# Re-thinking Modelling: a Call for the Use of Data Mining in Data-driven Social Simulation

Samer Hassan, Celia Gutiérrez, and Javier Arroyo

GRASIA: Grupo de Agentes Software, Ingeniería y Aplicaciones, Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, C/ Prof. José García Santesmases, 28040, Madrid (Spain)
{samer, cegutier, javier.arroyo}@fdi.ucm.es

**Abstract.** Along the last years, the Data-driven approach in Social Simulation is gaining more and more strength. Thus, the use of large collections of empirical data is becoming more frequent, and new demands emerge. This work attempts to contribute to the debate proposing the intense use of Data Mining for the improvement and development of Data-driven Agent-Based Models. Therefore, a methodological approach explaining why and when to use Data Mining is presented, with a formal description of each stage. Besides, a case study following the proposed approach is explained, showing its application step by step.

**Key words:** agent-based modelling, clustering, data-driven, data mining, social simulation

## 1 Introduction

The role Artificial Intelligence techniques have played in Agent-based Modelling has been rather discontinuous. Many agent-based models (ABM) tend to follow a KISS approach ('keep it simple, stupid') proposed by Axelrod [1], and thus elaborated auxiliary tools are not too valued. But even in the cases of complex data-driven ABM, the intense use of an auxiliary AI technique is quite rare.

Some of them simply do not fit well with the agent approach, such as case-based reasoning or natural language processing. However, others, in spite of their obvious utility, were not taken into account till very recently. This is the case of fuzzy logic (with interesting examples such as [2]) and data mining.

Data mining techniques (DM), by nature, depend on the availability of large amounts of data, which is processed and classified and/or clustered, extracting new knowledge such as hidden patterns. Although DM has been widely used in the multi-agent systems (MAS) field (section 2), it took quite long to find DM applications in social simulation. The main reason for this delay is the mentioned KISS approach: till the appearance and growing popularity of complex data-driven models, the ABM rarely used large amounts of empirical data.

Thus, only recently it makes sense to apply this technology, as it is explained in 2. However, this work not only shows a new example, but aims to provide some methodological guidelines for its general application in other data-driven social

simulations (section 3). Only afterwards a case study following the proposed methodology is presented, showing its application step by step. The paper ends with some concluding remarks of section 5.

## 2 Tracking the Common Ground of DM and ABM

According to [3], the applications of data mining in MAS can be divided in:

- Endogenous modeling: where DM is used to provide agents with some kind of intelligent behaviour, for instance [4].
- Exogenous modeling: where DM explores the system's output in order to reveal hidden relevant patterns or even system inconsistencies. Thus, the MAS could be re-designed and improved, or validated and supported by new evidence. A popular possibility is to analyze agent communication, as it tends to provide large amounts of data which can be explored [5].

The roles of DM in data-driven social simulation are similar to those proposed in [3] for the more general case of agent-based simulation, i.e., it can be applied to endogenous or exogenous modelling, as shown above. Unfortunately, nowadays there are few precedents of application of DM to data-driven social simulation.

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 [6], which uses 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 behavioural patterns, a classical context of the technique, which is boosted through the social network pattern-finding need [7]. 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 [8].

However, some works began to tackle the methodological implications of DM in agent-based social simulation (ABSS). In the classical social simulation approach [9], empirical data only play an important role on the validation step, letting the model building be guided mainly by a theoretical abstraction process. However, data-driven simulation encourages the use of data also in the design and initialisation stages, as explained in [10]. 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, data mining 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 [11] 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 data-driven 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 data mining can help to build data-driven simulation models 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 Social Simulation, which, as has been shown, are sparse at present. However, despite this scarcity, there is a lot of room for collaboration between both disciplines. This collaboration is especially indicated in the cases of great amounts of data with potentially useful information hidden within. This section aims to offer a formalised approach to describe how can data mining be applied to enhance the agent-based modelling process.

### 3.1 Shaping the approach

Thus, when data is abundant, the role of DM is to uncover this knowledge relevant for the simulation. Data may arise from two different sources: as a collection of empirical data with valuable information from the real world, or as a result of the simulations in the form of log files. Figure 1 shows a diagram that represents the iterative process of DM-assisted ABM that it is proposed. In the following paragraphs this methodological diagram will be explained.

The diagram describes the ideal situation, when there is plenty of empirical data. In such a context, data from past experiences and current situation are available. Thus, a desirable route to study a problem through data-driven simulation would be: first, to select and analyse empirical data from a past situation; second, load this data into the ABM; third, run the simulation; and finally validate it with a different data set corresponding to the current situation. However, frequently the environment under study does not have such abundance of data, or simply the modeller chooses to use statistical distributions to initialise the agents. In the case of the data-driven approach, those distributions are usually empirically-supported, typically by other research studies. Therefore, those statistics were previously inferred from empirical data (samples or raw data). Thus, the whole diagram stages of data collection and analysis of the initial point can be considered externalised.

Focusing on the diagram stages, the source data could be obtained by multiple 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 by a domain expert, usually the modeller. Besides, a data mining expert (who could also be the modeller) can provide

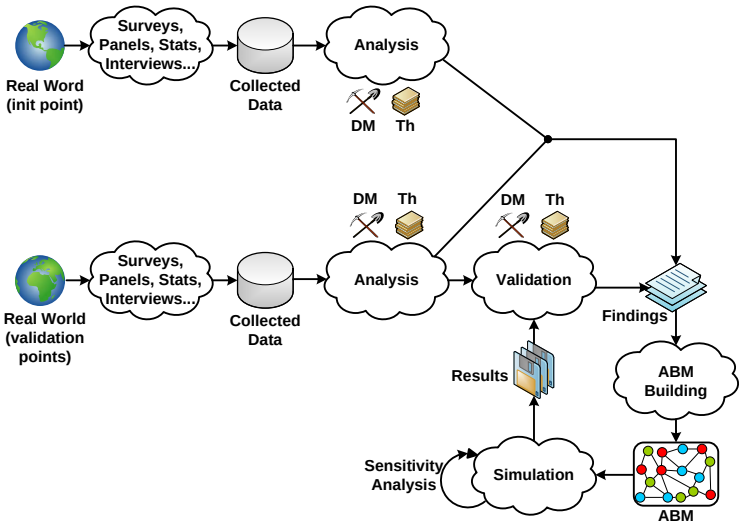


**Fig. 1.** Diagram of the iterative Data Mining-Assisted Agent-Based Modelling

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. The diagram analysis stage will be deeply detailed in the next subsection.

At the end of the analysis, 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. Another desirable task would be to adapt the simulation output to provide a log file that records the state of each agent throughout the simulation period, or at least the final state of the agents. 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 is an accurate representation of the real world state at the considered validation points. This is the *validation* stage, detailed below.

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 previous analysis, 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. As it can be observed in the diagram, the proposed approach is cyclical and it is expected to obtain an increase in the ABM accuracy with each iteration. The main aim of this

methodology is to gain a better understanding of the model and consequently of the situation being simulated with each iteration.

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

### 3.2 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, who may be the same person. As it is shown in works such as [10] and [11], the role of DM in data-driven ABM is becoming more prominent. However, such role must be specified and well-defined. This is the purpose of this section.

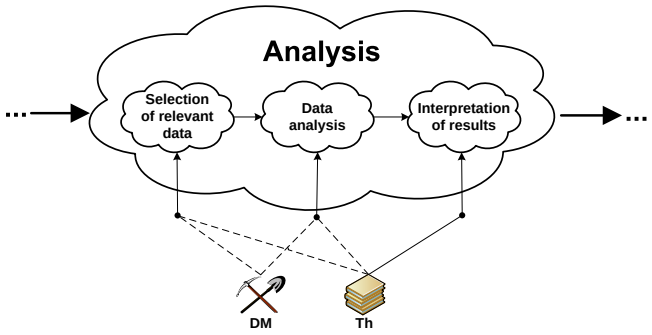


**Fig. 2.** The Analysis stage in iterative Data Mining-Assisted Agent-Based Modelling

Figure 2 shows a diagram that represents an in-depth focus of the *analysis* from Figure 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 (normalization, discretization...). The selection can be initially done by the domain expert, but DM can also be applied. For instance, to analyse if redundant variables have been chosen. Moreover, if the problem 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 the number of variables or individuals of the data is very high, then some kind of DM is essential, because a single human cannot handle such amounts of data. It must be guided by the domain theories, but at the end, they should be endorsed by the DM results. Some examples of the use of DM in the analysis sub-stage follow:

- *Cluster analysis*, useful to discover relevant groups of individuals in the population. The resulting groups could be used to discover hidden patterns which could be used to feed the model. E.g. to define the agents 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 characterize the agents.
- *Time series methods* can 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.
- As mentioned in [11], the DM technique *association rules* may also be applied to discover hidden relationships in categorical data. These relationships would be rather valuable to model the agent behaviour. Also, it could be checked if they are present in the simulated data as well.

Finally, the third sub-stage consists of 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 of the problem along with other domain theories and the results of other experiments.

## 4 A Case Study: the Mentat Model

### 4.1 Context of the ABM

As a case study of the methodology described in the previous section, the Mentat ABM [12] has been used. The aim of Mentat is to understand the evolution of multiple factors in Spain from 1980 to 2000, focusing on social values, specially religious and ideological values. This period is interesting because of the substantial shift in social values corresponding to the transition from the Catholic traditional values to the postmodern materialist values. The almost 40 years of dictatorship finished on 1975, when 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 are analogous to those found in its EU partners. Furthermore, the changes in Spain have been developed with a special speed and intensity during the period studied. The main factor proposed to explain the observed changes is demographic: 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).

Mentat hypothesises 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. In order to do that the ABM is completely data-driven, as explained in subsection 4.2. For

instance, the model is fed with empirical surveys and census for the agents initialisation. Besides, it includes additional data-based probabilities and equations to determine the agents micro-evolution.

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 [13].

The methodology in section 3 has been applied to an already existent ABM. Therefore, instead of the *ABM Building* stage, the model has been modified according to the DM Analysis previously done. The rest of the stages have been carried out normally, as it is explained in the following subsections.

### 4.2 Data Collection in the Model

The Mentat model has been developed following the data-driven approach step by step [13]. Thus, the main source of data is the European Values Study (EVS), 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 [13].

On the other hand, empirical data from the EVS-1999 are used for validating the model simulation. Both data sets are analysed, following the diagram stages, as it is described next.

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 The Analysis Stage

Following the methodology proposed in section 3, the EVS-1980 and EVS-1999 data sets are analysed. The analysis aims are to determine the set of variables

and information relevant to feed the ABM, and to gain some insight about the trends of some parameters and their evolution in the studied period and society. Both the DM and the domain expert collaborate to achieve these purposes.

In the studied case there is no target variable, such as a pre-classification of religious trends, and the number of individuals for each underlying trend is unknown. Thus, an exploratory method is suitable for this problem. This kind of methods will allow us to draw conclusions from the data that will be contrasted against the theoretical models and the domain expert knowledge.

In this case, cluster analysis is applied. It is used to identify groups (clusters) formed according to the social characteristics trends: individuals from the same group share similar characteristics; while individuals from different groups do not show relevant similarities.

The selection of variables plays an important role in the clustering process. Different subsets of variables may lead to different results. Thus, it should be considered the subset of variables whose results provide a good explanation of the studied phenomenon without redundancies or misleading variables.

The insight gained with the cluster analysis can also be used to validate the simulation results of the model. In this case, after the simulation, social groups similar to those found in the EVS-1999 data should appear in the simulated data.

The details of this process are given below.

**Selection of Relevant Data.** In both EVS data sets, there are almost a thousand variables, and the Spanish section of the EVS-1980 has around 2300 individuals, while the EVS-1999 almost 1200. Individuals were chosen by stratified or systematic random sampling to be representative of the Spanish population. Their high number is suitable for our purpose. However, the number of variables is 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 an algorithm that selects the best subset of variables or rely on the domain expert for the selection. The first option includes the use of a method that selects the subset of variables with the highest ability of prediction (if there is any target variable) and with the lowest degree of intercorrelation.

In our case, due to the availability of domain knowledge, the original set of variables was initially reduced to a subset of nine variables that has been successfully used in the previous experiments with Mentat [13]. These nine variables are described in Table 1. Later, it will be shown how some of these variables were ruled out.

Concerning the variables shown, *conf. church* and *church att.* are scaled from maximum to minimum degree, i.e. the mininum 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.

| 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 | – |

**Table 1.** Variables selected by the domain expert

This set of variables will be initially considered for the clustering analysis. However, in the following step it will be determined whether all these variables are relevant for the clustering or not.

**Data Analysis.** The clustering method deemed as suitable for our problem has been a prototype-based clustering, the *k-means* algorithm. As this algorithm requires to determine the number of clusters, a criterion to choose the right number is needed. In order to do so, the *k-means* algorithm is wrapped to make it return a probability distribution for each cluster and a log likelihood value of the result[1]. 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 will allow us to choose the more appropriate number of clusters $k$ for the observed data.

The right number of clusters is 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 is carried out. It is important to remark that, in order to avoid the undesirable effect of the different scales of numeric variables, they have been normalised.

The best subset of variables is chosen according to the expert criteria. The expert analyses the meaning of the resulting clusters and estimates if they match the sociological theories. In this case, three variables from the initial set shown in Table 1 are discarded:

[1] This approach corresponds to the *MakeDensityBasedClusterer* clustering algorithm in Weka [14]

– 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 included with the variable *age-stud.*).

**Interpretation of Results.** Each cluster is analysed in terms of the sociological characteristics that represents. The sociologist, in his sociological quantitative research, obtained a well-defined typology of the religious values (deeply explained in [13]) using a specific selection of variables: *confidence in church*, *church attendance*, and *is a religious person*. According to this classification, there are four groups (ecclesiastical, low-intensity, alternatives and non-religious) labeled by RLGTYPE, that vary in their religiosity level: for instance, in 1980 we can find the group of the ecclesiastical, the most religious, is the most populated.

In the clusters extracted by the DM process, the population is divided in five clusters instead of four, due to the participation of three more characteristics (*age*, *age-stud*, *ideology*), which produces some difference in the clustering process. However, 1980 clustering results are consistent with RLGTYPE. Next, a sociological and qualitative interpretation of each cluster meaning is provided (with its corresponding colour in the Figures 3 and 4):

– *Cluster A [grey]*: 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 [red]*: 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 [light blue]*: 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 [pink]*: the other non-religious cluster; left-winged as well, but with a quite low education and status because of their greater age.
– *Cluster E [dark blue]*: 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 1999 sample with the same selection of variables, we can find the same number of clusters and with the 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 assess 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 [13].

The main facts of this evolution, shown in Figures 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.

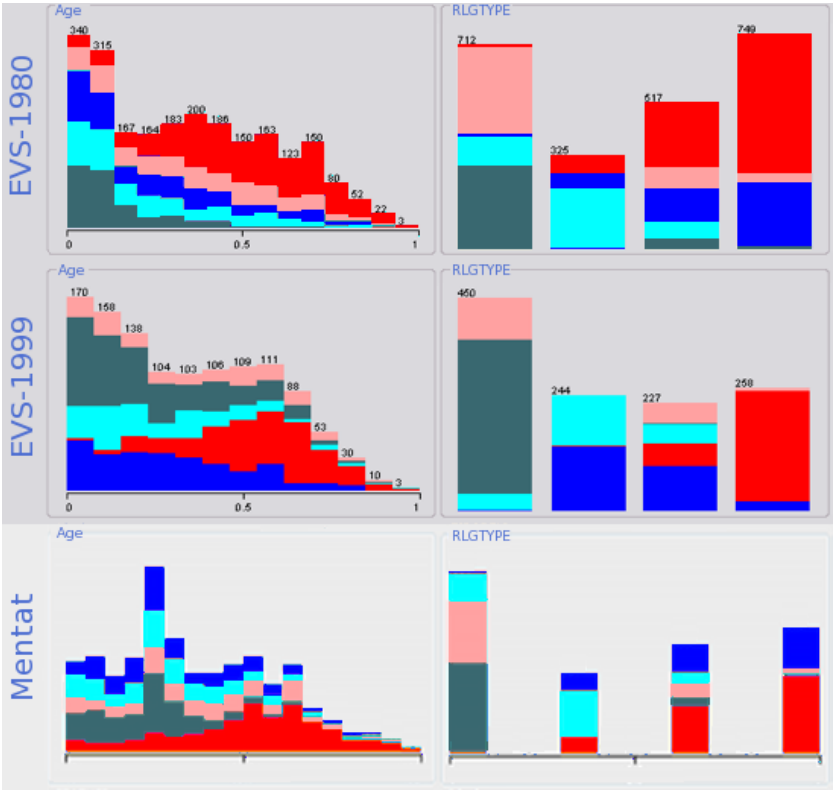


**Fig. 3.** A selection of the clustering results: comparison of the variables *Age* and *RLGTYPE* in the EVS-1980, EVS-1999 and the Mentat ABM.

### 4.4 Validating Mentat through DM

The last analysis stage, the interpretation of results, is useful for the validation of the simulation process. Here, the aim is to find the values patterns that produce the agent simulation and compare them with the classification and conclusions obtained previously.

For a correct comparison with the DM output, Mentat had to be slightly modified in several ways. Although Mentat was already taking into account the same selection of variables, the three religious ones were aggregated in RLGTYPE. For a more accurate comparison and to follow a parallel DM process, this variable was disaggregated. Besides, another output module was implemented to extract an appropriate table filled with the variables for each individual (instead of the classical set of statistics). Therefore, the same clustering process that was realised on the EVS was repeated on this data.

Although Mentat was already validated directly against the EVS percentages (and with very good results, as extensively exposed in [13]), the clustering comparison 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 section 4.3 are nicely observed here as well, as can be observed in Figures 3 and 4. Although Mentat shows 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 Mentat clusters population 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 statistics 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 Spain 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 Data-driven ABM through the use of Data-Mining. The proposed method has been framed and deeply de-

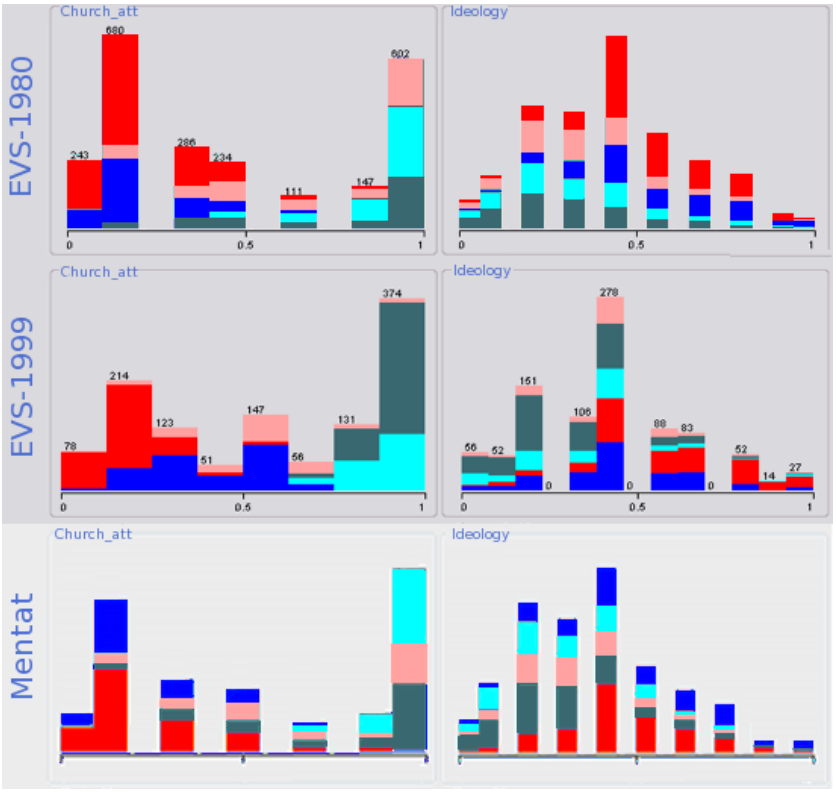


**Fig. 4.** A selection of the clustering results: comparison of the variables *Church attendance* and *Political Ideology* in the EVS-1980, EVS-1999 and the Mentat ABM.

scribed, stage by stage, with a focus on the Analysis process. Besides, a case study was presented, explaining how it was followed each of the stages in the context of a previously-implemented data-driven ABM.

The ideal models for this approach are the strongly data-driven ABM, usually complex, large and rather rich systems. The optimal situations need contexts with huge amounts of representative quantitative data (such as surveys) that cover the whole spectrum of the studied social problem. If the data-driven ABM focuses mostly on qualitative narrative data, DM does not make much sense. On the other hand, small KISS ABM, even if they are partly data-driven, may not require the deep DM analysis. Obviously, in a pure KISS ABM, not data-driven, DM cannot be applied.

The exposed methodology is not only useful to build new ABM, but also to re-think already implemented data-driven ABM. As it has occurred in the case study validation, the DM process may reveal new hidden facts that can, at least, be useful to strengthen the foundations of the model, and even may detect inconsistencies that would be overlooked otherwise.

**Acknowledgments.** We acknowledge support from the project *Agent-based Modelling and Simulation of Complex Social Systems (SiCoSSys)*, supported by Spanish Council for Science and Innovation, with grant TIN2008-06464-C03-01.

## References

1. Axelrod, R.: Advancing the art of simulation in the social sciences. Complexity **3**(2) (1997) 16—-22
2. Epstein, J.M., Mohring, M., Troitzsch, K.G.: Fuzzy-Logical rules in a Multi-Agent system. In: Proceedings of the ESSA'03: 1st Conference of the European Social Simulation Association, Groningen (2003)
3. Remondino, M., G., C.: Data mining applied to agent based simulation. In: Proceedings of the 19th European Conference on Modelling and Simulation, SCS Eur. Pub. (2005) 374–380
4. Sen, S., Sekaran, M. In: Multiagent coordination with learning classifier systems. Springer (1996) 218–233
5. Botía, J.A., Hernansaez, J.M., Gómez-Skarmeta, A.F.: Towards an approach for debugging mas through the analysis of acl messages. In: Multiagent System Technologies, Second German Conference, MATES 2004. Volume 3187 of Lecture Notes in Computer Science., Springer (2004) 301–312
6. Gostoli, U.: A cognitively founded model of the social emergence of lexicon. Journal of Artificial Societies and Social Simulation **11**(1) (2008) 2
7. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, California, ACM (2001) 57–66
8. Ahmad, M.A., Teredesai, A.: Modeling spread of ideas in online social networks. In: Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61, Sydney, Australia, Australian Computer Society, Inc. (2006) 185–190
9. Gilbert, N., Troitzsch, K.G.: Simulation for the Social Scientist. 1 edn. Open University Press (April 1999)
10. Hassan, S., Antunes, L., Pavon, J., Gilbert, N.: Stepping on earth: A roadmap for data-driven Agent-Based modelling. In: Proceedings of the Fifth Conference of the European Social Simulation Association (ESSA08), Brescia, Italy (2008)
11. Kennedy, C., Theodoropoulos, G., Sorge, V., Ferrari, E., Lee, P., Skelcher, C.: 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, Berlin, Heidelberg, Springer-Verlag (2007) 1098–1105
12. Hassan, S., Antunes, L., Pavón, J.: Mentat: A Data-Driven Agent-Based simulation of social values evolution. In: MABS 2009 Proceedings, Budapest, Springer (2009) To Appear in Springer LNAI.
13. Hassan, S., Antunes, L., Arroyo, M.: Deepening the demographic mechanisms in a Data-Driven social simulation of moral values evolution. In David, N., Sichman, J.S., eds.: MABS 2008. Volume 5269 of Lecture Notes in Artificial Intelligence (from the Lecture Notes in Computer Science)., Estoril, Portugal, Springer-Verlag (2008) 167–182
14. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2 edn. Elsevier (November 2005)