How Data Mining Techniques Can Improve Simulation Studies

Manel Saad Saoud, Abdelhak Boubetra, and Safa Attia

***Abstract*—Researchers take years and even decades of observation in order to analyze socio-economic phenomenon. Whereas the agent-based modeling simulation (ABMS) provides a new issue by offering the possibility to create virtual societies in which individuals and organizations are directly represented with their observed interactions. As it is known simulation generates and consumes a large amount of data. The analysis of these data which may contain implicit and hidden information will always remain a very difficult phase in the ABMS. As a solution to this problematic, the use of data mining techniques can contribute to the right analysis of the phenomena that emerges in these systems. In this paper we aim the investigation of agent-based modeling simulation and data mining techniques.**

***Index Terms*—Agent-based modeling and simulation, data mining, analysis behaviors, emergence.**

I. Introduction

The ABMS has demonstrated an efficacy in several scientific fields, especially in the socio-economic sciences where it creates artificial societies close to those of reality.

During the dynamic evolution of these artificial societies the agents involve a large amount of data. The search in this data is of great importance to study and analyze phenomena that emerges in these virtual societies. Nevertheless, the exploitation of the concealed relationships and the analysis of emerging behaviors which may enclose non-explicit and hidden information are judged as a difficulty in the multi-agent modeling simulation. As a solution to this problematic, the use of data mining techniques can contribute in the right investigation of these virtual societies.

In conjunction data mining will be used to extract from a large volume of data knowledge or hidden information, previously unknown, potentially useful and interesting information. The application of these techniques of data mining can uncover behavioral patterns and therefore non-trivial knowledge from the simulation data. The analysis of these results allows to correct and to improve the quality of the simulation study.

In spite of the investigation of simulation and data mining techniques is fairly a new area, various researches in different fields are talking about it. These studies are related to maintenance of aircraft engines [1], neurosciences [2], higher education [3], social sciences [4] and others.

This research suggests benefits to investigate knowledge involved in socio-economic simulation, using data mining methods.

This paper is organized as follows. Section I, reviews the ABMS. In Section II, we discuss about data mining methodology. Section III that follows describes the existing background about data mining and simulation integration. Our approach to the investigation of simulation process and data mining techniques is given in Section IV.

II. Agent-Based Modeling Simulation

ABMS of socio-economic phenomena is to model different societies with artificial agents by placing them in a virtual society simulated via a computer to observe their behaviors. Each agent represents an individual (or an organization) that can perceive and react to events and interact with other agents existing in the same environment, taking into account its beliefs, objectives, etc.

The ABMS is mainly based on a set of autonomous entities called agents. According to [5] an agent is a computer system having basically four properties: autonomy (the possibility to operate without the intervention of a human or another agent, and it has some sort of control over his actions and its internal state), social ability (agent interacts with other agents via an agent communication language). These actions are known as reactivity (the ability to perceive the environment and respond to changes that occur in it) and pro-activity (the possibility to demonstrate a behavior determined by its goals rather than as reaction according only to its environment).

These properties are the keystone of agent-based systems to model complex phenomena. However, before developing any multi-agent system it is important to determine the degree of reasoning of each agent. Related to [5] there are three main types of agents:

1) Cognitive agent
2) Reactive agent
3) Hybrid agent

With a simple and predefined behavior, reactive agent responds only to a simple environmental stimulus. Since it does not have a memory, neither a complete representation of the environment and other agents, it is not able to take account of his past actions and maintaining its internal state. Cognitive agent is an intelligent agent that has a necessary knowledge base to achieve its tasks and manage its interactions with other agents and the environment. It has explicit goals and plans to decide its actions. Hybrid agent combines the two previous types; it aims both the quick reflex of the reactive agent and the reflected behavior of cognitive agent.

In order to create a system similar to a human society and to correctly model the socio-economic phenomena, our

Manuscript received June 14, 2013; revised September 20, 2013.

The authors are with the Department of Computer Science, University of Bordj Bou Arreridj, Algeria (e-mail: m.saadsaoud@univ-bba.dz, a.boubetra@univ-bba.dz, s.attia@univ-bba.dz).

DOI: 10.7763/IJCTE.2014.V6.828

multi-agent system (MAS) is composed of two architectures of agents: cognitive and reactive. The system is illustrated in Fig 1.

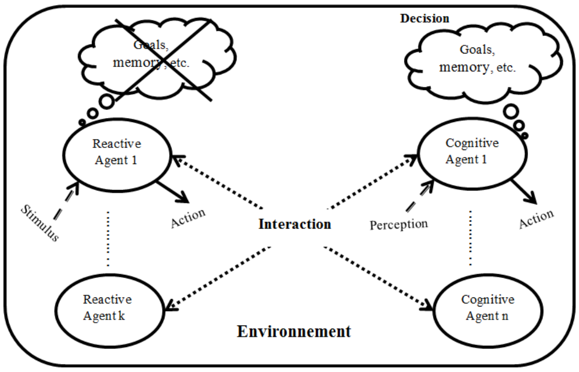

Fig. 1. A multi-agent system based on cognitive / reactive agents.

III. Temporal Data Mining

Generally the data produced by the simulation characterize the evolution of a given system over time, so we are talking about time series. In this work we will recover and keep these time series in a temporal database in order to analyze them using data mining techniques.

To exploit the hidden relationships and the emerging behaviors of this huge amount of temporal data, the use of data mining techniques (Knowledge Discovery in Database (KDD)) is the best approach that can be pursued since it represents the identification process of hidden and interesting information. In other words, it is the discovery and the extraction of knowledge from a large volume of data.

According to [6] the data mining process includes the following steps as shown in Fig. 2:

1) Pre-processing: Includes data selection, Data Cleaning, Data Transformation, management of missing values, etc.
2) Data Mining: The heart of the KDD process is to look for patterns structuring the data, discover the explanatory or predictive data models.
3) Post-processing: Evaluation and interpretation of knowledge extracted.

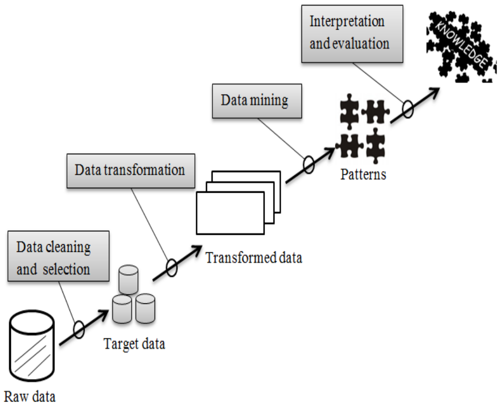

Fig. 2. The Data mining process.

The data mining algorithms are divided into two main classes [7].

1) Predictive algorithms (supervised algorithms) are to build predictive models from a database of tagged data. These models are used to forecast new untagged data.
2) Descriptive algorithms (Unsupervised algorithms) belong to the modeling of knowledge discovery. This task is descriptive rather than predictive and the goal is to detect trends in the current data without requiring prior learning.

Different techniques of these algorithms are used to analyze the simulation data [1], [8]:

1) The association rules: the goal of this method is to discover correlations (associations) between the hidden and interesting elements (items) in large databases. The association rules are of the form: If antecedent, then consequent.
2) Classification methods: refer to a set of statistical or heuristic methods that allow the separation of a set of entities (characterized by a representative attributes) into groups (classes) and the attribution of new entities in these predefined groups (classes). The final result is a set of rules used to assign new observations to one of these classes.
3) The classification methods include Bayesian Networks, Artificial Neural Networks and decision trees, etc.
4) Clustering: is considered as an unsupervised classification method since it allows the discovering of all database groups (clusters) in a way that maximizes the similarity between elements of the same group, and minimizes the similarity between elements of different groups.

IV. Related Works

There is a growing body literature regarding the investigation of data mining techniques and simulation studies.

The authors in [8] describe how the different data mining techniques can be used to assist in analyzing the performance of the predictions obtained from the simulation scenarios. In this work, the authors identified the clustering as a useful technical analysis and appropriate to analyze the data from the simulation.

One of the great challenges of simulation studies is to determine the set of inputs variables to produce the optimal outputs. The approach proposed in [9] provided the information (knowledge) of these variables and the logical relationships connecting them, using data mining tools. This information can be used as the basis of the development and the optimization of the simulation scenarios.

The study presented in [10] demonstrated an approach that uses the Monte Carlo simulation and the analysis of gray relationships taking into account the fuzzy data. The Monte Carlo simulation is used as a data mining technique to measure the impact of fuzzy decision trees models (using categorical data) compared to the decision tree based on continuous models. The proposed approach is applied to a case study with a set of typical commercial data.

In [11], the authors studied the ways in which data mining techniques could be successfully applied to the agent-based modeling and simulation, in order to exploit the hidden relationships and the emerging behaviors. They distinguished between endogenous and exogenous application of data mining techniques. Endogenous modeling aims to use data mining techniques to improve the performance of agents participating in the simulation. On the other hand, exogenous modeling is primarily focused on the use of data mining techniques to analyze and reveal interesting trends in the simulation results that could help to better model the system behavior.

To develop and maintain the parameters of the estimated costs, a methodology for the life cycle cost was developed in [1]. Authors combined simulation and data mining to help decision makers of the US Department of Defense to take decisions related to the aircraft engine maintenance.

The proposed approach in [12] is based on a dynamic knowledge extraction model that identifies the relevant inputs of the simulation and discovers the impact of their relationship on the system performance. This model also allows the use of information learned during the optimization process due to separate good from bad solutions, ensure that only promising solutions should be evaluated in future iterations.

In [3] the authors proposed an intelligent framework based on the integration of simulation and data mining techniques where the outputs of the simulation are transferred to a data warehouse using the steps of the ETL (Extract, Transform, and Load). This data warehouse can produce: different data stores, multidirectional cubes or simple aggregate data. The data mining techniques are applied intelligently on these data sets to facilitate the extraction of relevant information and knowledge from this huge amount of data. The proposed framework has been validated through two case studies, the first one on car market demand simulation, and the second one was presented in order to demonstrate how to apply data mining and simulation to ensure quality in higher education.

In the neurosciences data integration is a very difficult phase, because it is necessary to organize a large amount of data around some functional hypotheses. On the other hand, it is often noted that the simulation provides explicit hypotheses for a particular system, so it can provide an organizational orientation which can be exploited to form important hypotheses [2]. In this study the authors have developed a neural query system (NQS) in the NEURON simulator by providing: a system of relational database, a research function and data mining tools. NQS is used in the simulation to manage, control and evaluate the model parameters. More importantly, it is used to extract knowledge from the simulation data in order to compare them with neurophysiology.

The data mining techniques are also applied to simulation based multi-objective optimization [13], this work investigates the ways of extracting knowledge from the simulation, in order to obtain information can support decision makers to take the right decisions in order to optimize the manufacturing process.

In [14] the authors presented an overview of bi-directional integration of Data Mining (DM) and the agent-based modeling and simulation (ABMS) [application of DM in ABMS / application of ABMS in DM]. This study proposes a conceptual framework and a presentation of the advantages that can be provided by this integration.

In [4] a methodological approach for multi agents modeling and simulation using data mining techniques was presented. The authors proposed an intensive use of these techniques of data mining to improve and develop agent-based models. Such use is illustrated by a study of a mental model for understanding the evolution of several factors in Spanish society from 1980 to 2000, focusing on social values, especially religious and ideological values.

Comprehension of the correlation between inputs and outputs parameters of the simulation is critical to correctly analyze the behavior of the simulated system [15]. In this paper, an approach that combines simulation and data mining techniques has been proposed. The authors found firstly the relationships and the correlation between the input and output parameters, then they added some additional information to the outputs using data mining techniques, such information will be used later to classify the results of the simulation.

In [16] a methodology based on the integration of clustering techniques in the design of multi-agent simulation. To use the observation of real word agents to model simulated agents this approach was proposed. It has been implemented experimentally on a simple case study involving the simulation of human activities in a university building lobby.

V. Proposed Approach

Simulation can produce a huge amount of data sets. Analyzing and exploiting such data sets is considered as a very difficult challenge for simulation studies. In conjunction data mining techniques are credible tools that extract from a large data, sets of knowledge, relationships, anomalies, etc. Our approach aims to investigate the ABMS and data mining techniques as shown in Fig. 3.

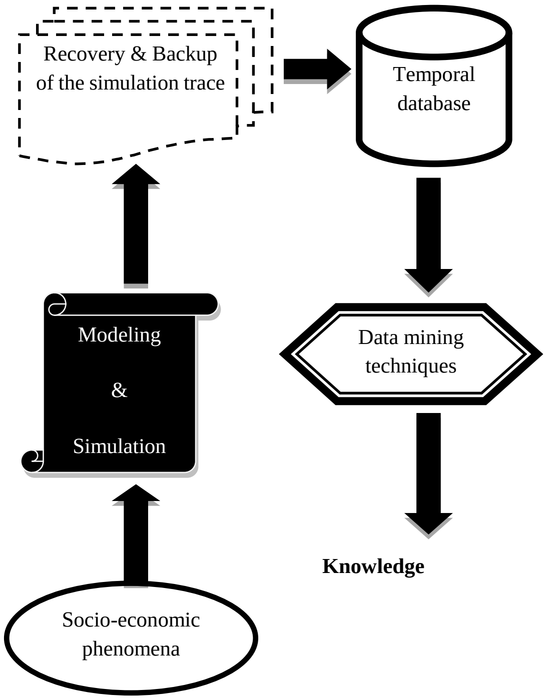

Fig. 3. The proposed approach based on the integration of simulation and data mining.

As a first step of our proposed approach, simulation data are collected and stored in a temporal database in order to facilitate the task of data mining techniques which will be applied later on these stored data.

Our temporal database represents the history of each agent (perception, action, interaction, etc.) as well as all changes of states of our system, as illustrated in Fig. 4.

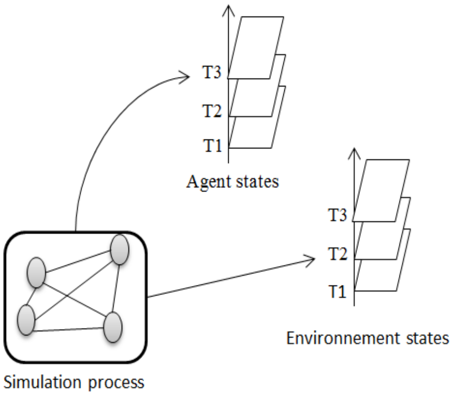

Fig. 4. Temporal data of our simulation process.

Our model is based primarily on the creation of an artificial agent (which we named: Collector Agent). The role of this latter is to recover the simulation trace by collecting the data generated during the simulation as well as its results.

B. Inputs Analysis

Most of previous studies mentioned above handle only the analysis of outputs simulation, whereas the quality of inputs data affects the outputs quality. In this paper we will work on the inputs and the outputs of the simulation. Hence, two databases are manipulated; the first one is used to save the original data (inputs) and the second will be generated during the progress of simulation as shown in Fig. 5.

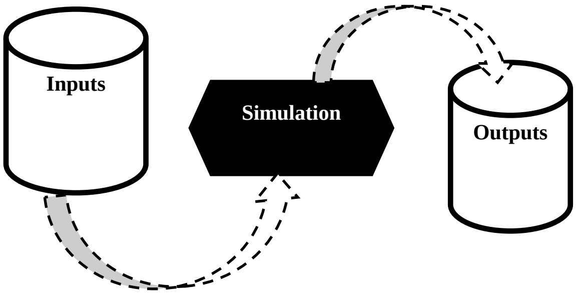

Fig. 5. Inputs/outputs of the simulation system.

Data collected for simulation is often susceptible to be noisy, incomplete, inconsistent, corrupted, inaccurate, etc. To assure the quality of data and, consequently, of our simulation study, raw data will be prepared by applying the following methods as illustrated in Fig. 6:

1) Data cleaning.
2) Missing data management.
3) Data transformation.
4) Data selection.

The key step to get desired results of this preparation process is data cleaning, where it represents the process of detecting and correcting or removing of errors in datasets.

The key step to get desired results of this preparation process is data cleaning, which it represents the process of detecting and correcting or removing errors in datasets.

Missing data is a major problem in socio-economic sciences. This absence of information affects the efficiency of data analysis methods. Therefore, we will create a management that assures the handling of missing data without losing the power of our original data.

For a good mining we will transform or consolidate our data into an appropriate form by performing normalization and aggregation operation.

In the last step of this preparatory process, target datasets and relevant data to the analysis task are retrieved from database.

After the application of these methods on our raw data, we will obtain a database prepared and ready to be used by the simulation process.

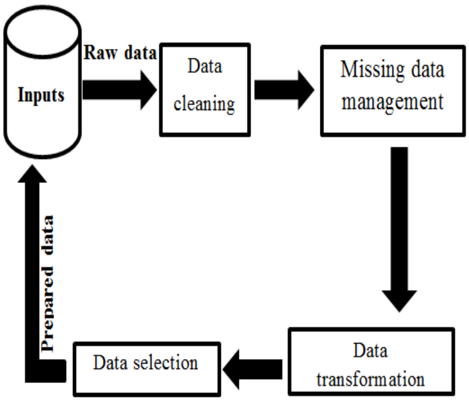

Fig. 6. Data preparation process.

VI. Conclusion and Perspectives

In this paper we have pointed out the investigation of data mining techniques and the agent-based modeling and simulation of socio-economic phenomenon. Besides, it has been reviewed how ABMS can benefit from analysis methods to deal with its open problems, represented mainly in the enormous amount of used and generated data.

Moreover, the analysis of such huge quantity of data is a big problem that continues to plague simulation studies. Data mining provides powerful techniques that can be used to tackle this problem.

This paper presents an approach based on the collector agent. This agent is responsible for simulation data collection.

Also, we proposed a process of input and output data handling.

As a future work, we plan, on the one hand, to implement this approach in a multi agent modeling and simulation environment, and, on the other hand, to integrate our collector agent in this environment.

References

[1] M. Painter, M. Erraguntla, G. Hogg, and B. Beachkofski, “Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management,” in *Proc. the 2006 Winter Simulation Conference*, pp. 1253-1260.
[2] W. Lytton, “Neural query system: Data-Mining from within the NEURON simulator,” *Neuroinformatics*, vol. 4, pp. 163-176, 2006.
[3] M. Alnoukari, A. El Sheikh, and Z. Alzoabi, “An integrated data mining and simulation solution,” *Handbook of Research on Discrete*

Event Simulation Environments: Technologies and Applications, ch. 16, 2008.

[4] J. Arroyo, S. Hassan, C. Gutiérrez, and J. Pavón, "Re-Thinking simulation: A methodological approach for the application of data mining in agent-Based modeling," *Computational & Mathematical Organization Theory*, vol. 16, pp. 416-435, December 2010.

[5] M. Wooldridge and N. R. Jennings, "Intelligent agents theory and practice," *Knowledge Engineering*, vol. 10, no. 2, 1995.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *American Association for Artificial Intelligence*, *AI Magazine*, pp. 37–54, 1996.

[7] M. Gracia, I. Roman, F. Penalvo, and M. Bonilla, "An association rule mining method for estimating the impact of project management policies on software quality, development time and effort," *Expert Systems with Applications*, vol. 34, pp. 522–529, 2006.

[8] C. Morbitzer, P. Strachan, and C. Simpson, "Data mining analysis of building simulation performance data," *Building Services Engineers Res.Technologies*, vol. 25, no. 3, pp. 253–267, 2004.

[9] T. Brady, and E. Yellig, "Simulation data mining: A new form of computer simulation output." in *Proc. the 2005 Winter Simulation Conference*, 2005.

[10] D. Wu, and Z. Y. Dong, "Data mining and simulation: A grey relationship demonstration," *International Journal of Systems Science* vol. 37, no. 13, pp. 981–986, October 2006.

[11] M. Romondino and G. Corrend, "MABS Validation through repeated executing and data mining analysis," *International Journal of Simulation Systems, Science & Technology*, vol. 7, no. 6, 10–21, 2006.

[12] M. Better, F. Glover, and M. Laguna, "Advances in analytics: Integrating dynamic data mining with simulation optimization," *IBM. Journal of Research and Development*, vol. 51, no. 3/4, 2007.

[13] C. Dudas, N. G. Amos, and H. Boström, "Information extraction from solution set of simulation-Based multi-Objective optimisation using data mining," in *Proc. Industrial Simulation Conference*, 2009.

[14] O. Baqueiro, Y. Wang, P. McBurney, and F. Coenen, **"**Integrating data mining and agent based modeling and simulation," in *Proc. the 9th Industrial Conf. on Advance in Data Mining. Application and Theoretical Aspects*, pp. 220-231, 2009.

[15] S. Ghasemi, M. Ghasemi, and M. Ghasemi, "Knowledge discovery in discrete event simulation output analysis," in *Proc. INCT 2011*, CCIS 241, © Springer-Verlag Berlin Heidelberg, pp. 108–120, 2011.

[16] I. Saffar, A. Doniec, J. Boonaert, and S .Lecoeuche, "Multi-Agent simulation design driven by real observations and clustering techniques," in *Proc. 23rd IEEE International Conf. on Tools with Artificial Intelligence - ICTAI 2011*, Boca Raton, Florida, USA, November 7-9, 2011.

Manel Saad Saoud was born in Bordj Bou Arreridj, Algeria in 1988. She received her computer science bachelor degree from Bordj Bou Arreridj University in 2009. She received her master's degree from Bordj Bou Arreridj University in 2011. Now she is a Ph.D student at the University of Bordj Bou Arreridj. Her research interests include Data mining and Agent Based Modeling and Simulation.

Abdelhak Boubetra was born in Bordj Bou Arreridj Algeria in 1959. He is an associate professor and the Head Research Group of computer simulation in the computer Science department of the University of Bordj Bou Arreridj in Algeria. He received a computer science engineer degree from the University of Constantine (Algeria), a master Philosophy degree from the University of Lancaster (U.K) and a doctorat degree from the University of Setif (Algeria) in computer science. His research interests include data bases, software engineering, and green IT.

Safa Attia was born in Setif, Algeria in 1989. She received her computer science bachelor degree from Bordj Bou Arreridj University (Algeria) in 2009. She received her master's degree from Bordj Bou Arreridj University in 2011. Now she is a Ph.D student at the University of Bordj Bou Arreridj. Her research interests include Agent Based Modeling and Simulation.