (Bell and Mgbemena 2018), both micro-level and macro-level data of individuals and populations are integrated. Micro-level data enables agents to make autonomous decisions, while macro-level data captures the collective outcomes of these individual decisions, shaping the overall system dynamics.

To inform *environment* design (components, characteristics, structure) in ABMs, spatial data is commonly used (8 out of 9 studies). Since environment in ABMs often involve spatial dimensions, GIS data representing geographical locations is frequently applied to model environmental structures, such as the layout of a city, street network or land use zones. These data types are classified as quantitative-micro as they contain agent-level information and are empirically collected from GIS databases. For instance, as demonstrated by Rosés et al. (2021) in a study simulating crime patterns in an urban environment, environmental components are extracted from spatial data such as GIS layers, public transport stops, and school locations, likely sourced from government databases.

As noted earlier, spatial data and temporal data are interconnected, with some GIS data containing both time and location information. Even though this is not always explicitly stated, ABM components can be extracted from these data types through spatial and temporal analysis. Of course, not limited to extracting environmental components, spatial and temporal data are also used to extract components of agents, depending on the nature of the system. An example can be seen in the study by Zhang and Tan (2024), where GIS data is used to extract both agents' components and environmental components, as well as interaction rules. An example of temporal data can also be seen in the same study, where charging event logs of taxis are used to extract and model agent behavior over time.

When it comes to extracting *interactions* (agent-agent, agent-environment, topology), some literature does not disclose the extraction methods or data sources, while others mention machine learning, probabilistic (rule-based) methods, hybrid models, and decision trees. There is no clear pattern in terms of the data types used for extracting interaction rules. The data types mainly include quantitative, spatial, and temporal data, with quantitative and spatial being the most common. The use of these data types does not appear to be strongly tied to the specific extraction methods employed, as both probabilistic and machine learning methods are applied across various data types.

4.3.1 Mapping Data to ABM Components: Potential Practical Guidance for DDABM Development

Based on our synthesis, we propose preliminary practical guidance to assist DDABM modelers in choosing suitable data types for key ABM components, as shown in Table 3. This mapping helps identify what data is typically used to extract each component.

Furthermore, we recommend incorporating dynamic real-time data when possible, following the DDDS approach. It helps improve the model's accuracy and enables continuous refinement, which static data alone cannot support. When real-time data is unavailable, combining micro-level and macro-level static data can provide a more complete view of agent behavior and support more effective modellings, especially for capturing both individual decisions and aggregate trends. We also recommend extracting data from both quantitative and qualitative (especially knowledge from experts or experience from real users) sources, when possible, as these data types can complement each other and improve the accuracy of ABMs.

Table 3: Recommended data types for mapping ABM Components.

Component	Relevant Data Type
Agents	Quantitative data (e.g., surveys); Qualitative data (e.g., interviews) for decision rules and contextual behaviors; Temporal data (e.g., activity log) to extract agent behavior over time.
Environment	Spatial data (e.g. GPS data), especially important for modeling agent movement.
Interaction Rules	Qualitative data (e.g., expert input); Micro-level quantitative data (e.g., activity logs) Spatial data for agent-environment interactions.

4.4 Challenges and Opportunities for Data-driven ABMs

We identified a lack of structured approaches for data-driven extraction of ABMs in existing studies. Notably, many studies do not consistently document how data is integrated or how ABM components are derived from data. For example, none of the reviewed studies explicitly describe input data properties, granularity, or data types, nor do they detail methods for extracting all ABM components. This gap limits transparency and hinders both replicability and the extension of existing work. While the ODD (Overview, Design concepts, Details) Protocol by Grimm et al. (2020) provides a standard framework for describing ABMs, it currently does not capture detailed information on data integration or components derivation. To address this gap and better support DDABM, we propose extending the ODD Protocol to include the elements such as data sources and characteristics, as well as model (components) derivation approaches. This extension can improve the transparency and replicability of DDABM studies.

Furthermore, while SDDS approaches still dominate current data-driven ABM studies, the shift toward DDDS presents a valuable opportunity for improving model accuracy and adaptability. By incorporating real-time data, DDDS allows agent behaviors to evolve in response to changing conditions, making simulations more aligned with real-world complexity. This is especially important when we adopt ABMs as underlying core models of DTs for MASs, where continuous refinement and adjustments are needed to reflect the real-world complexities of agent interactions, environmental changes, and decision-making processes.

5 SUMMARY AND OUTLOOK

This study provides an overview of the emerging field of Data-Driven Agent-Based Modeling (DDABM), focusing on the data requirements for deriving Agent-Based Models (ABMs) from real-world data. DDABM offers a novel approach to model development by integrating dynamic and static data sources to represent agents, their interactions, and environmental components. Our findings reveal that while most studies primarily use static data, there is a shift toward dynamic data-driven approaches, which hold potential for more adaptive and accurate models. However, challenges remain, the lack of standardized frameworks and structured approaches for data-driven ABMs hinders reproducibility and transparency.

This study emphasizes the need for clearer mappings between data types and ABM components, as well as the utilization of dynamic data sources for dynamic model refinement.

The future of DDABM lies in developing flexible methodologies applicable across various domains, moving beyond problem-specific approaches. Shifting from static Data-Driven Simulations (SDDS) to Dynamic Data-Driven Simulations (DDDS) offers significant promise, especially when using data-driven ABMs as the core models of Digital Twins (DTs) for Multi-Agent Systems (MASs), where real-time data updates are crucial for accurate, responsive simulations and continuous model refinements.

REFERENCES

Aloui, A., H. Hachicha, and E. Zagrouba. 2024. “Multi-Agent Based Framework for Cooperative Traffic Management in C-ITS System:” In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence Vol.1*, edited by Rocha, A. P., Steels, L., and van den Herik, J., 420–427. Setúbal: SCITEPRESS – Science and Technology Publications.

Ambra, T. and C. Macharis. 2020. “Agent-Based Digital Twins (ABM-Dt) In Synchromodal Transport and Logistics: The Fusion of Virtual and Pysical Spaces.” In *2020 Winter Simulation Conference (WSC)*, 159–69 https://doi.org/10.1109/WSC48552.2020.9383955.

Bae, J. W., C.-H. Lee, J.-W. Lee, and S. H. Choi. 2024. “A Data-Driven Agent-Based Simulation of the Public Bicycle-Sharing System in Sejong City.” *Simulation Modelling Practice and Theory* 130(1): 1–15.

Bell, D. and C. Mgbemena. 2018. “Data-Driven Agent-Based Exploration of Customer Behavior.” *SIMULATION* 94(3): 195–212.

Brown, D. G., R. Riolo, D. T. Robinson, M. North, and W. Rand. 2005. “Spatial Process and Data Models: Toward Integration of Agent-Based Models and GIS.” *Journal of Geographical Systems* 7(1): 25–47.

Bruch, E. and J. Atwell. 2015. “Agent-Based Models in Empirical Social Research.” *Sociological Methods & Research* 44(2): 186–221.

Chaib-draa, B. and J. P. Müller, eds. 2006. *Multiagent Based Supply Chain Management* (Studies in Computational Intelligence, Vol. 28). Berlin, Heidelberg: Springer.

Flamino, J., W. Dai, and B. K. Szymanski. 2019. “Modeling Human Temporal Dynamics in Agent-Based Simulations.” In *Proceedings of the 2019 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, 99–102. Chicago IL USA: ACM.

Ghorbani, A., G. Dijkema, and N. Schrauwen. 2015. “Structuring Qualitative Data for Agent-Based Modelling.” *Journal of Artificial Societies and Social Simulation* 18(1): 1–6.

Grimm, V., S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, et al. 2020. “The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism.” *Journal of Artificial Societies and Social Simulation* 23(2): 1–20.

He, Y. and W. Shen. 2025. “Agent-Based Digital Twins for Collaborative Machine Intelligence Solutions.” *IET Collaborative Intelligent Manufacturing* 7(1): 1–4.

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): 1–35.

Jamali, R. and S. Lazarova-Molnar. 2024. “A Comprehensive Framework for Data-Driven Agent-Based Modeling.” In *2024 Winter Simulation Conference (WSC)*, 620-631 https://doi.org/10.1109/WSC63780.2024.10838766

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*, edited by Frydenlund, E., Shafer, S., and Kavak, H.,1–12. San Diego: Society for Computer Simulation International.

Lazarova-Molnar, S. 2024. “A Vision for Advancing Digital Twins Intelligence: Key Insights and Lessons from Decades of Research and Experience with Simulation:” In *Proceedings of the 14th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, 5–10. Setúbal: SCITEPRESS – Science and Technology Publications.

Lee, H. M., R. Jamali, and S. Lazarova-Molnar. 2025. “A Conceptual Framework for Digital Twins of Multi-Agent Systems.” *Procedia Computer Science* 257: 321–328.

Lee, S., S. Jain, K. Ginsbach, and Y.-J. Son. 2021. “Dynamic-Data-Driven Agent-Based Modeling for the Prediction of Evacuation Behavior during Hurricanes.” *Simulation Modelling Practice and Theory* 106: 1–26.

Li, Z., X. Guan, R. Li, and H. Wu. 2016. “4D-SAS: A Distributed Dynamic-Data Driven Simulation and Analysis System for Massive Spatial Agent-Based Modeling.” *ISPRS International Journal of Geo-Information* 5(4): 42.

Lounis, M. and D. K. Bagal. 2020. “Estimation of SIR Model’s Parameters of COVID-19 in Algeria.” *Bulletin of the National Research Centre* 44(1): 1–6.

Lü, J., G. Chen, and X. Yu. 2011. “Modelling, Analysis and Control of Multi-Agent Systems: A Brief Overview.” In *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, 2103–2106.

Macal, C. M. and M. J. North. 2014. “Tutorial on Agent-Based Modeling and Simulation.” In *Agent-Based Modeling and Simulation*, edited by Simon J. E. Taylor, 11–31. London: Palgrave Macmillan UK.

Marah, H. and M. Challenger. 2024. “MADTwin: A Framework for Multi-Agent Digital Twin Development: Smart Warehouse Case Study.” *Annals of Mathematics and Artificial Intelligence* 92(4): 975–1005.

Mariani, S., M. Picone, and A. Ricci. 2022. “About Digital Twins, Agents, and Multiagent Systems: A Cross-Fertilisation Journey.” In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers*, edited by Francisco S. Melo and Fei Fang, 114–129. Cham: Springer International Publishing.

Monti, C., M. Pangallo, G. De Francisci Morales, and F. Bonchi. 2023. “On Learning Agent-Based Models from Data.” *Scientific Reports* 13(1): 1–12.

Moyaux, T., Y. Liu, G. Bouleux, and V. Cheutet. 2023. “An Agent-Based Architecture of the Digital Twin for an Emergency Department.” *Sustainability* 15: 1–13.

Niloofar, P., S. Lazarova-Molnar, F. Omitaomu, H. Xu, and X. Li. 2023. “A General Framework for Human-in-the-Loop Cognitive Digital Twins.” In *2023 Winter Simulation Conference (WSC)*, 3202–3013 https://doi.org/10.1109/WSC60868.2023.10407598

Paudel, R. and A. Ligmann-Zielinska. 2023. “A Largely Unsupervised Domain-Independent Qualitative Data Extraction Approach for Empirical Agent-Based Model Development.” *Algorithms* 16(7): 1–12.

Rahman, N. 2014. “Temporal Data Update Methodologies for Data Warehousing.” *Journal of the Southern Association for Information Systems* 2(1): 25–41.

Ravaioli, G., T. Domingos, and R. F. M. Teixeira. 2023. “A Framework for Data-Driven Agent-Based Modelling of Agricultural Land Use.” *Land* 12(4): 1–17.

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: 1–15.

Sajjad, M., K. Singh, E. Paik, and C.-W. Ahn. 2016. “A Data-Driven Approach for Agent-Based Modeling: Simulating the Dynamics of Family Formation.” *Journal of Artificial Societies and Social Simulation* 19(1): 1–14.

Santos, C. H. dos, J. A. B. Montevechi, J. A. de Queiroz, R. de Carvalho Miranda, and F. Leal. 2022. “Decision Support in Productive Processes through DES and ABS in the Digital Twin Era: A Systematic Literature Review.” *International Journal of Production Research* 60(8): 2662–2681.

Squazzoni, F., and R. Boero. 2005. “Does Empirical Embeddedness Matter? Methodological Issues on Agent-Based Models for Analytical Social Science.” *Journal of Artificial Societies and Social Simulation* 8(4)): 1–6.

Stary, C. 2021. “Digital Twin Generation: Re-Conceptualizing Agent Systems for Behavior-Centered Cyber-Physical System Development.” *Sensors* 21(4): 1–24.

Tian, G. and Z. Qiao. 2014. “Modeling Urban Expansion Policy Scenarios Using an Agent-Based Approach for Guangzhou Metropolitan Region of China.” *Ecology and Society* 19(3): 1–14.

Venkatramanan, S., B. Lewis, J. Chen, D. Higdon, A. Vullikanti, and M. Marathe. 2018. “Using Data-Driven Agent-Based Models for Forecasting Emerging Infectious Diseases.” *Epidemics* 22: 43–49.

Walsh, S. J., G. P. Malanson, B. Entwisle, R. R. Rindfuss, P. J. Mucha, B. W. Heumann, P. M. McDaniel, et al. 2013. “Design of an Agent-Based Model to Examine Population-Environment Interactions in Nang Rong District, Thailand.” *Applied Geography* 39: 183-198.

Yang, Liu and K. H. van Dam. 2022. “Data-Driven Agent-Based Model Development to Support Human-Centric Transit-Oriented Design.” In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers*, edited by Francisco S. Melo and Fei Fang, 60–66. Cham: Springer International Publishing.

Yang, Lu and N. Gilbert. 2008. “Getting Away from Numbers: Using Qualitative Observation for Agent-Based Modeling.” *Advances in Complex Systems* 11(02): 175–185.

Zhang, Y. and J. Tan. 2024. “A Data-Driven Approach of Layout Evaluation for Electric Vehicle Charging Infrastructure Using Agent-Based Simulation and GIS.” *SIMULATION* 100(3): 299–319.

AUTHOR BIOGRAPHIES

HUI MIN LEE is a PhD student at the Institute of Applied Informatics and Formal Description Methods at Karlsruhe Institute of Technology. Her research interests include data modeling and simulation, digital twins, and especially data-driven agent-based modeling and simulation. Her email address is hui.lee@kit.edu.

SANJA LAZAROVA-MOLNAR is a Professor at both 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, with an emphasis on reliability and energy efficiency. She develops advanced methodologies to optimize complex systems and leads several European and national projects in these areas. Prof. Lazarova-Molnar holds leadership roles in IEEE and The Society for Modeling & Simulation International (SCS), where she currently serves as SCS Representative to the Winter Simulation Conference (WSC) Board of Directors. She was Proceedings Editor for WSC in 2019 and 2020 and serves as Associate Editor for *SIMULATION: Transactions of The Society for Modeling and Simulation International*. Her email address is lazarova-molnar@kit.edu.