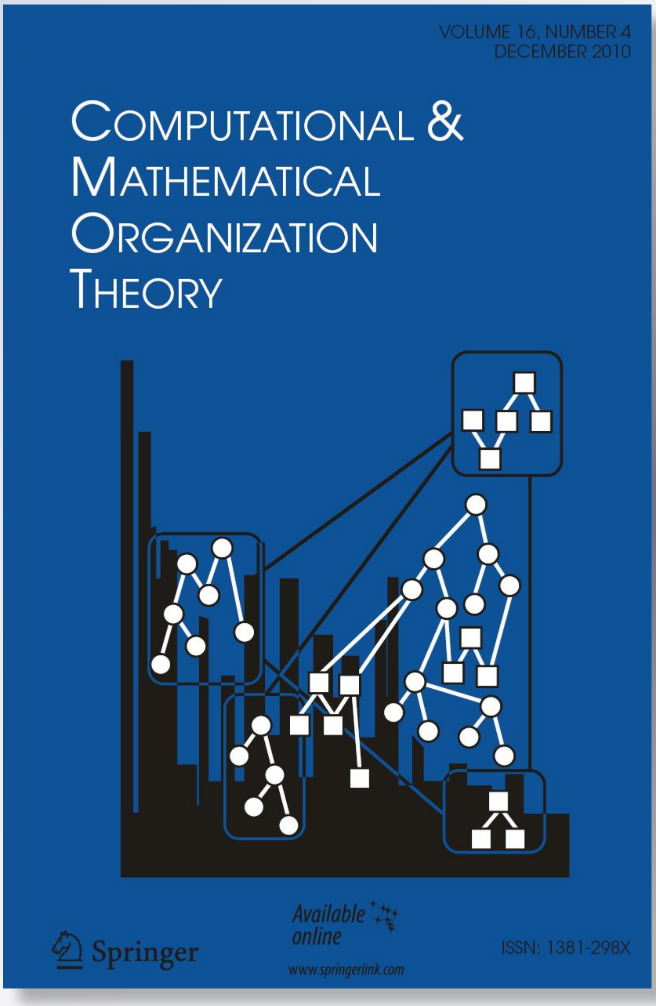
Re-thinking simulation: a methodological approach for?the?application of data mining in agent-based modelling

Computational and Mathematical Organization Theory

ISSN 1381-298X
Volume 16
Number 4

Comput Math Organ Theory
(2010) 16:416-435
DOI 10.1007/s10588-010-9078-y



Comput Math Organ Theory (2010) 16: 416–435
DOI 10.1007/s10588-010-9078-y

Re-thinking simulation: a methodological approach for the application of data mining in agent-based modelling

Javier Arroyo · Samer Hassan · Celia Gutiérrez · Juan Pavón

Published online: 21 October 2010

Abstract Agent-based models informed by empirical data are growing in popularity. Many models make extensive use of collected data for the development, initialisation or validation. In parallel, models are growing in size and complexity, generating large amounts of output data. On the other hand, Data Mining is used to extract hidden patterns from large collections of data using different techniques. This work proposes the intense use of Data Mining techniques for the improvement and development of agent-based models. It presents a methodological approach explaining why and when to use Data Mining, with a formal description of each stage of the corresponding process. This is illustrated with a case study, showing the application of the proposed approach step by step.

Keywords Agent-based modelling · Clustering · Data-driven · Data mining · Social simulation

1 Introduction

There are recent efforts for injecting data into agent-based models (ABM) (Hassan et al. 2010). An increasing number of models follow a data-driven approach (Boero and

J. Arroyo (✉) · S. Hassan · C. Gutiérrez · J. Pavón
GRASIA: Grupo de Agentes Software, Ingeniería y Aplicaciones, Dpto. Ingeniería del Software e Inteligencia Artificial, Facultad de Informática, Universidad Complutense de Madrid,
28040 Madrid, Spain
e-mail: javier.arroyo@fdi.ucm.es

S. Hassan
e-mail: samer@fdi.ucm.es

C. Gutiérrez
e-mail: cegutier@fdi.ucm.es

J. Pavón
e-mail: jpavon@fdi.ucm.es

Squazzoni 2005), using a variety of empirical methods in combination with ABM (Janssen and Ostrom 2006). For instance, the IMAGES project (Deffuant et al. 2002) explores opinion dynamics from a data-driven approach, using both quantitative and qualitative data; or Galán's water demand model (Galán et al. 2009b), that makes an intensive use of data, specially coming from databases and survey data.

Those models can use empirical data in multiple ways: for building agent's characteristics directly from data types; for initialising the simulation configurations; for validating the results. Besides, this trend comes together with the growing size and complexity of models (Galán et al. 2009a), generating large amounts of output data. Frequently, those collections of data (input or output) are complex both qualitatively (their relations and characteristics) and quantitatively. Given the size and complexity of these data sets, statistical tools are needed to extract knowledge from them. In some fields, statistical methods are routinely applied to analyse the output of the simulation. This is the case of agent-based artificial stock markets, where statistical tests are used to check if the prices exhibit it several well-known empirical features of real stock markets, such as predictability, volatility and volume relations (LeBaron et al. 1999). However, the benefits from the application of techniques of statistical learning and Data Mining (DM) could be extended to other fields and to other data sets used in the development of the ABM. This work aims to explore the use of DM in the process of development of agent-based models, proposing a methodological approach for its application, result from our experience.

Section 2 reviews the use of DM in multi-agents systems and agent-based models, while Sect. 3 provides methodological guidelines for the general application of DM in agent-based models. Afterwards, a case study following the proposed guidelines is presented, showing their application step by step (Sect. 4). In the final section, some conclusions are extracted from the exposed approach.

2 Data mining for agents

As it is explained by Remondino and Correndo (2006), DM techniques have been used in the development of multi-agent systems, either as a component of the agents or as a tool for their design. In the first case, DM techniques are encapsulated as a component or as a resource that is used by agents to extract knowledge that will guide their behaviour, i.e., to support their decisions (Sen and Sekaran 1996). In the second case, DM techniques are used in the design process, usually to explore the system's output in order to reveal hidden patterns or even system inconsistencies. Thus, a MAS could be re-designed and improved, or validated and supported by new evidence. This is specially useful when the amount of information is huge, as it is the case in massive MAS to analyse agent communication (Botía et al. 2004).

Although DM could be applied similarly in social simulation, nowadays there are few precedents. Some tangential examples can be mentioned, though. Inherited from the mentioned MAS examples, there are ABM which analyse the agent communication and behaviour through DM techniques, such as Gostoli (2008), who use decision trees (a DM technique) to allow agents to acquire knowledge based upon experience.

On the other hand, in the field of social networks, closely related to ABM, there are two areas where DM specially finds a fertile ground. The first one is the consumer behavioral patterns, a classical context of the technique, which is boosted through the social network pattern-finding need (Domingos and Richardson 2001). The other area, which is quickly growing in popularity, is the analysis of online social networks as their electronic nature provides huge amounts of data to explore (Ahmad and Teredesai 2006).

However, some works began to tackle the methodological implications of DM in agent-based modelling. In the classical social simulation approach (Gilbert and Troitzsch 1999), empirical data only play an important role on the validation step, letting the model building be guided mainly by a theoretical abstraction process. Moreover, nowadays the majority of the models do not use formal methods for validation such as statistical techniques, as found by Heath et al. (2009), who also push for introducing statistical methods and tools into ABM. On the other hand, data-driven simulation encourages the use of data also in the design and initialisation stages, as explained in Hassan et al. (2010). As a natural extension of this approach, the new knowledge found in the data set through DM can also join the mentioned data to influence the ABM design. In fact, DM is included in the catalogue of methods suggested to deal with data, but no examples are provided and few details about it are given.

In Kennedy et al. (2007) it is proposed the use of a DM technique, association rules, to validate the simulation output and analyse the real-world data. These rules would discover unexpected relationships among the categorical variables both in the simulation and empirical data. The model is validated checking the existence of inconsistencies among the rules generated for each data set. However, again no results of the approach are shown.

The aforementioned works suggest that ABM can benefit from a more intensive use of data and data mining in the whole process. However, these are only the first steps in a promising area: a more comprehensive approach showing how DM can help to build simulation models informed by data can be proposed. This is the aim of the next section.

3 Towards a methodology for data mining-assisted agent-based modelling

The previous section reviewed the DM applications in the field of ABM, which, as has been shown, are sparse at present. More generally, as is shown in the review by Heath et al. (2009), statistical techniques are seldom used in ABM for validation purposes. This fact is surprising given the widespread use of statistical techniques in other simulation contexts (Ripley 1987). The authors surmise that the reasons behind are that in some cases ABM is used to simulate systems whose output cannot be statistically analysed and that the ABM modellers use different validation criteria than those used in other simulation contexts. Heath et al. (2009) consider that this situation highlights the need to explore the use of statistical validation in ABM in order to strengthen the proposed models.

Following this line, this section aims to provide a methodological approach where statistical methods are used to assist not only the validation of an ABM, but the whole development process. Statistical learning and DM methods are especially indicated to extract useful information from great amounts of data. These data may be the available information of the real world or may be obtained as a result of the simulation runs. In the first case, the extracted information should guide the building of the model. While in the second case, it should be used for validation purposes. In both cases, statistical methods can be applied to enhance the agent-based modelling process, as it is shown below.

3.1 Methodology in ABM: a review

The methodology to build a simulation model in social sciences has been widely documented. In the classical approach exposed by Gilbert and Troitzsch (1999), data arise from two different sources: there is data collected from the real world and there is data that results of the simulation process. The first kind of data may be survey data on the variables of interest and is used to build the model. While the second kind of data may appear in the form of log files and is used to validate the model by means of, for example, statistical tests.

In the particular context of ABM simulation the same concern about the definition process of simulation can be found. Edmonds (2001) attempts to define the ABM modelling framework. This framework is later complemented by Galán et al. (2009a) that assign the roles usually involved in the ABM modelling process, i.e. the thematician, the modeller, the computer scientist and the programmer, to each of the stages of the process. Heath et al. (2009) work also concerns the ABM development process, with the aim to emphasise the role of validation at both a conceptual and an operational level.

We share the same concerns of these works about the modelling process and we firmly believe that ABM can take profit of the intensive use of data to build and validate the model (Hassan et al. 2010). Thus, we propose our view about the ABM methodology emphasising on the role of data and, more precisely, on the benefits derived of their systematic use to propose and refine the model.

3.2 Introducing the DM-based methodological approach

The current widespread use and power of computers makes possible to store large amounts of data and to analyse them to extract useful information. Data are more easily available and they allow us to propose more robust simulation models. In order to do so, DM and statistical methods are the tools needed to uncover the information relevant for the simulation process. Figure 1 shows a diagram that represents the iterative process of DM-assisted ABM that it is proposed.

The diagram assumes that there is enough empirical data to guide the process. More precisely, it is assumed that there are available data of the real world at several periods of interest and that the aim is to analyse the evolution of a phenomena of interest through those periods. An initial and a validation point should be available, but it may be possible to have several validation points that will allows us to check the deviation between the simulation and the real world at several time points. The process described by the diagram is specially indicated for simulations whose aim is to provide accurate predictions about the system, that is, the *Predictor* simulations, according to the framework proposed by Heath et al. (2009). However, in some cases, it is not possible to apply the proposed approach. This is the case of simulation studies where there is no data available from the real world or where the phenomenon of interest cannot be validated in a quantitative way but in a qualitative manner. These studies would mostly correspond to *Generator* simulations, according to the Heath et al. taxonomy, where the purpose of the simulation is to generate plausible hypotheses about the behaviour of the real system. In the case of the *Mediator* simulations the proposed methodological approach may be applicable, depending on the availability of data.

In some cases, the diagram stages of data collection and analysis of the initial point can be considered externalised. This is the case of the data-driven approach shown in Hassan et al. (2009a) where the distributions used to initialise the agents are empirically-supported by other research studies where empirical data was analysed. There are many other cases where data is available but is not analysed. The proposed methodological approach provides a way to incorporate the knowledge of these data into the model.

The desirable route to study through agent-based modelling a problem where enough data is available would be: first, to select and analyse empirical data from past situation; second, use the resulting information to design the ABM and, possibly, to use some of the data to initialise the ABM; third, run the simulation; and validate the simulation results against the real world data in the different validation points. The proposed process is cyclical. It means that the information learned in each iteration can be used to refine the ABM in case that the validation is not satisfactory. As a result, it is expected to obtain an increase in the ABM accuracy with each iteration and it is possible to study the effect of different set-ups in the results. In the following paragraphs the sequence of the methodological diagram is detailed.

3.3 The stages of the process flow

The source data could be obtained by different means: surveys, panels, interviews, statistics, or by any other social research tool. The resulting *data collections* should contain, if not an accurate depiction of the real world at the initial and validation points, at least some interesting facts that can guide the model building stage. In order to extract these facts, both data collections should be analysed. In the classical methodology, following to Galán et al. (2009a), this task is done by the thematician. However, we consider that a data mining expert (although such role might be played by the same thematician or modeller if they could) can provide a valuable help in the analysis process. Both experts should work in close collaboration and the knowledge provided by each of them should eventually reach consensus, i.e. data mining should confirm the theories of the domain expert or, alternatively, the expert's theories may be refined according to the DM conclusions. Similarities can be drawn between this part of the diagram and the marketing research process (which often uses DM). A marketing research is conducted to achieve an increased understanding of the subjects relating to marketing products and services. The marketing research process consists of a systematic gathering, recording, and analysis of data about the subject matters. ABM studies based on surveyed data or in interviews can follow a similar approach to that followed by marketing research to uncover hidden knowledge.

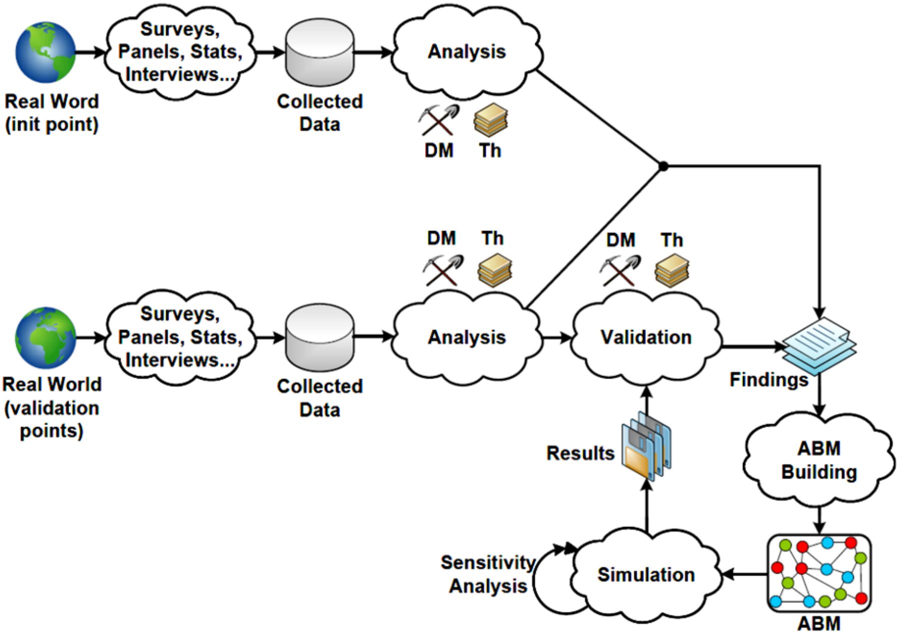


Fig. 1 Diagram of the iterative data mining-assisted agent-based modelling

Following the diagram in Fig. 1, once the data is analysed (the analysis stage will be detailed in-depth in the next subsection), the resulting *findings* are used to *build the ABM*: formalise, design, implement and initialise it. Once the model is ready, the *simulation* process is carried out. In this stage it is recommended to fine tune the implemented system by means of a sensitivity analysis or at least by intensive testing. These stages are extensively documented in the literature (Gilbert and Troitzsch 1999; Gilbert 2008), together with other works that deepen the methodological pitfalls in ABM and how to avoid them (Richiardi et al. 2006).

After the simulation, the *validation* stage determines if the model performed in the manner intended, i.e. if it accurately represents the real system. This stage is considered critical (Balci 2004; Sargent 2007), but it is too often disregarded in practise, according to the survey in Heath et al. (2009). This survey shows that statistical validation is not commonly used in ABM. However, in the specific subdomain of agent-based computational economics, such validation is far more common.

Statistical methods used to validate ABMs usually consist of hypothesis tests or confidence intervals. However, a richer analysis can be conducted if the model generates a log file that records not only the final state of the agents but also the evolution of their state throughout the simulation period. This log file can be analysed to determine if the agents state evolved in a consistent way. Moreover, it is useful to check if the final state of the simulation is an accurate representation of the real world state at the considered validation points.

The more data is stored in the log file, the more possibilities exist to apply statistical and DM methods to validate our model. For example, if enough historical data exist, cross-validation can be applied. That is, part of the data can be used to build the model and the remaining data can be used to check whether the model is accurate or not. This would be very useful to determine if *predictor* simulation models are useful to forecast.

At the *validation* stage, the domain expert is required again to extract trustworthy conclusions from the data. And again, DM can aid this task substantially. If the conclusions obtained in the validation agree with those obtained in the analysis of the real world data, then the model will be considered a good representation of the simulated process. If they do not, the conclusions obtained in the validation should be incorporated to the previous base of knowledge (coined 'findings'), and the ABM should be redesigned according to them. The iterations are intended to provide a better understanding of the model with each iteration and consequently of the situation being simulated. The use of DM techniques to analyse simulation results is also encouraged in Remondino and Correndo (2006).

In the next section, more information is given about the *analysis* stage and the role that the domain and DM experts play.

3.4 The analysis stage

The analysis stage is central to the approach that is proposed in this work. This stage involves both a data mining and a domain expert. As it was shown in Sect. 2, the role of DM in ABM is becoming more prominent. Thus, such role should be specified and well-definition, which is the purpose of this section.

Figure 2 shows a diagram that represents an in-depth focus of the *analysis* stage from Fig. 1. This stage is divided into three sub-stages which resemble the three steps of DM: preprocessing, analysis, postprocessing. In the first sub-stage, the aim is to select the variables and information that are relevant to the problem, and transform them appropriately for the following sub-stage (normalisation, discretisation, etc.). The selection can be initially done by the domain expert, but DM can also be applied, e.g. to analyse if redundant variables have been chosen. Moreover, if the problem

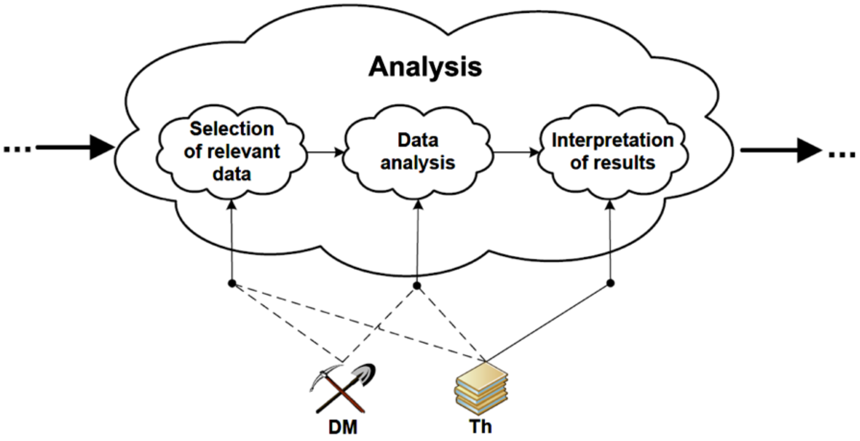


Fig. 2 The analysis stage in iterative data mining-assisted agent-based modelling

of interest is linked to a (dependent) variable, DM can also detect if relevant (independent) variables have been overlooked by the expert. If the dependent variable is numerical, this problem could be solved by a regression method, while if it is categorical, a classification method should be applied.

The second sub-stage is the *data analysis* sub-stage. If either the number of variables or the number of individuals within the data is very high, then some kind of DM is essential, as a single human cannot handle such amounts of data. Thus, the analysis must be guided by the domain theories, but at the end, they should endorse the DM results.

DM tasks are classified into descriptive and predictive. The descriptive is used to characterise the properties of the considered data, while the predictive facilitates inferring the behaviour of new data based on the current data set. The possible tasks carried out with DM include concept or class description, classification, prediction, extraction of frequent patterns and associations, etc.

Some examples of the use of DM methods in the analysis sub-stage follow:

- *Cluster analysis*, useful to discover relevant groups of individuals in the population. The resulting groups (*clusters*) could be used to discover hidden patterns which could be used to feed the model, e.g. defining the agent micro-behaviour.
- *Principal component analysis* can be used to reduce a high number of possibly correlated variables to a smaller number of uncorrelated variables that represent the underlying dimensions which describe the individuals. Thus, the resulting variables can be used to characterise the agents.
- *Time series methods* could be applied to study how variables of interest evolved over time. This would be useful to define a model that follows a similar evolution and to check whether the simulation evolved like the social process.
- The DM technique *association rules*, following Kennedy et al. (2007), may also be applied to discover hidden relationships in categorical data. These relationships would be rather valuable to model the agent behaviour. Moreover, it facilitates the check of whether these relationships are also present in the simulated data.
- Graphs make possible to represent sophisticated structures and their interactions, being able to characterise social networks. *Graph mining* methods are useful for analysing these structures and, for instance, extract frequent subgraph patterns.

In Remondino and Correndo (2006) the use of some other methods, such as the analysis of variance and regression analysis, are suggested.

Finally, the third sub-stage consists on the *interpretation of the results* in order to make decisions based on them. This process must be done by the domain expert. The findings obtained are added to the findings base along with other domain theories and the results of other experiments.

4 A case study: the Mentat model

4.1 Definition of the case study

The methodological guidelines described in the previous section for the application of DM techniques in ABM are illustrated with a specific model of a social system, Mentat[1] (Hassan et al. 2009b). The aim of Mentat is to understand the evolution of multiple factors in Spanish society from 1980 to 2000, focusing on social values, specially religious and ideological values. This period corresponds to the transition to democracy in Spain, and it is interesting because of the substantial shift in social values corresponding to the transition from the Catholic traditional values to the postmodern materialistic values.

After forty years of dictatorship, finished on 1975, the country was far from Europe on all indicators of progress, including the predominant moral values and modernisation level. However, the observed evolution of moral values since then has been analogous to its EU partners. Furthermore, the changes in Spain have been developed with a special speed and intensity during these two decades. Mentat focuses on demographics evolution as the main factor proposed to explain the observed changes: the change in the age structure of the population and the influence of a younger generation. The model aims to simulate the effect of cross-generational changes, focusing on these 'vertical' ones rather than on 'horizontal' influences (as it would be the case of homophily direct effects).

The hypothesis of the Mentat model is that values are influenced by a range of factors, including demography, economy, political ideology, religiosity, family/friend relationships, reproduction patterns, and stage in the life course. The simulation attempts to match the real evolution of several social indicators over time, expecting to provide an insight of the social process of values change in Spain.

There are multiple sources of empirical data to study this problem, as explained in Sect. 4.2. Thus, surveys, census and secondary data were used to design and build the Mentat model. Moreover, its complexity degree facilitates providing a rich output. For both reasons, it was an interesting case study for illustrating the methodological approach exposed in Sect. 3.

The simulation has been configured with a population of 3000 agents and simulated for a period of 20 years (from 1980 to 2000). The agents are able to grow, communicate, establish friendship and couple relationships, reproduce and die. These social relationships form a complex social network among the agents, which evolves over time both in topological structure and in link strength: agents-nodes appear (being birth) and disappear (dying), allowing new links to be established and old links to break; depending on the similarity of two friends, their friendship will strength over time (following a logistic curve). Besides, the spouse choice is determined selecting from the friends group. And both partners will give their characteristics to the newborns. A deep insight of the system is provided in Hassan et al. (2008).

4.2 Setting up the experiment

The Mentat model is strongly data-driven, using a variety of sources of empirical data (Hassan et al. 2008). The main data source is the European Values Study (EVS),

[1] The Mentat model has been programmed in Java using the Repast framework (Collier 2001). It simulates 3019 agents along 1000 steps (20 years) + X steps of warming-up, with an $X = \{100 \rightarrow 1000\}$. It uses a Torus-Grid of 98 × 98 cells, with a fixed density of 3.2 cells/agent. Additional details concerning its theoretical foundations or technical details can be found in Hassan et al. (2009b).

a survey repeated every 10 years in the most part of European countries (1980, 1990, 1999), which handles around a thousand variables.

The agents are initialised with data corresponding to the Spanish section of the EVS-1980: 2303 individuals that turn to be 2303 agents. Additional 717 agents have been introduced to cover the miss of underage individuals: as they cannot do surveys, they do not appear in EVS. But in order to keep a representative demographic pyramid of Spain, the correct percentage of individuals must be introduced using some statistical methods, as explained in Hassan et al. (2008).

On the other hand, empirical data from the EVS-1999 are used for validating the model simulation. Following Fig. 1, the EVS-1980 is the dataset used in the initial point of the procedure, while EVS-1999 is used here as the validation point. Both data sets will be analysed following the stages in Fig. 2 (see Sect. 4.3).

Besides, other sources of data, such as statistical equations (for the demographic dynamics) and qualitative studies (for the micro behaviour) were introduced into the model. For instance: age-related probabilities of having children (for example, a woman in her forties will have less chance than a 23 years old); regression equations to determine whether an agent searches for a partner or not; and time-varying transition matrices for life expectancy and the fertility rate (the birth rate in Spain fell from 2.2 in 1980 to 1.19 in 2000).

4.3 Applying the methodological guidelines

The methodological guidelines shown in Sect. 3 illustrate the general case: building a model from scratch and applying data mining in several stages of the modelling process. However, in the Mentat case the model was already built. Due to this fact, the methodological guidelines have been used to refine the model, and the *ABM Building* stage has been turned into a brief *ABM Review* stage. In such stage, the software has been adapted in order to provide a proper output to be analysed by the DM methods. As a result, the output module implemented yields a data table, filled with the variables for each individual, that will make possible to analyse in detail simulation results. The rest of the stages, Data Collection, Analysis, Simulation and Validation, are carried out normally. More details will be given in the following subsections.

By following the methodological guidelines shown in Sect. 3, we try to shed some light on how the data is structured and its evolution in the studied period. Note that by data we mean both simulated and real-world data. More precisely, the aims are the following:

- Find groups of similar individuals in the data that could be related to conceptually consistent groups in the sociological problem under study
- Determine which variables are relevant to characterise the aforementioned groups
- Analyse whether the simulated data and the real-life data have the same structure, i.e., the sociological groups that can be found in both data sets are identical in terms of size and features

According to the aims pursued, the nature of the analysis is exploratory, i.e. the aim is not to test some hypotheses but to help to understand better the data structure. Moreover, since we want to group the individuals and they are not labelled by a target variable that assigns individuals to previously defined groups, we cannot use a classification method such as decision trees, linear classifiers or nearest-neighbours methods. Thus, an unsupervised learning method, i.e. a method suitable to analyse unlabelled examples, will be used.

Cluster analysis is an appropriate method to find groups in a data set. More precisely, cluster analysis divides a set of observations into subsets, called clusters, so that individuals in the same cluster are similar according to a similarity criterion, while individuals in different clusters do not show relevant similarities. In our case study, the aim will be to identify groups of people with similar social characteristics.

In cluster analysis, the variables used to represent the individuals play a key role. If non-relevant variables are considered, these "extra" dimensions could worsen the quality of the resulting clusters, e.g. if non-relevant variables, such as eye colour or weight, are considered along with other relevant variables, then the similarity criterion among individuals will also take these features into account and consequently will distort the results. In our case study, the variables considered for the cluster analysis will be those that are used to represent the sociological features of the individuals in the agent simulation.

The insight gained with the cluster analysis can also be used to validate the simulation results of the model. In our case, the validation of the simulation will consist of analysing the groups of individuals using cluster analysis in both the survey data and the simulated data in the validation points. If the model is correct, social groups similar to those found in the EVS-1999 data should appear in the simulated data.

Both the DM and the domain expert collaborate to achieve these purposes. More details are given in the next subsections.

4.3.1 Selection of relevant data

First, it is important to remark that we are working with two data sets from the European Values Study (EVS) surveys: EVS-1980 and EVS-1999. The first one corresponds to the initial point of the simulation and the last one to the validation point of the simulation. The simulation period goes from 1980 to 1999. Thus, the simulation results can be compared with the EVS-1999 survey data.

In the EVS, individuals were chosen by stratified or systematic random sampling to be representative of the Spanish population.[2] The Spanish section of the EVS-1980 has around 2300 individuals, while the EVS-1999 almost 1200. Their high number is suitable for our simulation framework.[3]

In the EVS data sets, there are almost a thousand variables. The number is clearly too high and most of them are irrelevant to the social values trends, specially to analyse the religious and ideological values. Thus, it is required to select only the appropriate ones. In order to do that, there are two options: to use a data analysis

[2] The EVS performed a random sampling for each European country under study, but in this work only the Spanish section is considered.

[3] Note that, as the whole data set is supposed to be representative of the Spanish population, it is not recommended to divide the data sets into training and test sets. However, as the purpose of our analysis is exploratory, it is not necessary to assess our data analysis method with a test set.

Table 1 Variables selected by the domain expert

Name	Description	Source type	Range
Sex	Male (M) or Female (F)	Categorical	–
Age	Age	Numeric	≥18
Age stud.	Age the individuals finished their studies	Numeric	≥5
Marital status	Married, divorced, widowed, etc.	Categorical	–
Status	Socioeconomic status	Numeric	Real
Ideology	Political self-positioning	Ordinal	1–10
Conf. church	Degree of confidence in church as an institution	Ordinal	1–4
Church att.	Frequency of attendance to church services	Ordinal	1–7
Relig. person	Religious (R), Non-Religious (NR) or Atheist (A)	Categorical	–

method that selects the best subset of variables or to rely on the domain expert for the selection. In the first option, if there is a target variable, it could be used as a method to select the subset of variables with the highest ability of prediction. Some of the classification methods, such as decision trees and ridge regression, enjoy embedded feature selection method. If no target variable is available, then choosing the variables with the lowest degree of intercorrelation or mutual information are reasonable choices. For some data analysis methods there are available methods to guide variable selection, see Steinley and Brusco (2008) for the case of cluster analysis.

In our case, due to the availability of domain knowledge, the original set of variables had been previously reduced to a subset of nine variables that was successfully used in the previous experiments with Mentat (Hassan et al. 2008). These nine variables are described in Table 1. Later, it will be shown how some of these variables were ruled out by the interpretation of the clustering results made by the domain expert.

Concerning the variables shown, *conf. church* and *church att.* are scaled from maximum to minimum degree, i.e. the minimum value represents the maximum degree and viceversa; *status* is estimated applying principal component analysis (PCA) for categorical and ordinal data to extract the socioeconomic status from survey variables such as occupation, range of incomes, etc.; in *ideology*, the extreme left is represented by 1 while 10 is extreme right.

The set of nine variables was initially considered for the clustering analysis that is explained in the next section.

4.3.2 Data analysis

The clustering method deemed as suitable for our problem has been a prototype-based clustering: the *k-means* algorithm. This algorithm requires to determine the number of clusters (prototypes), before the clustering is carried out. Thus, a criterion to choose the right number is needed. In order to do so, the *k-means* algorithm was wrapped to make it return a probability distribution for each cluster and a log likelihood value of the result.[4] The log likelihood is a measure of the suitability of the observed data to the normal probability distributions defined by the mean and standard deviation of the k clusters estimated with the *k-means* algorithm. This value allowed us to choose the more appropriate number of clusters k for the observed data.

The right number of clusters was determined looking for the "elbow" in the graph that represents the number of clusters in the x-axis and the resulting log likelihood in the y-axis. The optimal number of clusters is the one at which the log likelihood slope decreases significantly. By doing so, there is a trade-off between the complexity of the model (i.e. the number of clusters) and the likelihood of the results.

Following this approach, several cluster analysis with different variables subsets were carried out. It is important to remark that, in order to avoid the undesirable effect of the different scales of numeric variables, they were normalised.

The best subset of variables was chosen with the help of the domain expert criteria. The domain expert analysed the meaning of the resulting clusters and estimated if they match the sociological theories. In this case, three variables from the initial set shown in Table 1 were discarded:

- if *sex* and *age* are considered, some of the resulting clusters are spurious. E.g. a cluster of widowed old women which is a whole cluster is completely irrelevant from the sociological problem perspective.
- *status* is deemed redundant as the clusters obtained without it are essentially the same. From the sociological point of view, it happens because of the strong correlation among the socio-economic status and the studies level (already represented by the variable *age-stud.*).

4.3.3 Interpretation of results

The interpretation of the clustering results was performed with the aid of the domain expert. In his previous sociological research, the Mentat domain expert obtained a well-defined typology of the religious values (deeply explained in Hassan et al. 2008) based on the values of three variables: *conf. church*, *church att.*, and *is a relig. person*. According to this classification, there are four groups of individuals: ecclesiastical, low-intensity, alternatives and non-religious. In order to shed light on the clustering results, all the individuals will be labelled according to this typology, that will be denoted as RLGTYPE. This new variable will allow us to study our approach from the sociological theory, but it is important to remark that it was not used in the cluster analysis.

In the cluster analysis carried out, apart from the variables used to define the RLGTYPE typology, three more characteristics (*age*, *age-stud*, *ideology*) were used (after the pruning of the previous subsection). Using the procedure described in the previous section, the number of clusters obtained in the three data sets was five, instead of the four categories of RLGTYPE. This is due to the participation of new variables, which produces some differences in the clustering process. However, according to the domain expert, clustering results are consistent with the RLGTYPE typology.

[4] This approach corresponds to the *MakeDensityBasedClusterer* clustering algorithm in Weka (Witten and Frank 2005).

Figures 3 and 4 will help us to interpret the results of the cluster analysis in the three data sets (the two EVS and the simulation results). The figures show the distribution of the individuals of each cluster between the categories of four variables: *age*, *ideology*, *church att.* and RLGTYPE. Individuals from each cluster are represented by a different pattern. The comparison of the distributions for the same variable makes possible to give a sociological and qualitative interpretation to each cluster. The conclusions drawn from the domain expert are the following:

- *Cluster A [white]*: this cluster represents the group of individuals with the following common characteristics: they are rather left-winged and strongly non-religious (mainly atheists and agnostics), typically young people from a middle-high social class (and thus, with high education and economic levels).
- *Cluster B [tiles]*: on the opposite side of the spectrum, these individuals are very right-winged and ecclesiastical religious. They are usually old married/widowed people with a quite low level of income and education.
- *Cluster C [bricks]*: this group matches very well the 'alternatives' type of RLGTYPE, where left-winged and quite young people still consider themselves as religious but with no trust in the Church institution.
- *Cluster D [dots]*: the other non-religious cluster; left-winged as well, but with a quite low education and status because of their greater age.
- *Cluster E [stripes]*: these are religious and right-winged, but with a high level of education and economic status, which make them clearly different from B.

After studying the EVS-1980 and the EVS-1999 results, we can find the same number of clusters and with an equivalent meaning. However, the percentage of each group varies substantially, in both the theoretical typology RLGTYPE or the resulting clusters. The sociological interpretation of this evolution assesses that there has been a transformation of the Spanish society in these 20 years, whose most conservative and religious population (Cluster B) has dropped in favour of the most liberal (Cluster A). At the same time, the groups with intermediate levels of values (Clusters C, D, and E) have experimented slight differences. These results are consistent with the sociological theories concerning the Spanish religiosity map evolution (Hassan et al. 2008).

The main facts of this evolution, shown in Figs. 3 and 4, and taking into account the clusters meaning, are summarised here:

1. The religiosity strength of the population falls. The most orthodox people (B) represent a lower percentage and the liberal ones (A) reach the first position in population. This is due to the individualisation values, characteristic of Europe, which grew quickly in Spain.
2. The ideological spectrum twists to the left, while the educational and economic levels are substantially increased for the whole population, because of the fast economic progress in those years. Thus, the 'poor' clusters, B and D, are affected and fall, while A and E rise.
3. The newest type of religiosity, the 'alternatives' represented by C, rise because of the increase in the youngsters representation.

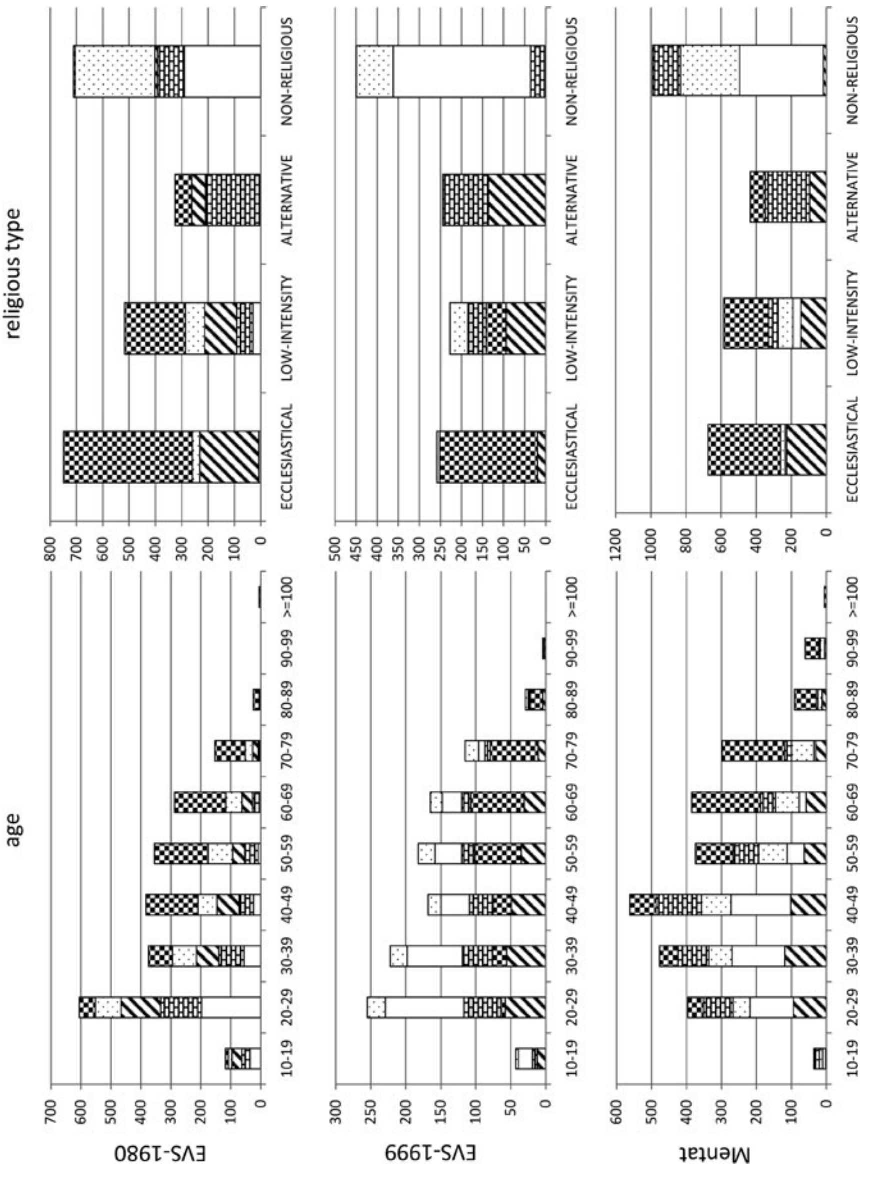


Fig. 3 A selection of the clustering results: comparison of the variables *Age* and *RLGTYPE* in the EVS-1980, EVS-1999 and the Mentat ABM

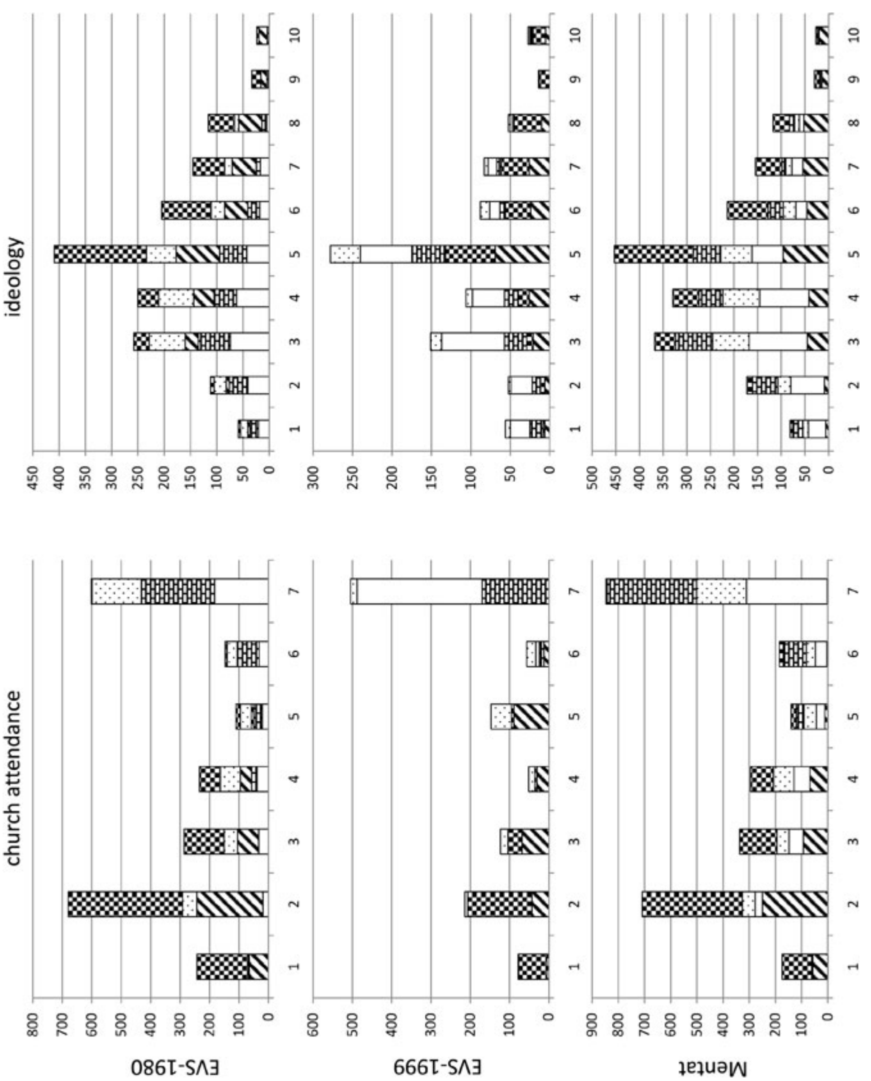


Fig. 4 A selection of the clustering results: comparison of the variables *Church att.* and *Ideology* in the EVS-1980, EVS-1999 and the Mentat ABM

4.4 Validating Mentat through DM

In this case, the validation of the simulation process will consist in comparing the clusters of the results of the agent simulation with the clusters obtained for the EVS-1999. It is important to remark that Mentat was already validated directly against the EVS percentages and with pretty successful results, as extensively exposed in Hassan et al. (2008). However, the validation carried out with the clustering results revealed hidden knowledge that led to new findings.

The number of clusters found and their sociological meaning were the same, which already constitutes a good beginning, as the same patterns can be found. Besides, similar evolving trends of each cluster between 1980 and 1999 are observed in the ABM. Thus, A and C grow, while B and D decrease. Moreover, the three theoretical observations regarding the values evolution mentioned in the *Interpretation of Results* of Sect. 4.3 are nicely observed here as well, as can be observed in Figs. 3 and 4. Although Mentat results show a more leftist ideology, this is normal considering that the Spanish twist to the right in mid-90s (with the change of government) is quite impossible to predict with the basis of 1980.

The ideal situation would be that the distributions in the Mentat clusters match the ones found in EVS-1999. However, this is not the case. Although the sums of the most liberal clusters, A and D, is equivalent in the ABM and in the EVS-1999, each cluster percentage is rather different. This explains why this inconsistence was not found in the previous validation with the EVS: the aggregated statistic and the RLGTYPE (which considers just one irreligious group) were not accurate enough to find the hidden differences. After a deep analysis of the causes of this inconsistence, the problem was detected: the representation differences are directly due to the less quantity of young people in the simulation. Although the difference is not too high and sociologically consistent (in 1999 a drop in the Spanish population had began), it distorts the clusters where young people constitute a big proportion: specifically, the non-religious ones. This error should be solved in future versions of Mentat, reviewing the demographical data inputs that are considered (not only the EVS, but Spanish official data as well) and adjusting it for a better match of the population pyramid.

5 Concluding remarks

This work has presented a methodological approach to agent-based modelling through the use of Data Mining. It was reviewed how agent-based models can be benefited by formal techniques for dealing with the variety and size of data sources, and the rich outputs of complex models. DM provides a powerful collection of resources that the modeller can find useful in several points of the modelling process. In order to provide a framework for the application of DM, an iterative DM-assisted modelling process has been defined, taking into account the roles of both the domain expert and the DM expert throughout the process.

A case study was presented in order to illustrate the different stages of the approach with a rich example. The description of each part of the process has revealed several limitations and potentials of this approach.

The most important limitation for these guidelines is the availability of data. Purely theoretical models that neither are fed by empirical data nor produce large amounts of data will not find any benefit from DM. However, such models are just part of the spectrum of social simulation: in the survey of Heath et al. (2009), they constitute just a 40% of the total. Therefore, there is room for methodologies focused on the descriptive models that consume and produce data.

Related to the data used, another limitation to consider is the availability of, specifically, quantitative data as a requirement for the application of DM. Qualitative data, although potentially useful for the design of the ABM (Yang and Gilbert 2008), cannot be processed by DM techniques. For instance, the Mentat model uses both qualitative and quantitative data sources, but only the quantitative survey was considered. As long as the quantitative data covers all the spectrum of the input of the ABM, the qualitative information used in the design can be safely ignored in the DM treatment.

Another important limitation is the weight of the DM expert role in the process. Due to the wide diversity of techniques available in the realm of DM, the expert's presence is fundamental for its successful selection and application (as it was noted in multiple points of the process in the Mentat model). Depending on the model chosen, a different group of techniques will be needed. It is important to consider that each technique has several implications that can affect the conclusions extracted. Therefore, a fluent communication between the DM expert and the modeller is highly encouraged.

On the other hand, the engagement of the domain expert in the process has revealed essential. Even though most agent-based models restrict the role of the domain expert to the specification of the social phenomena to model, in the methodological approach exposed this role is frequently used along the process. Besides, only if the domain expert fully understands the whole process carried out, will be able to successfully analyse and interpret the meaning of the results, as was the case with the meaning of the clusters in our case study (Sect. 4.3.3).

Thus, the ideal models for the application of this approach are the data-driven ABM, usually complex, large and rather rich systems, such as the urban dynamic model of Galán et al. (2009b) or the well-known model on the collapse of the Anasazi civilization (Dean et al. 2000). As it was stated, the optimal contexts would have large amounts of representative quantitative data that cover the whole spectrum of the studied social problem.

The key utility of the DM process is the potential of revealing new hidden facts in implemented models. At least, this would strength the foundations of the model, and even might detect inconsistencies that would be otherwise overlooked.

Acknowledgements This work has been developed with support of the program *Grupos UCM-Comunidad de Madrid* with grant CCG07-UCM/TIC-2765, and the project *Agent-based Modelling and Simulation of Complex Social Systems (SiCoSSys)*, TIN2008-06464-C03-01, funded by the Spanish Council for Science and Innovation. We also thank the support from the *Programa de Creación y Consolidación de Grupos de Investigación UCM-BSCH, GR58/08*.

References

Ahmad MA, Teredesai A (2006) Modeling spread of ideas in online social networks. In: Proceedings of the fifth Australasian conference on data mining and analystics, vol 61. Australian Computer Society, Inc, Sydney, pp 185–190

Balci O (2004) Quality assessment, verification, and validation of modeling and simulation applications. In: WSC '04: proceedings of the 36th conference on winter simulation, pp 122–129. Winter simulation conference

Boero R, Squazzoni F (2005). Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science. J Artif Soc Soc Simul 8(4):6

Botía JA, Hernansaez JM, Gómez-Skarmeta AF (2004) Towards an approach for debugging mas through the analysis of acl messages. In: Multiagent system technologies, second German conference, MATES 2004. Lecture notes in computer science, vol 3187. Springer, Berlin, pp 301–312

Collier N (2001) Repast: an extensible framework for agent simulation. In: Swarmfest 2000: proceedings of the 4th annual swarm user group meeting, March 11–13, 2000. Utah State University, Logan, Utah, p 17

Dean JS, Gumerman GJ, Epstein JM, Axtell RL, Swedlund AC, Parker MT, McCarroll S (2000) Understanding anasazi culture change through agent-based modeling. In: Dynamics in human and primate societies: agent-based modeling of social and spatial processes. Oxford University Press, London, pp 179–205

Deffuant G, Huet S, Bousset JP, Henriot J, Amon G, Weisbuch G (2002) Agent based simulation of organic farming conversion in allier departement. Complex Ecosyst Manag, pp 158–189

Domingos P, Richardson M (2001) Mining the network value of customers. In: Proceedings of the seventh SIGKDD international conference on knowledge discovery and data mining. ACM, San Francisco, pp 57–66

Edmonds B (2001) The use of models—making MABS actually work. In: Multi-agent-based simulation. Springer, Berlin, pp 15–32

Galán JM, Izquierdo LR, Izquierdo SS, Santos JI, del Olmo R, López-Paredes A, Edmonds B (2009a) Errors and artefacts in agent-based modelling. J Artif Soc Soc Simul 12(1):1

Galán JM, López-Paredes A, del Olmo R (2009b) An agent-based model for domestic water management in valladolid metropolitan area. Water Resour Res 45(5):W05401

Gilbert N (2008) Agent-based models. Thousand Oaks, Sage

Gilbert N, Troitzsch KG (1999) Simulation for the social scientist, 1st edn. Open University Press, Maidenhead

Gostoli U (2008) A cognitively founded model of the social emergence of lexicon. J Artif Soc Soc Simul 11(1):2

Hassan S, Antunes L, Arroyo M (2008) Deepening the demographic mechanisms in a data-driven social simulation of moral values evolution. In: David N, Sichman JS (eds) MABS. Lecture notes in artificial intelligence (from the Lecture notes in computer science), vol 5269. Springer, Estoril, pp 167–182

Hassan S, Antunes L, Pavón J (2009a) A data-driven simulation of social values evolution. In AAMAS 2009 proceedings, Budapest. doi:10.1145/1558109.1558282

Hassan S, Antunes L, Pavón J (2009b) Mentat: a data-driven agent-based simulation of social values evolution. In: MABS 2009 proceedings, Budapest. Springer (Springer LNAI, doi:10.1007/978-3-642-13553-8_12)

Hassan S, Pavón J, Antunes L, Gilbert N (2010) Injecting data into agent-based simulation. In: Takadama, K, Deffuant, G, and Cioffi-Revilla, C (eds) The second world congress on social simulation (tentative), Springer Series on Agent Based Social Systems. Springer, Washington. doi:10.1007/978-4-431-99781-8_13

Heath B, Hill R, Ciarallo F (2009) A survey of agent-based modeling practices (January 1998 to July 2008). J Artif Soc Soc Simul 12(4):9

Janssen MA, Ostrom E (2006) Empirically based, agent-based models. Ecol Soc 11(2):37

Kennedy C, Theodoropoulos G, Sorge V, Ferrari E, Lee P, Skelcher C (2007) Aimss: an architecture for data driven simulations in the social sciences. In: ICCS '07: proceedings of the 7th international conference on computational science, part I. Springer, Berlin, pp 1098–1105

LeBaron B, Arthur WB, Palmer R (1999) Time series properties of an artificial stock market. J Econ Dynam Control 23(9–10):1487–1516

Remondino M, Correndo G (2006) MABS validation through repeated executing and data mining analysis. Int J Simul Syst Sci Technol 7(6):10–21

Richiardi M, Leombruni R, Saam N, Sonnessa M (2006) A common protocol for agent-based social simulation. J Artif Soc Soc Simul 9(1):15

Ripley BD (1987) Stochastic simulation. John Wiley and Sons, New York

Sargent RG (2007) Verification and validation of simulation models. In: WSC '07: proceedings of the 39th conference on winter simulation. IEEE Press, New York, pp 124–137

Sen S, Sekaran M (1996) Multiagent coordination with learning classifier systems. Springer, Berlin, pp 218–233

Steinley D, Brusco M (2008) Selection of variables in cluster analysis: an empirical comparison of eight procedures. Psychometrika 73:125–144

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Elsevier, Amsterdam

Yang L, Gilbert N (2008) Getting away from numbers: using qualitative observation for agent-based modeling. Adv Complex Syst 11(2):175–186

Javier Arroyo has a research position in University Complutense de Madrid. His research interests are Forecasting, Agent-Based Modelling and Symbolic Data Analysis. He holds a PhD in Computer Science by the Universidad Pontificia Comillas de Madrid (Spain).

Samer Hassan is a researcher in Social Simulation with a multidisciplinary background in Computer Science, Artificial Intelligence, Political Science and Sociology. He holds a PhD in Computer Science by the Universidad Complutense de Madrid (Spain).

Celia Gutiérrez is Associate Professor at Universidad Complutense of Madrid. Her research is focused in Data Mining. She holds a PhD in Computer Science.

Juan Pavón is Full Professor at Universidad Complutense of Madrid and the leader of the GRASIA research group, where he has been involved in several research projects on the application of multi-agent systems technology, in particular, on software engineering, distributed control, web services personalisation, knowledge management, and social simulation.