Framework of Experimental Design and Data Mining in Multi-agent Simulation

Fa Zhang, Shi-Hui Wu, and Zhi-Hua Song

***Abstract*—Multi-agent based simulation (MABS) is an important approach for studying complex systems. The Agent-based model often contains many parameters, these parameters are usually not independent, with differences in their range, and may be subjected to constraints. How to use MABS investigating complex systems effectively is still a challenge. The common tasks of MABS include: summarizing the macroscopic patterns of the system, identifying key factors, establishing a meta-model, and optimization. We proposed a framework of experimental design and data mining for MABS. In the framework, method of experimental design is used to generate experiment points in the parameter space, then generate simulation data, and finally using data mining techniques to analyze data. With this framework, we could explore and analyze complex system iteratively. Using central composite discrepancy (CCD) as measure of uniformity, we designed an algorithm of experimental design in which parameters could meet any constraints. We discussed the relationship between tasks of complex system simulation and data mining, such as using cluster analysis to classify the macro patterns of the system, and using CART, PCA, ICA and other dimensionality reduction methods to identify key factors, using linear regression, stepwise regression, SVM, neural network, etc. to build the meta-model of the system. This framework integrates MABS with experimental design and data mining to provide a reference for complex system exploration and analysis.**

***Index Terms*—Agent, complex systems, data mining, simulation, uniform design.**

I. Introduction

There are many complex systems in the real world [1]. Complex system is composed of many elements or components and there are nonlinear interactions among them. The evolution of a complex system is difficult to predict. Typical complex systems include ant colony, human brain, multinational corporation, financial market, and the Internet. The research of complex systems has been a hot topic in the scientific community in recent decades [2], though there has not a consensus on what a complex system is and how to measure complexity [3].

Manuscript received October 25 2021; revised December 23, 2021. This work was supported by the National Natural Science Foundation of China under Grant No. 71571190.

Fa Zhang is with the Beijing Institute of Technology, Zhuhai School, Zhuhai, China (e-mail: richter2000@163.com).

Shi-hui Wu and Zhi-hua Song are with the Equipment Management and UAV Engineering College at the Air Force Engineering University, Xi'an, China (e-mail: wu_s_h82@sina.com, szhele@163.com).

The basic requirement for a complex system generally accepted by the scientific community is only "beyond reductionism", which is said in a methodological sense. Complex systems have some properties such as emergence, stability, adaptability, and criticality [4]. The understanding of complex system cannot be obtained only by understanding its parts. Therefore, it is necessary to invent method which is "beyond reductionism". Multi-Agent based simulation (MABS, or ABM, MABS, IBM, etc.) is just a method satisfied with "beyond reductionism". MABS uses a microscopic perspective to build model, decompose the system into a number of agents (individuals), describe the properties and behaviors of the agents, and the agents interact with each other. Using simulation, the macro state of the system emerges from the bottom up. MABS communicates the microscopic and the macroscopic level of the system, can be used to explore the mechanism of complex systems, and could help to deepen the understanding of the system [5]. Since the 1980s, MABS has developed rapidly and has been applied in many fields such as physics, chemistry, biology, and social science, and has become an important method in complex system research [6].

In general, agent-based model contains many heterogeneous agents with autonomous capabilities. Each type of agent has its properties and behavior rules, and agents interact in a specific way. In initial stage of simulation, a number of agents are distributed in environment (physical space or logical space), and the environment also has its parameters. When simulation is running, the agents runs concurrently and interacts with each other according to the interaction rules to form the dynamic of the system. There are usually large number of variables in MABS, which are set up at the initial stage and keep constant during a simulation run, these variables are called parameters (or factors in experimental design). The parameters of MABS may have characteristics as follow:

A. There Are Many Parameters in MABS

MABS generally contains many parameters, the common types include:

- Individual parameters, parameters describing the attributes or behavior of the agent, such as the age threshold of the agent, the behavior excitation threshold in decision rules, etc.
- Population parameters, such as the initial number of agents, the timing of agent injection, etc.
- Environmental parameters, such as the range of physical space, type of network topology, grid division scale, etc.
- Simulation parameters, such as parallel/serial of the

DOI: 10.7763/IJMO.2022.V12.791

simulation engine, clock advance mode, time step, etc.

B. Parameters are Often Difficult to be Calibrated

The parameters of multi-agent models are often derived from assumptions. Many of the parameters are conceptual and cannot find a direct relationship with real physical/social processes. These parameters cannot be observed and can only be given a rough range.

C. Parameters are Often Heterogeneous

The parameters often come from different sources, with obvious differences in their data types, value ranges, and granularity. For example, the variable may be continuous or discrete, and the data type may be 0-1, integer, real, or characters. Their range and resolution are often quite different.

D. Sometimes the Parameters ARE not Independent and Need to Meet Some Constraints

The parameters may have internal connections and do not satisfy the independence assumption. Some parameters need to meet complex constraints.

Therefore, the parameter space of complex systems is a high-dimensional, mixed, and constrained space. Exploring the parameter space is an important task of complex system research, and it is the basis for discovering the system's macroscopic pattern, screening key factors, and finding micro-macro relationship. How to effectively explore the parameter space of MABS is an important problem faced by complex system simulation.

Since the 1970s, a few scholars have combined experimental design with simulation and developed simulation experimental design theory [7], which provides a theoretical basis for reducing the number of simulation experiments and analyzing simulation results. In recent years, big data and machine learning have developed rapidly, and data analysis capabilities have been greatly enhanced. Some scholars have realized the value of data mining to simulation. Remondino and Correndo tried to apply data mining to agent-based simulation [8]. Saoud and Boubetra analyzed the data processing problems in simulation and designed a data collection agent [9] . Patel, Abbasi, etc. combined exploratory analysis and data mining for the analysis of agent-based simulation results [10]. Sitova and Peceskar proposed a data farming and knowledge discovery framework for simulation results [11]. Shao, Ye, etc. developed a machine learning based simulation data mining approach to realize global performance evaluation [12]. These studies are very valuable and provide basic ideas for the combination of data mining and simulation. However, in the process of simulation research, experimental design and data analysis cannot be separated and need to be considered as a whole. From this perspective, we proposed an iterative simulation research framework that integrates experimental design and data mining to improve the efficiency of MABS.

II. Common Tasks of MABS

Using simulation to study the system has different purposes and forms different types of simulation tasks. Common tasks of simulation include V&V, what-if analysis, optimization, and risk analysis, etc. As a subtype of simulation, MABS has roughly the same research goals and tasks, but has some characteristics of its own. In reality, MABS has two common way of uses. One is to explore the mechanism of complex systems and deepen the understanding of complex systems. The other is to build model of a real system, evaluate and predict the system, or find the optimal solution that meet specific measures. Common tasks of MABS are listed as follows:

A. Summarize the Macro Pattern of the Complex Systems

The evolution of complex systems is diverse, dynamic, and uncertain. MABS can help to explore the parameter space of the system comprehensively, discover the macro pattern of the system, and master the characteristics of the system. According to whether it involves time, the macro pattern of the system is divided into static and dynamic.

Static pattern: Use measures $\boldsymbol{y}$ (scalar or vector) to describe it, which can be continuous or discrete. For example, in the simulation of infectious diseases, macro patterns such as extinction, epidemic, and pandemic can be discovered according to the infection rate.

Dynamic patten: summarized according to the dynamic characteristics of time series data $\boldsymbol{y}(t)$. Such as exponential growth (decay), periodic oscillation, chaos and others.

B. Identify the Key Factors Affecting the System

In MABS, the macroscopic characteristics of the system depend on the microscopic characteristics and interaction rules of the system. There are many factors at the micro level, and it is necessary to screen and identify the key factors that affect the macro mode of the system, and judge whether there is an interactive effect between the factors.

C. Build the Meta-Model of the Simulation System

The simulation model is essentially a mathematical transformation $\boldsymbol{y} = f(\boldsymbol{x}) + \varepsilon$, but the function f is implicit. Given $\boldsymbol{x}$, the measure $\boldsymbol{y}$ can only be obtained through simulation run. In order to simplify the description of the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$, a meta-model $\boldsymbol{y} = f'(\boldsymbol{x})$ can be established, which is an approximation of the simulation model. The meta-model can compute $\boldsymbol{y}$ more quickly, or it can be embedded into a more complex simulation system as a build block.

D. Simulation Optimization

In MABS with application background, it is often necessary to find a set of parameters that makes the objective function optimal (maximum or minimum). For example, in the simulation of prevention and control policy of infectious disease, in order to minimize the number of infected people, it is necessary to find the optimal control parameters.

III. Experimental Design and Data Analysis

There are many types of simulation. According to the two dimensions of random/determined and static/dynamic, Kleijnei divides simulation models into four types: Deterministic and static, Random and static, Deterministic and dynamic, Random and dynamic [7]. When using MABS for complex system simulation, it belongs to the “Random and dynamic”. Using the black box point of view, the MABS simulation model can be viewed as:

$$y = F(x, c, r_0)$$

where $y = (y_1, y_2, \dots, y_m)$ is the output variable, $x = (x_1, x_2, \dots, x_n)$ is the controllable input variable (parameter, factor), c is the uncontrollable parameter (environmental variable), the random number seed r_0 used to generate random number stream.

The basic process of using MABS for complex system research is: according to the purpose of study, design a simulation experiment, run the simulation model to collect output data, then analyze the simulation output data, and finally make the conclusions about the complex system. The process of simulation experimental design and data analysis are shown in Fig. 1. In most cases, this process may be repeated many times. In addition, simulation research often involves V&V, which can be regarded as a simulation task, following the same process.

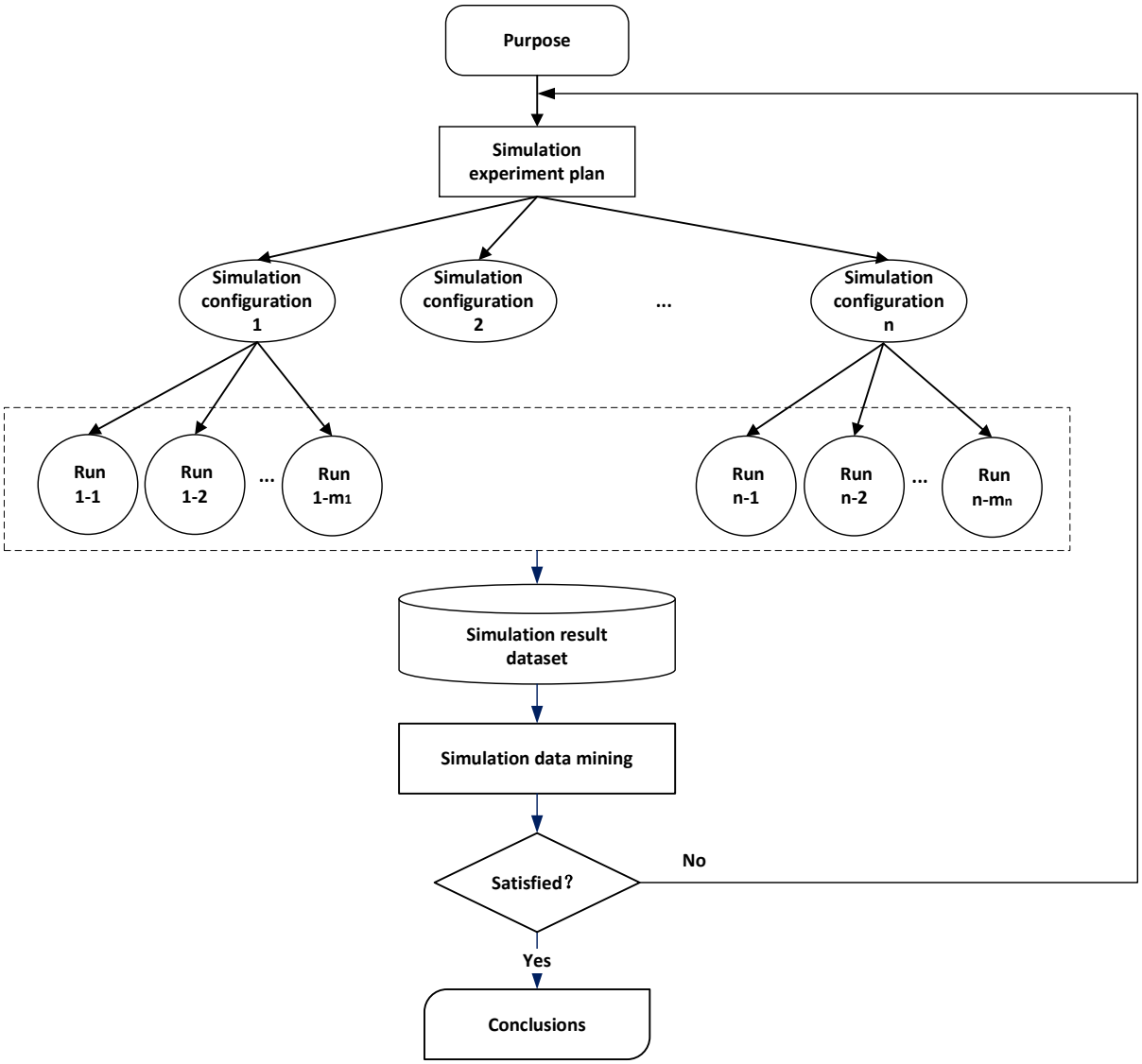

Fig. 1. The process of using MABS to investigate complex system.

As shown in Fig.1, we determine the specific purpose firstly according to the simulation task. No matter what kind of research purpose, it will eventually be transformed into an exploration of the relationship between simulation input and output. Based on the input-output transformation of, we then select a set of variables from the controllable input x to form the experimental factors, denoted as $x_S = \{x_1, x_2, \dots, x_s\}$, and select a set of variables from the simulation output y to form the performance index (response variable), denoted as $y_R = \{y_1, y_2, \dots, y_r\}$. Also need to define the range of each factor, for continuous factor $x_i \in [a_i, b_i]$. For discrete factors, without loss of generality, they are denoted as $x_i \in \{1,2,3, \dots, q_i\}$. According to the condition in reality, sometimes the factor x is required to satisfy a set of constraints, which is denoted as:

$$f_j(x) \le 0, j = 1,2, \dots, t$$

For continuous factor x, the domain of the simulation experiment is denoted as:

$$S = \{x | x \in [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_s, b_s]$$
$$and\ f_j(x) \le 0, j = 1,2, \dots, t\}$$

An important step in simulation research is to make a simulation experiment plan. The simulation experiment plan $D(n, S)$ is a point set composed of n points which is selected from the experimental domain S. Obviously, the number of

possible experimental plan is very large, so experimental design theory is often adopted, and the research purpose is achieved with as few experiments as possible. There are many experimental design methods, such as Monte Carlo, Orthogonal Design, Uniform Design, Latin Hypercube Sampling, partial Factorial Design, etc.

In experimental design theory, the level of each factor is often determined in advance, and the mapping relationship $l_i = z_i(x_i)$ is established between the factor x_i and its factor level $l_i = \{1,2,\dots,q_i\}$, so the experimental domain is expressed as:

$$S = \{l | l \in \{1,2,\dots,q_1\} \times \{1,2,\dots,q_2\} \times \cdots \times \{1,2,\dots,q_s\} \quad and\ f_j(z^{-1}(l)) \le 0, j = 1,2,\dots,t\}$$

Especially, when each factor takes q levels, denote $C = \{1,2,\dots,q\}^s$. Since the mapping $l_i = z_i(x_i)$ is relatively simple, x_i is often used to represent the factor level l_i in the following text, and no strict distinction is made anymore.

The simulation experiment plan $D(n,S)$ contains n experimental points, and each experimental point determines a combination of all factors, which is called a simulation configuration. Because of the randomness in the random dynamic model, each simulation configuration needs to be repeated many times, and a sample of the response variable is collected for each run, and the response variable under the configuration is obtained after statistical processing.

After the execution of the simulation experiment plan is finished, the simulation result dataset $DS = \{R_i, i = 1,2,\dots,n\}$ can be obtained, and the record of the dataset saves both the level of factora and the value of the response variable. Each data record is $R_i = (x_1, x_2, \dots, x_s, \widetilde{y_1}, \widetilde{y_2}, \dots, \widetilde{y_r})$, where $\widetilde{y_j}$ is the statistical result of y_j in m_i replications.

Traditionally, statistical methods are used to analyze simulation output data. With the rapid development of big data techniques, data mining (including statistical analysis, machine learning, deep learning, etc.) can now be used for data analysis to gain an understanding of the complex system.

In the simulation framework, experimental design and data analysis are the most important components, and there is a close relationship between them. On the one hand, different experimental design methods have different assumptions and the generated data have different characteristics, which restrict the subsequent data analysis methods. On the other hand, the data analysis method to be used has specific requirements for data, and appropriate experimental design methods must be used to generate data. Therefore, it is necessary to integrate experimental design and data mining.

IV. Uniform Design Under Any Constraints

Experimental design help to achieve the research purpose with fewer experiments. With the widespread of simulation, some experimental methods suitable for simulation have emerged. Kleijnen pointed out that there are many factors in the simulation experiment, there are uncertainties, and the calculation load is large, but the simulation experiment is a pseudo experiment conducted on the computer, which has some advantages compared with the physical experiment. Including: simulation experimental factors are easy to change, we can explore a larger range of parameters, use pseudo-random numbers, do not need to consider randomization and blocking, and pay more attention to sequential design [7]. Therefore, the simulation experimental design should not completely copy the classic experimental design, but choose and adjust according to the characteristics of the simulation experiment.

In MABS, the number of experimental factors is large, the range of factor is wide, factors are not independent, and sometimes there are constraints among factors. Among many experimental design methods, the uniform design (UD) proposed by KaiTai Fang et al. only considers the uniform dispersion of experimental points in the experimental domain, and the number of experiments required is proportional to the number of factors [13]. When the number of factors is large and the level of factors is more, the uniform design can meet the requirements with more fewer experiments. Therefore, UD is suitable for use in MABS, but has the following shortcomings:

A. In MABS, not Only Consider the Uniform Dispersion of Experimental Points

Uniform design selects a small number of experimental points in the experimental domain, and these experimental points are evenly distributed in the experimental domain. If the system model is relatively simple, such as linear or quadratic function, this design is reasonable. However, complex system are often discontinuous, non-smooth, and the landscape may be rugged, and there are multiple peaks/valleys and sudden changes in some areas. If only a small number of uniformly dispersed experimental points are selected, the true characteristics of the model may not be presented, which may lead to incorrect inferences.

B. There are Complex Constraints among Experimental Factors

There may be any constraint among the factors of a complex system, and there may be various forms of constraints, such as nonlinear functions. The experimental domain of uniform design is often a n-dimensional hypercube. In the mixture design of UD, $\sum_i x_i = 1, x_i \ge 0$, and linear or idempotent constraints are allowed. In MABS, this assumption is not satisfied. If a number of experimental points are first generated with a uniform design, and then the experimental points that do not meet the constraints are deleted, it may lead to too few experimental points and insufficient data for analysis.

Unlike physical experiments, simulation experiments can accept more experimental points. With the rapid improvement of computing power, more attention is paid to the comprehensive and fine filling of the experimental domain, and the primary goal is to reveal the true characteristics of the model. In scenarios there are arbitrary constraints among factors, we proposed a method that combine uniform experimental design and random selection, which could make better space filling and generate controllable experimental points, and facilitate the use of data mining for data analysis.

Suppose the simulation model has s factors, $\boldsymbol{x} = (x_1, x_2, \dots, x_s)$, and each factor can be mapped to the interval $[0,1]$ through a group of linear transformations, then suppose $0 \le x_i \le 1, i = 1,2,\dots,s$. $C = [0,1]^S$ is an s-dimensional

unit hypercube. It is required that x satisfy some constraints $g_j(x) \leq 0, j = 1,2,\dots,t$. Then the experimental domain S is the space which is satisfy the constraint g_j in the unit hypercube $C = [0,1]^S$, as follow:

$$S = \{\boldsymbol{x} | \boldsymbol{x} \in [0,1]^s and\ g_j(\boldsymbol{x}) \leq 0, j = 1,2,\dots,t\}$$

In the field of uniform design, some scholars have studied the mixing problem with constraints. In the mixing problem $\sum_i x_i = 1$, the experimental domain is a simplex. For the non-rectangular experimental domain, Chuang and Hung used the switching algorithm to construct an approximately uniform design [14]. For complex constraints, Liu and Liu proposed a uniform design algorithm that satisfy the complex constrained mixture problem [15]. Ning designed a nearly uniform design construction [16] for the flexible region, $\{(x_1, x_2, \dots, x_s): |x_1|^m + |x_2|^m + \cdots + |x_s|^m \leq 1\}$. What we are facing is the experimental design with any constraints. There are no restrictions on the shape of the experimental domain. A new algorithm needs to be designed. Furthermore, we hope that experimental data meets data mining needs.

There are many different measures of uniformity. We uses the central composite discrepancy (CCD) proposed by Chuang and Hung to measure the uniformity of experimental points. For the design $\mathcal{P}$ of n points on the experimental domain S, the simplest case is to divide each dimension into two, so that S is divided into 2^s small areas, and the CCD is approximately expressed as [14]:

$$CCD_2(n, \mathcal{P}) \approx \left\{ \frac{1}{N} \sum_{i=1}^{n} \frac{1}{2^s} \sum_{t=1}^{2^s} \left| \frac{N(S_t(x_i), \mathcal{P})}{n} - \frac{N(S_t(x_i))}{N} \right|^2 \right\}^{1/2}$$

The algorithm for generating $n+m$ experimental points in the experimental domain S with any constraints is as follows. Here n and m are parameters given in advance. These n uniformly distributed experimental points ensure the spatial coverage of the experimental domain, and these m randomly distributed experimental points help to show the complex characteristics of the response surface.

The Algorithm

Step 1:

Let $N \gg n$

Find a uniform design $\mathcal{P}(N, C) = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ on $C = [0,1]^S$

Step 2：

Let $\mathcal{P}(N', S) := \emptyset$

for i=1 to N

if $\boldsymbol{x}_i \in \mathcal{P}(N, C)$ satisfied $g_j(\boldsymbol{x}_i) \leq 0, j = 1,2,\dots,t$:

$\mathcal{P}(N', S) \leftarrow \mathcal{P}(N', S) \cup \{\boldsymbol{x}_i\}$

end if

end for

if $not\ |\mathcal{P}(N', S)| \gg n$:

return Step1and increase N

end if

Step 3:

Select n points from $\mathcal{P}(N', S)$ as initial design C_{design} , e.g. $C_{design} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$

Step 4:

$i := 0$, $N_{design} := C_{design}$

while i=0 or $N_{design} \neq C_{design}$

$i := i + 1$, $C_{design} := N_{design}$

for j =1 to n

$\boldsymbol{x}^* = argmin_{\boldsymbol{x} \in \mathcal{P}(N', S) - N_{design}} CCD(n, N_{design} - \{\boldsymbol{x}_j\} \cup \{\boldsymbol{x}\})$

if $CCD(n, N_{design} - \{\boldsymbol{x}_j\} \cup \{\boldsymbol{x}^*\}) < CCD(n, N_{design})$:

$N_{design} := N_{design} - \{\boldsymbol{x}_j\} \cup \{\boldsymbol{x}^*\}$

end if

end for

end while

Step 5：

Let $\mathcal{P}(m, S) := \emptyset$, i=0

while i<m

Random select $\boldsymbol{x} \in C$

if $\boldsymbol{x}$ satisfied with $g_j(\boldsymbol{x}) \leq 0, j = 1,2,\dots,t$:

$\mathcal{P}(m, S) \leftarrow \mathcal{P}(m, S) \cup \{\boldsymbol{x}\}$

$i := i + 1$

end if

end while

Step 6：

Output $C_{design} \cup \mathcal{P}(m, S)$

V. Data Mining for Simulation

Traditionally, statistic method is used for simulation output data analysis, which usually requires the data to meet the independence assumption and some prior knowledge about the system is required. With the development of data science, the fourth paradigm of scientific research suitable for data-intensive problems has emerged. Using the fourth paradigm, we can regard the simulation model as a data generation machine, drive the model to produce input-output datasets, and then use machine learning to conduct exploratory analysis and mining of the datasets to obtain a comprehensive and in-depth understanding of the system.

The agent-based model is a simulation of a complex system at the micro level. It can also be viewed as a data generation mechanism. There is a time advance mechanism inside simulation model. Agents interact in parallel in the time according to the interaction rules to generate the system state. As the simulation clock advances, a trajectory of the system state is formed. Given a simulation configuration x , the model can be regarded as a state transition function in discrete time:

$$S_{t+\Delta t} = Sim_x(S_t, \Delta t)$$

The state trajectory is collected and transformed after data collection, and the response variable $y = \{y_1, y_2, \dots, y_r\}$ is obtained. Each simulation configuration and response variable form a pair of input-output data (x, y).

In the complex system simulation, there isn't sufficient prior knowledge of the system, so no model assumptions are suitable. According to the algorithm in the previous section, the uniform experimental design method is used to produce a few experimental points, and a set of data is obtained to meet the space filling requirements. Then, a number of simulation configurations are randomly generated using the Monte Carlo method to meet the comprehensive requirements. Combine the two parts of the data to get the dataset $DS = \{(\boldsymbol{x}, \boldsymbol{y})^{(1)}, (\boldsymbol{x}, \boldsymbol{y})^{(2)}, \dots, (\boldsymbol{x}, \boldsymbol{y})^{(n)}\}$. This kind of dataset has sufficient coverage of the experimental space, and the amount

of data is not very much, so it is suitable for obtaining an understanding of the system through data mining. Different simulation tasks may use different data mining methods. The following discusses the choice of data mining methods for common MABS tasks.

A. Discover the Macro Pattern of a Complex System

Based on the purpose, select some attributes from the dataset DS as feature vectors, and perform cluster analysis on the feature vectors. For static macro pattern, cluster analysis can be performed directly on the dataset. If the feature vector is continuous, distance-based methods such as k-mean and k-medoids are often used. For discrete feature vectors, hierarchical clustering methods are often used, such as AGNES, DIANA, Chameleon, etc. Through cluster analysis, the macroscopic patterns of complex systems are obtained. For the dynamic macro pattern of the system, time series analysis or dynamic data mining methods such as data flow mining and empirical mode decomposition are used.

B. Identify the Key Factors Affecting the System

Screening key factors from a large number of factors is an important step of understanding complex systems. From the perspective of data mining, identifying the key factor is dimensionality reduction. There are many dimensionality reduction methods, such as wavelet transformation, Principle Component Analysis (PCA), Independent Component Analysis (ICA), Self-Organizing map (SOM), etc. There are also a large number of feature selection methods in data mining, including forward selection, backward elimination, and optimize selection. Through dimensionality reduction or feature selection, several factors that have great impact on the system, i.e. key factors, are discovered.

C. Build the Meta-Model of the System

The simulation model is an implicit description of the relationship between the input parameter $\boldsymbol{x}$ and the output $\boldsymbol{y}$. Through mining the dataset, the meta-model of the simulator can be established. There are many types of meta-models, such as linear regression, stepwise regression, Kriging, CART, SVM, neural network, etc. Different meta-models have differences in simplicity, effectiveness, computational efficiency, and interpretability. Choose the appropriate meta-model according to the requirements and use it as an approximation of the simulation model.

D. Parameter Optimization

For the optimization task, given the objective function $z = f(\boldsymbol{y})$, find $\boldsymbol{x}^*$ that can produce a minimal (maximal) z. Because simulation can only produce $\boldsymbol{y}$ from $\boldsymbol{x}$, it cannot get $\boldsymbol{x}$ from $\boldsymbol{y}$. It is necessary to combine simulation with a certain optimization mechanism such as evolutionary algorithm, grid search, etc. to realize the optimization of parameters. But simulation optimization requires a lot of computing. It is possible to use the parameter optimization method in data mining to find the optimal parameters based on the simulation data set. Some data mining tools provide grid search and heuristic search such as genetic algorithm, which can quickly obtain approximate optimal solutions.

VI. Conclusions

Research on complex systems is becoming more and more popular. As an important method of complex system research, MABS has been successfully applied in many fields. How to improve the research efficiency of MABS, there are still some problems. This paper analyzed the characteristics of MABS and point out that MABS has many factors, large difference in the range of factors, heterogeneity of factors, and subject to complex constraints among factors. In simulation research, it is necessary to carry out experimental design based on experimental design theory to minimize experimental points. At the same time, it is necessary to relaxes the requirements for data and uses data mining to analyze simulation data. This paper proposes a MABS framework that combines experimental design and data mining. It could help to improve the efficiency of complex system research and gain understanding of complex systems. There are many experimental methods and data mining methods. Experimental design and data mining are closely related. There are still many problems to study, such as improving experimental design methods based on data mining requirements, or revealing limitation of data mining for simulation data processing, etc.

Conflict of Interest

The authors declared that they have no conflicts of interest to this work. We do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Author Contributions

F. Zhang formulated the question and proposed a UD based algorithm. S. H. Wu studied the simulation optimization problem. Z. H. Song studied the application of data mining in simulation. All authors had approved the final version.

Acknowledgment

This research was supported by National Natural Science Foundation of China (No. 71571190). The authors appreciate the financial supports deeply.

References

[1] N. Goldenfeld and L. P. Kadanoff, "Simple lessons from complexity," *Science*, vol. 284, no. 5411, pp. 87-89, 1999.
[2] R. Preiser, "Identifying general trends and patterns in complex systems research: An overview of theoretical and practical implications," *Systems Research and Behavioral Science*, vol. 36, no. 5, pp. 706-714, 2019.
[3] A. F. Siegenfeld and Y. Bar-Yam, "An introduction to complex systems science and its applications," *arXiv*, 2020.
[4] C. Szabo, "Complex systems modeling and analysis," in *Proc. the 2019 Winter Simulation Conference*, 2019, pp. 1495-1503.
[5] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," in *Proc. National Academy of Sciences of the United States of America*, vol. 99 no. suppl 3, pp. 7280-7287, 2002.
[6] S. Abar, G. K. Theodoropoulos, P. Lemarinier, and G. M. P. O'Hare, "Agent based modelling and simulation tools: A review of the state-of-art software," *Computer Science Review*, vol. 24, pp. 13-33, 2017.
[7] J. P. C. Kleijnen, *Design and Analysis of Simulation Experiments*, Springer, 2015.
[8] M. Remondino and G. Correndo, "Data mining applied to agent-based simulation," in *Proc. 19th European Conference on Modeling and Simulation*, 2005, pp. 387-392.

[9] M. S. Saoud, A. Boubetra, and S. Attia, "How data mining techniques can improve simulation studies," *International Journal of Computer Theory and Engineering*, vol. 6, no. 1, pp. 15-19, 2014.
[10] M. H. Patel, M. A. Abbasi, M. Saeed, and S. J. Alam, "A scheme to analyze agent-based social simulations using exploratory data mining techniques," *Complex Adaptive Systems Modeling*, vol. 6, no. 1, 2018.
[11] I. Šitova and J. Pečerska, "Approach to integration of data mining techniques in simulation results analysis," *Information* Technology *and Management Science*, vol. 21, pp. 86-92, 2018.
[12] Y. Shao, Y. Liu, X. Ye, and S. Zhang, "A machine learning based global simulation data mining approach for efficient design changes," *Advances in Engineering Software*, vol. 124, pp. 22-41, 2018.
[13] K.-T. Fang, M.-Q. Liu, H. Qin, and Y.-D. Zhou, *Theory and Application of Uniform Experimental Designs*, Science Press, 2018.
[14] S. C. Chuang, and Y. C. Hung, "Uniform design over general input domains with applications to target region estimation in computer experiments," *Computational Statistics and Data Analysis*, vol. 54, no. 1, pp. .219-232, 2010.
[15] Y. Liu and M.-Q. Liu, "Construction of uniform designs for mixture experiments with complex constraints," *Communications in Statistics - Theory and Methods*, vol. 45, no. 8, pp. 2172-2180, 2015.
[16] J. H. Ning, W. W. Yin, and L. Peng, "Nearly uniform design construction on flexible region," *Acta Mathematicae Applicatae Sinica*, vol. 36, no. 3, pp. 557-565, 2020.

Fa Zhang is a professor at Beijing Institute of Technology, Zhuhai School. He was born in Hebei, P.R. China on 1970. He received his B.S. in mechanical engineering from the Air Force Engineering University (1992). He received his Ph.D. degree in management science and engineering from Xi'an Jiao-tong University (2006). His research interests include modeling and simulation, emergency management and data mining.

From 2006 to 2009, he was engaged in postdoctoral research at Management School, Xi'an Jiao-tong University. From 2010 to 2016, He worked at the Air Force Engineering University. He joined Beijing Institute of Technology, Zhuhai School in 2017. He has published more than 50 papers and five monographs. He has served as cochair of program committee for some international conferences.

Shihui Wu was born in Wuhan (Hubei, China) in 1982. In 2004, he received a bachelor's degree in equipment management from the Air Force Engineering University. He received the M. S. and Ph.D. degrees in management science and engineering from Air Force Engineering University, Xi'an, P.R. China in 2007 and 2010, respectively.

He is currently an assistant professor in Air Force Engineering University. His research interests focus on decision theory, simulation optimization, discrete event simulation, and so on.

Zhihua Song was born in Hebei, China, in 1982. In 2004, he received a bachelor's degree in missile engine from the Air Force Engineering University. In 2010, he received a Ph.D. degree in operations research from the Air Force Engineering University.

He is an associate professor at the Air Force Engineering University. He has published a number of monographs, including: Modeling and Simulation of equipment management (Beijing China: Electronic Industry Press,2020), weapon equipment data mining technology (Beijing, China: National Defense Industry Press,2018), Fundamentals of operational research (Xi'an China: Xidian University Press,2020), etc. His research interests include intelligent decision making, modeling and simulation.