Patel *et al. Complex Adapt Syst Model (2018) 6:1*
https://doi.org/10.1186/s40294-018-0052-8

Complex Adaptive Systems Modeling

METHODOLOGY Open Access

A scheme to analyze agent-based social simulations using exploratory data mining techniques

M. Hammad Patel[1], Mujtaba Ahmed Abbasi[1], M. Saeed[1] and Shah Jamal Alam[2*]

*Correspondence:
sj.alam@sse.habib.edu.pk
[2] School of Science & Engineering, Habib University, Karachi 75290, Pakistan
Full list of author information is available at the end of the article

Abstract

Exploring and understanding outputs from agent-based models is challenging due to a relatively higher number of parameters and multidimensionality of the generated data. We use a combination of exploratory and data mining techniques to understand data generated from an existing agent-based model to understand the model's behavior and its sensitivity to initial configurations of its parameters. This is a step in the direction of an ongoing research in the social simulation community to incorporate more sophisticated techniques to better understand how different parameters and internal processes influence outcomes of agent-based models.

Keywords: Agent-based modeling, Exploratory data analysis, Data mining

Background

Agent-based models simulating social reality generate outputs which result from a complex interplay of processes related to agents' rules of interaction and model's parameters. As such agent-based models become more descriptive and driven by evidence, they become a useful tool in simulating and understanding social reality. However, the number of parameters and agents' rules of interaction grows rapidly. Such models often have unvalidated parameters that must be introduced by the modeler in order for the model to be fully functional. Such unvalidated parameters are often informed by the modeler's intuition only and may represent gaps in existing knowledge about the underlying case study. Hence, a rather long list of model parameters is not a limitation but an inherent feature of descriptive, evidence-driven models that simulate social complexity.

Theoretical exploration of a model's behavior with respect to its parameters in particular those that are not constrained by validation is important but have been, until recently, limited by the lack of available computation resources and analysis tools to explore the vast parameter space. An agent-based model of moderate complexity will, when run across different parameters (i.e., the total number of configurations times the number of simulation runs) generates output data that could easily be on a scale of gigabytes and more. With high performance computing (HPC), it has become possible for agent-based modelers to explore their models' (vast) parameter space, and while generating

this simulated 'big data' is becoming (computationally) cheaper, analyzing agent-based model's outputs over a (relatively) large parameter space remains a big challenge for researchers.

In this paper we present a selection of practical exploratory and data mining techniques that might be useful to understand outputs generated from agent-based models. We propose a simple schema and demonstrate its application on an evidence-driven agent-based model of inter-ethnic partnerships (dating and marriages), called 'DITCH'. The model is available on OpenABM[1] and reported by Meyer et al. (2014). In the analysis reported in this paper, we focus on the dynamics and interplay of the key model parameters and their effect on model output(s). We do not consider the model's validation in terms of the case studies on which it is based.

The next section ("Analyzing agent-based models: a brief survey" section) reviews selected papers that have previously addressed the issue of analyzing agent-based models. "A proposed schema combining exploratory, sensitivity analysis and data mining techniques" section present a general schema to analyze outputs generated by agent-based models and gives an overview of the exploratory and data mining techniques that we have used in this paper. In "Illustration: implementing the proposed schema on the 'DITCH' agent-based model" section, we present an overview of the DITCH agent-based model and discuss its parameters with their default values that have been reported by Meyer et al. (2014). This section also describes the experimental setup and results and finally, "Conclusions and outlook" section concludes with next steps in this direction.

Analyzing agent-based models: a brief survey

Agent-based models tend to generate large volumes of simulated data that is dynamic and high-dimensional, making them (sometimes extremely) difficult to analyze. Various exploratory data analysis (EDA) and data mining (DM) techniques have been reported to explore and understand a model's outcome against different input configurations (e.g., Villa-Vialaneix et al. 2014). These techniques include heat-maps, box and whisker plots, sensitivity, classification trees, the K-means clustering algorithm and ranking of model parameters' in influencing the model's outcomes.

Several papers have proposed and explored data mining techniques to analyze agent-based simulations. One such is by Remondino and Correndo (2006) where the authors applied 'parameter tuning by repeated execution', i.e., a technique in which, multiple runs are performed for different parameter values at discrete intervals to find parameters that turn out to be most influential. The authors suggested different data mining techniques such as regression, cluster analysis, analysis of variance (ANOVA), and association rules for this purpose. For illustration, Remondino and Correndo (2006) presented a case study in which a biological phenomenon involving some species of cicadas was analyzed by performing multiple runs of simulations and aggregating the results. In another work, Arroyo et al. (2010) proposed a methodological approach involving a data mining step to validate and improve the results of an agent-based model. They presented a case study in which cluster analysis was applied to validate simulation results of the 'MENTAT' model. Their aim was to study the factors influencing the evolution in

[1] http://www.obenabm.org.

a Spanish society from 1998 to 2000. The clustering results were found to be consistent with the survey data that was used to initially construct the model.

Edmonds et al. (2014) used clustering and classification techniques to explore the parameter space of a voter behavior model. The goal of this study was to understand the social factors influencing voter turnout. The authors used machine learning algorithms such K-means clustering, hierarchical clustering, and decision trees to evaluate data generated from the simulations. Recently, Broeke et al. (2016) used sensitivity analysis as the technique to study the behavior of agent-based models. The authors applied OFAT ('One Factor at a Time'), global, and regression-based sensitivity analysis on an agent-based model in which agents harvest a diffusing renewable source. Each of these methods was used to evaluate the robustness, outcome uncertainty and to understand the emergence of patterns in the model.

The above cited references are by no means exhaustive but provide some interesting examples of the use of data mining techniques in analyzing agent-based models. In the next section, we give an overview of some of the EDA and sensitivity analysis (SA) techniques used in this paper. "Illustration: implementing the proposed schema on the 'DITCH' agent-based model" section of this paper further discusses the EDA, SA and DM techniques vis-à-vis the analysis of simulated outputs of an agent-based model.

A proposed schema combining exploratory, sensitivity analysis and data mining techniques

We propose a schematic approach as a step towards combining different analysis techniques that are typically used in the analysis of agent-based models. We present a methodological approach to use exploratory, statistical and data mining techniques for analyzing the relationships between inputs and output parameters of an agent-based model. Applying the appropriate technique (or a set of techniques) to analyze a model's behavior and parameters sensitivity is the key to validate and predict any real word phenomena in an agent-based model. In "Illustration: implementing the proposed schema on the 'DITCH' agent-based model" section, we demonstrate the application of various exploratory data analysis, sensitivity analysis, and data mining techniques to understand the impact of various input parameters on the model output.

Figure 1 shows a schema that combines exploratory, statistical and data mining techniques to analyze outputs of agent-based models. We first begin with a broader, exploratory analysis of a selected model's input variables (parameters) to understand their effect on the given agent-based model's outputs. This is a typical way of understanding agent-based models, where a wider range of parameters are explored to visually see their relationship with the model outputs. Performing model sensitivity analysis follows next. With many input parameters, understanding outputs through eyeballing is difficult. Hence, techniques such as partial rank correlation coefficient (PRCC) help to measure 'monotonic relationships between model parameters and outputs' (Morino et al. 2008). The use of data mining techniques further allows to find patterns in the generated output across a wider range of a model's input parameters.

Next, we present an overview of some of the techniques that may be applied for each step in the schema, as shown in Fig. 1.

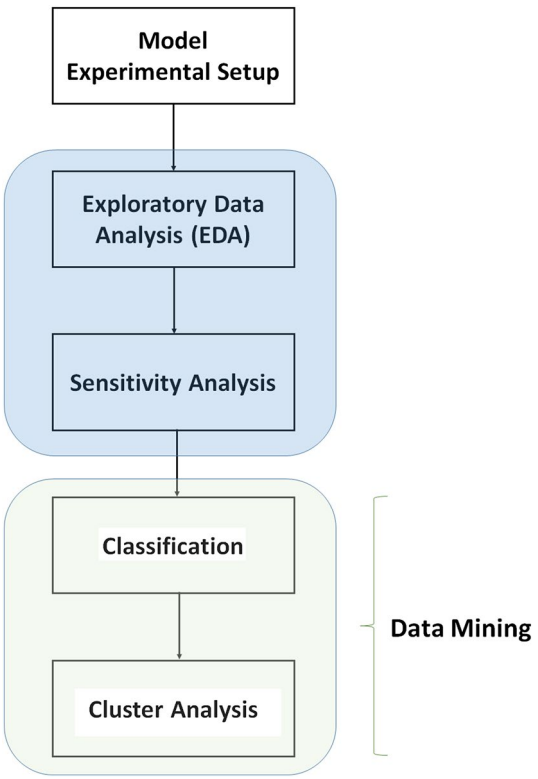

Fig. 1 A schema for analyzing outputs generated by agent-based (social simulation) models using a combination of exploratory, statistical and data mining techniques

Exploratory data analysis

Data analysis in exploratory data analysis (EDA) is typically visual. EDA techniques help in highlighting important characteristics in a given dataset (Tukey 1977). Choosing EDA as a starting point in our proposed schema provides a simple yet effective way to analyze relationship between our model's input and output parameters. Graphical EDA techniques such as box and whisker plots, scatter plots, and heat maps (Seltman 2012) are often reported in the generated data from an (agent-based) simulation. Heat maps are (visually) often good indicators of patterns in the simulated output when parameter values change, whereas, the scatter maps are good often indicators to highlight association between two independent variables (model parameters) for a particular dependent variable (model output). Box and whisker plots on the other hand, summarize a data distribution by showing median, the inter-quartile range, skewness and presence of outliers (if any) in the data. Other techniques such as histograms and violin plots are used to describe the full distribution of an output variable for a given input parameter configuration(s) and are more descriptive than box and whisker plots (Lee et al. 2015).

In this paper, we used the *ggplot2* package in *R* to generate heat maps and box and whisker plots for output variables against the most influential parameters having variations. The result as shown in "Illustration: implementing the proposed schema on the 'DITCH' agent-based model" section highlights the tipping points in heat maps where the percentage of dependent variable changes significantly. In order to explore the

variation in output across the varying parameters, box plots were plotted for different parameter configurations. The results produced while plotting box plots can thus be used to identify the subset of a dataset contributing more in increasing the proportion of a target variable.

Sensitivity analysis

The purpose of performing sensitivity analysis is to study the sensitivity of input parameters of our ABM in generating the output variables, and thus, provide a more focused insight than exploratory analysis techniques. Several techniques may be used to perform sensitivity analysis. For instance, for the results reported in "Illustration: implementing the proposed schema on the 'DITCH' agent-based model" section, we performed multiple sensitivity analysis techniques such variable importance method, recursive elimination method, and PRCC (partial rank correlation coefficient).

Following step 2 of the proposed schema (Fig. 1), we identify two useful methods that are used in the analysis in "Illustration: implementing the proposed schema on the 'DITCH' agent-based model" section: variable importance and recursive feature elimination.

Variable importance

For a given output variable, ranking of each input variable (model parameter) with respect to its importance can be estimated by using model information (training data set). Variable importance thus quantifies the contribution of each input variable (parameter) for a given output variable. The method assumes a linear model, whereby the absolute value of each model parameter is used to train the dataset to generate importance of each input variable. In our case, we used *caret* package in *R*, which constructs a linear model by targeting a dependent attribute against the number of input attributes and then ranking with respect to their estimated importance.

Recursive feature elimination

The recursive feature elimination (aka RFE) method builds many models based on the different subsets of attributes using the *caret* package in *R*. This part of analysis is carried out to explore all possible subsets of the attributes and predicting the accuracy of the different attribute subset sizes giving comparable results.

Using data mining to analyze ABM outputs

There is a growing interest in the social simulation on the application of data mining techniques to analyze multidimensional outputs that are generated from agent-based simulations across a vast parameter space. In this section, we present an overview of some of the common datamining techniques that have been used to analyzed agent-based models' outputs.

Classification and regression trees

A classification/regression tree is based on a supervised learning algorithm which provides visual representation for the classification or regression of a dataset (Russell and Norvig 2009). It provides an effective way to generalize and predict output variables for

a given dataset. In such trees, nodes represent the input attributes, and edges represent their values. One way to construct such a decision tree, is by using a divide-and-conquer approach to reach the desired output by performing a sequence of tests on each attribute node and splitting the node on each of its possible value. The process is repeated recursively, each time selecting a different attribute node to split on until there are no more nodes left to split and a single output value is obtained.

K-Means clustering

K-Means clustering is one of the widely implemented clustering algorithms and have been used to analyze agent-based models, e.g., Edmonds et al. (2014). It is often used in situations where the input variables are quantitative and a squared Euclidean distance is used as a dissimilarity measure to find clusters in a given dataset (Friedman et al. 2009). The accuracy of the K-means clustering algorithm depends upon the number of clusters that are specified at the initialization; depending upon the choice of the initial centers, the clustering results could vary significantly.

Illustration: implementing the proposed schema on the 'DITCH' agent-based model

In this section, we present an overview of the 'DITCH' agent-based model followed by a description of the experimental setup through which the data was generated. We then report analysis of the generated output using the techniques introduced in the previous section.

An overview of the DITCH agent-based model (Meyer et al. 2014)

We have used the DITCH ("Diversity and Inter-ethnic marriage: Trust, Culture and Homopily") agent-based model by Meyer et al. (2014, 2016) for our analysis. Written in NetLogo,[2] the model simulates inter-ethnic partnerships leading to cross-ethnic marriages reported in different cities of the UK and is evidence-driven.

Agents in the DITCH model are characterized by traits that influence their preferences for choosing suitable partner(s) over the course of a simulation run. The model assumes heterosexual partnerships/marriages within and across different ethnicities.

Agents' traits in the DITCH model (source: Meyer et al. 2016):

- Gender {Male, Female}: Agents choose partners of opposite gender.
- Age {18–35}: Preference based on a range with (default) mean 1.3 years and (default) standard deviation of 6.34 years.
- Ethnicity (w, x, y, z): Agents have a preference for selecting partners of their own ethnicity or a different ethnicity.
- Compatibility (score: 0–1): Agents prefer partners with a compatibility score that is closer to their own.
- Education (levels: 0–4): Agents are assigned different levels of education, which influences their partner selection.

[2] http://www.netlogoweb.org/.

Environment Agents in the DITCH model are situated in a social space where they interact with each other and use their pre-existing social connections to search for potential partners. The choice of a potential partner depends upon an agent's aforementioned traits as well as other model parameters which we will discuss later on. Once a partnership is formed, agents then date each other to determine if the partner's education and ethnicity satisfy their requirements. They continue dating for a specified period, after which they reveal their compatibility scores to each other; if the scores are within their preferred range, they become marriage partners. Once a marriage link is formed, agents remain in the network without searching for any more potential partners. There is no divorce or break-up of marriages in the model. The model runs on a monthly scale, i.e., a time step/tick corresponds to 1 month in the model.

DITCH model parameters Following are the model parameters that setup the initial conditions at the start of a simulation run.

- *ethproportion*: Proportions of different ethnicities in the agent population.
- *num-agents*: Total number of agents. The population remains constant during simulation.
- *love-radar (values: 1, 2, 3)*: Defines the search range by an agent for a potential partner in its network as the 'social distance'.
- *new-link-chance*: Probability that two unconnected agents will form a new link during a simulation run.
- *mean-dating/sd-dating*: Mean and standard deviation of an agent's dating period (in years).
- *sd-education-pref*: An agent's tolerance for the difference in education level vis-à-vis its romantic partner.

Experimental setup

Initialization of ethnic proportions The DITCH model uses the UK census data of 2001 as a basis for the parameter *ethproportion*. In all of our simulation experiments reported in this paper, the following four cases were used; based on the four UK cities differentiated with respect to the proportion of different ethnicities (Meyer et al. 2014):

I. *Newham*, London (Super-diverse[3]): White: British (WB): 49.8%; Asian/Asian British: Indian (A/ABI): 17.9%; Asian/Asian British: Bangladeshi (A/ABB): 13.0%; Black/Black British: African (B/BBA): 19.3%.
II. *Birmingham*, W. Midlands (Cosmopolitan): WB: 75.53%; Asian/Asian British: Pakistani (A/ABP): 12.26%; A/ABI: 6.57%; Black/Black British: Caribbean (B/BBC): 5.64%.
III. *Bradford*, W. Yorkshire (Bifurcated): WB: 80.3%; White: Other (WO): 1.6%; A/ABI: 2.8%; A/ABP: 15.3%.
IV. *Dover*, Kent: WB: 98.17%; WO: 1.83%.

[3] The case labels 'super diverse', 'cosmopolitan', 'bifurcated' and 'parochial' are taken from Meyer et al. (2014, 2016); as reported in their original paper.

We conducted experiments using the *BehaviorSpace* tool in NetLogo, which allows exploring a model's parameter space. The approach we used is also called "Parameters Tuning by repeated execution", i.e., varying one input parameter at a time while keeping the remaining parameters unchanged (update-threshold, second-chance-interval) (Remondino and Correndo 2006).

The DITCH model generates several outputs and a complete description is reported by its developers in Meyer et al. (2014; 2016). In the analyses reported in this paper, we have focused on one output variable as the primary output: *crossethnic*, which is the percentage of cross-ethnic marriages in the system. The values taken for this variable were at the end of a simulation run (120 time steps; 10 years) and averaged over 10 replications per parameter configuration.

Given our resource constraints, we performed the experiments in two phases: In the first phase, we looked into the model's sensitivity to scale (in terms of the number of agents) and the extent to which agents search their potential partners in the network (i.e., *love-radar*). In the second phase, we explored the model's parameters specific to expanding agents' social network and those related to agents' compatibility with their potential partners.

Phase-I We first explored the model by varying two parameters with 10 repetitions for a total of 600 runs. All other parameters remained unchanged. Each simulation ran for 120 ticks (10 years).

Phase-II In the second phase, we kept the number of agents fixed to 3000 (see "Conclusions and outlook" section about the discussion on this choice). We then varied the other five model parameters for the four UK cities' ethnic configurations (see Table 1); for a total of 9720 runs. Each simulation ran for 120 ticks (10 years).

Simulation results and analyses

Here we present the results of the simulation experiments. For box plots and heat maps, we used *R*[4] and its *ggplot2* package. For regression/parameters importance analyses and for cluster analyses, we used *R*'s *caret* and *cluster* packages respectively. For classification trees, we used *Weka3* software.[5]

Results from simulation experiments (Phase-I)

In Phase-I, we varied the number of agents and the three values for the model parameter *love-radar*. For the rest of parameters, default values were used as reported in Meyer et al. (2016). The purpose for running experiments in Phase I was to gain a broader sense of the model's outcomes, in particular, the outcome of interest, which is the percentage of cross-ethnic marriages happening over a course of 10 years. Primarily, we were interested in testing the model's sensitivity to scale (the number of agents) and the availability of potential partners once the social distance (*love-radar* parameter) increases (Table 2).

To summarize the results, we generated the box and whisker plots and heat maps (Janert 2010; Seltman 2012; Tukey 1977), to explore variation in output across the two varying parameters and within each parameter configuration when repeated 10 times.

[4] https://cran.r-project.org/.

[5] http://www.cs.waikato.ac.nz/ml/weka/.

Table 1 Model parameters and their range of values that were explored in Phase-I of simulation experiments

Parameter	Values explored	Description
num-agents	1000, 2500, 5000, 7500, 10,000	The number of agents in the model
love-radar	1, 2, 3	The diameter of an agent's ego network through which potential partners are sought

Table 2 Model parameters and their range of values that were explored in Phase-II of simulation experiments

Parameter	Values explored	Description
love-radar	{1, 2, 3}	Represents the social distance with respect to an agent's ego network through which potential partners are sought
new-link-chance	{0.25, 0.5, 0.75}	The probability for an agent to form a new link at each time step (month)
sd-education-pref	{1, 2, 3}	"Standard deviation of the normal distribution governing the agents' preference for difference in education level (mean is always 0)."—Meyer et al.
mean-dating	{1, 2, 3}	"Mean and standard deviation (in years) of the normal distribution governing the duration of the agents' dating period."—Meyer et al.
sd-dating	{1, 2, 3}	

The number of agents was kept fixed at 3000

Figure 2 clearly indicates that the average percentage of cross-ethnic marriages across all the four cases (UK cities) is sensitive to the number of agents in the system. In particular, there is a sharp decrease in the average percentage of cross-ethnic marriages when the number of agents increases from 1000 to 2500, which is more evident in the case of Newham, where ethnic diversity was greatest in contrast to the case of Dover, where 98%

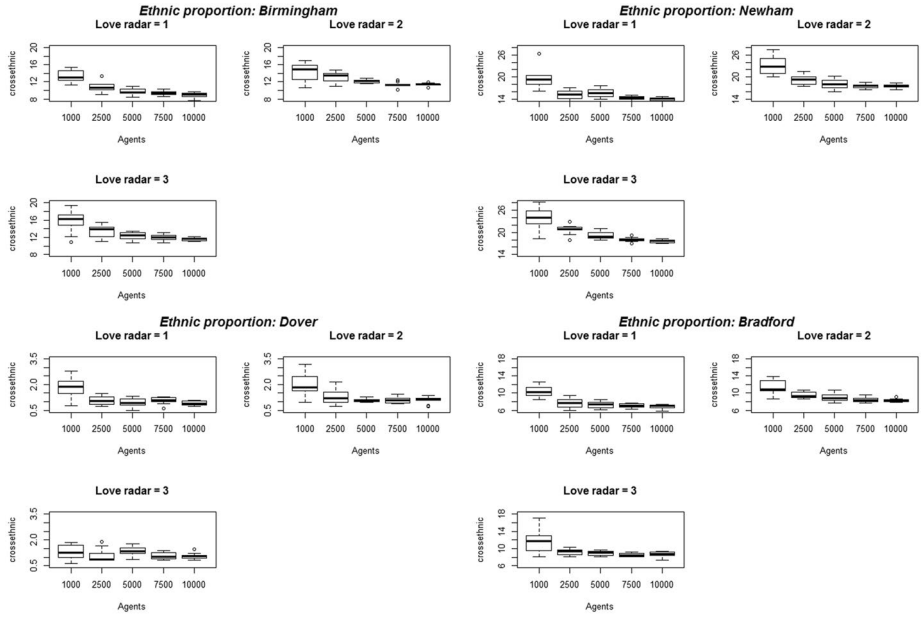

Fig. 2 Box and whisker plots showing the output variable *crossethnic* (percentage of cross-ethnic marriages) for the four cities in the UK across two parameters *num-agents* and *love-radar* in the DITCH model (with box plots drawn across the five different values of *num-agents*)

of the agent population belonged to the White ethnic group. While sensitivity to scale is observed, the observed decline goes much slower and levels off as the number of agents reaches to 10,000.

For a fixed size of agent population, the *love-radar* parameter in the DITCH model does influence the percentage of cross-ethnic marriages for all the four cases (UK cities). This is unsurprising as increasing the value of this parameter enables agents with a wider search space to find potential partners and thus the possibilities for finding a potential partner belonging to a different ethnic group increases as well. However, the relation with increasing the values of love-radar in the model is nonlinear for the output variable *crossethnic* for all the four cases (see Fig. 3). In Newham, which has the greatest ethnic diversity among all the four cities considered, the percentage of cross-ethnic marriages in-creases as the allowable social distance (value of the *love-radar* parameter) increases, whereas in case of Bradford and Dover, an increase in the *love-radar* from 1 to 2 results in an increase in the average cross-ethnic marriages but a further increase from 2 to 3 results other-wise. The heat map plot shown in Figure S1 in Additional file 1: Appendix further highlights this effect.

From an exploratory analysis of simulations from Phase-I, it is clear that the DITCH model is sensitive to the number of agents in the system. As the effect dampens when the agent population increases further on, we fix the number of agents to be 3000 for simulation experiments in Phase-II. In case of the *love-radar*, the observed nonlinear relation indicates that other model parameters that were kept fixed in Phase-I also contribute to the output. Thus, a further exploration and a deeper analysis of the four model parameters is presented next.

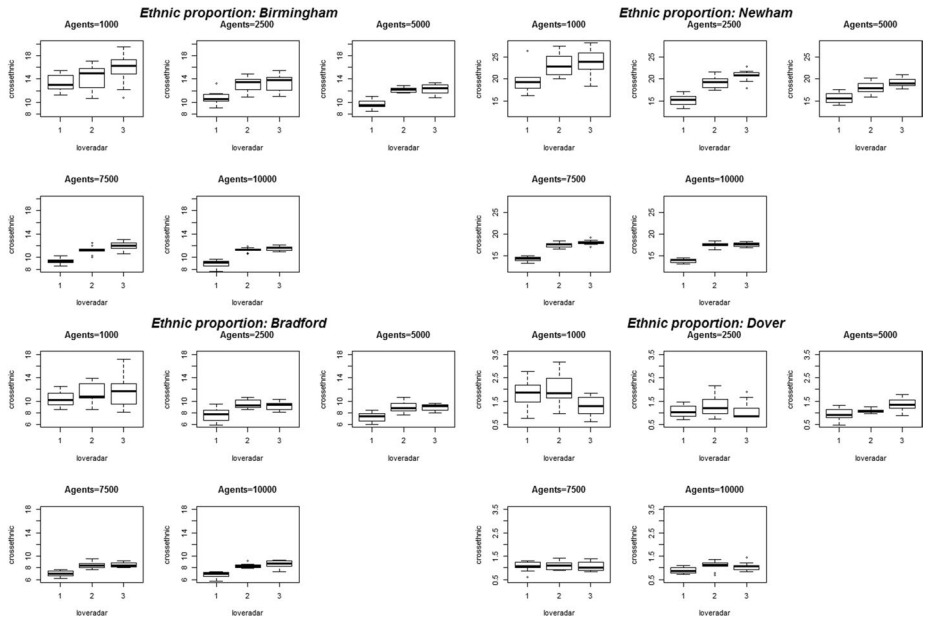

Fig. 3 Box and whisker plots showing the output variable *crossethnic* (percentage of cross-ethnic marriages) for the four cities in the UK across two parameters *num-agents* and *love-radar* in the DITCH model (with box plots drawn across the three different values of *love-radar*)

Results from simulation experiments (Phase-II)

In Phase-II, we fixed the agent population at 3000 and ran simulations across different values of the five other model parameters, as described in the previous section. Here we demonstrate the use of several predictive and data mining techniques that might be useful in exploring and analyzing outputs generated from agent-based models.

First, we estimate the 'importance' of parameters by building a predictive model from the simulated data Brownlee (2016). For instance, importance of parameters can be estimated (subject to the underlying assumptions) using a linear regression model. We used the *caret* package in *R* for this purpose. The method ranks attributes by importance with respect to a dependent variable, here *crossethnic* (the percentage of cross-ethnic marriages) as shown in Fig. 4 (left). As Fig. 4 (left) shows, the model parameters *love-radar* and the *new-link-change* were identified as the most important parameters while the parameter *mean-dating* was ranked last. Figure 4 (right) shows the RMSE (root mean square error) when identifying the predictive model's accuracy in the presence and absence of model parameters through the automated feature selection method. Again, *love-radar* and *new-link-chance* were found as most significant (as the top two independent variables). Having identified *love-radar* and *new-link-chance* as two most important parameters, we explore variation in the generated dataset for the four cases (UK cities) with respect to these two parameters as shown in the box plots in Fig. 5.

As Fig. 5 shows, increasing the value of *love-radar* parameter does result in increasing average cross-ethnic marriages in the DITCH model. Increasing chances of new links formation also contributes albeit less significantly. The variations observed in the box and whisker plots also suggest the role of other three parameters, which seem to play a role when the values of *love-radar* and *new-link-chance* are increased (see heat map in Figure S2 in Additional file 1: Appendix).

Evaluating partial rank correlation coefficients We further explored a subspace of the parameter space to identify the most admissible parameters by evaluating partial rank correlation coefficients (PRCC) for all output variables (Blower and Dowlatabadi 1994). The rationale behind calculating the PRCC is that for a particular output, not all input parameters may contribute equally. Thus, to identify the most relevant parameter(s), PRCC could be useful. One major advantage of identifying the top most relevant parameters based on the PRCC is that given a large parameter space, if only few input parameters have a significant contribution for a particular output, it reduces the dimensionality

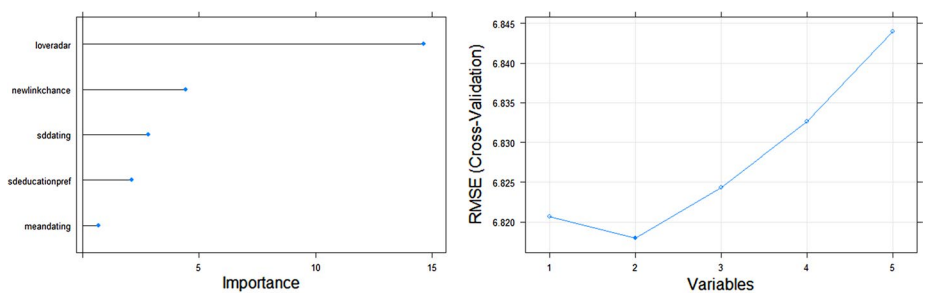

Fig. 4 Ranking of the five parameters as predictors of the output variable *crossethnic*. Left: the *importance* ranking of model parameters. Right: RMSE score against different models built using the automatic feature selection algorithm

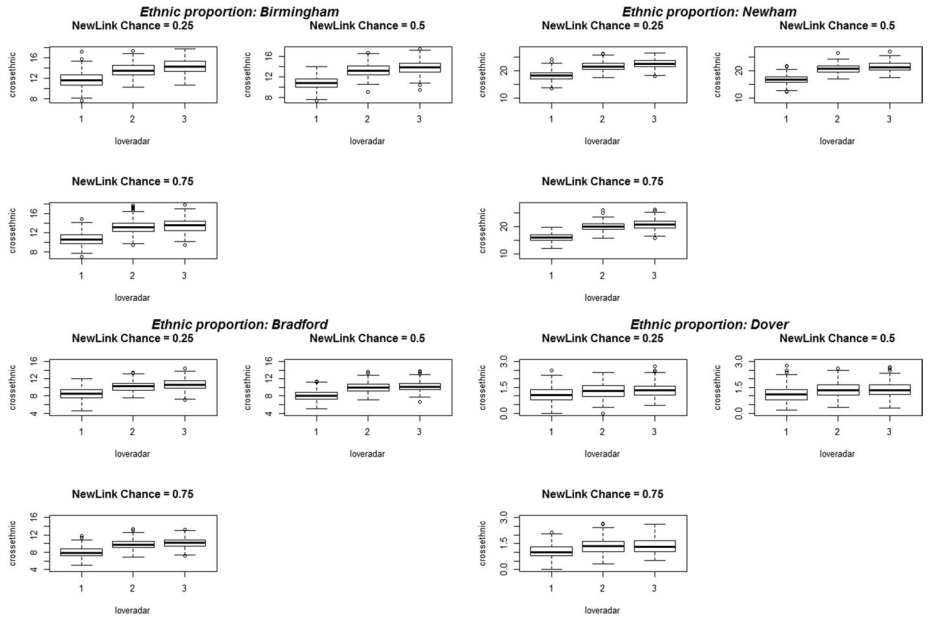

Fig. 5 Box and whisker plots showing the output variable *crossethnic* (percentage of cross-ethnic marriages) for the four cities in the UK across two parameters *love-radar* and *new-link-chance* in the DITCH model (with box plots drawn across the three different values of *love-radar*)

of parameter space significantly. For our analysis, we calculated the PRCCs for all output variables using a package in *R* called *knitr*.[6] Table 3 shows the top three contributing inputs for each output variable when the PRCC was estimated.

Following our proposed schema, we proceed with generating a classification and regressing tree using *Weka*'s decision tree builder as shown in Fig. 6.

The decision tree shown in Fig. 6 was built using *Weka*'s *REPTree* algorithm.[7] It is a 'fast decision tree learner and builds a decision/regression tree using information gain/variance reduction' (Hall et al. 2011). Since here we are predicting the *cross-ethnic* parameter, which is a continuous variable, the *REPTree* algorithm uses variance reduction to select the best node to split. We used the five varied parameters to build the tree shown in Fig. 6, in which the DITCH model parameters *love-radar*, *sd-education-pref*, *mean-dating*, *new-link-chance*, *sd-dating* were the predictors while the output parameter *cross-ethnic* was the target variable. We set the *minNum* (the minimum total weight of the instances in a leaf) property of the classifier to 200 to avoid overfitting. The resulting tree had the following accuracy/error metrics on the test/unseen data.

Mean Absolute Error: 0.9582
Root Mean Squared Error: 1.2995

As the constructed tree shows (Fig. 6), ethnic diversity (or the lack of) in the agent population was the strongest determinant of cross-ethnic marriages. Once again, *love-radar* was found to be the second most important determinant, especially, in situations where some ethnic diversity existed. When the value of the *love-radar* was set to 1 (i.e., only immediate neighbors in the social network were sought), it alone determined

[6] https://cran.r-project.org/web/packages/knitr/index.htmls.

[7] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html.

Table 3 PRCC values of the top three contributing input parameters against each output variable

	Output	Input-variable-1	Input-variable-2	Input-variable-3	Value of input1	Value of input2	Value of input3
1	Cross-ethnic marriages overall	*loveradar*	*newlinkchance*	*sddating*	0.17161477	− 0.04572220	0.03247614
2	Number of agents per ethnicity	*loveradar*	*newlinkchance*	*sddating*	− 0.00556323	0.00274108	0.00262957
3	Married agents per ethnicity	*loveradar*	*newlinkchance*	*sddating*	0.28927404	0.11014048	− 0.47564036
4	Cross-ethnic marriages per ethnicity	*loveradar*	*newlinkchance*	*sddating*	0.26181551	− 0.05297639	0.0114786

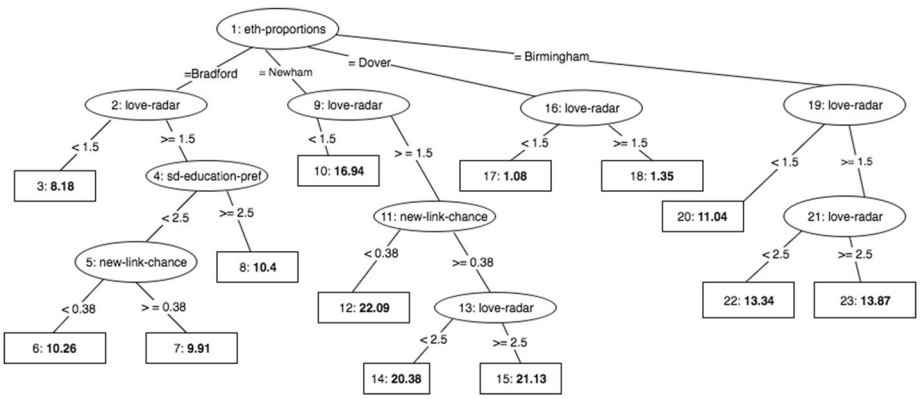

Fig. 6 Classification and regression tree generated from the simulated data across the five parameters in the DITCH model against the average percentage of cross-ethnic marriages shown in the leaf nodes

the percentage of cross-ethnic marriages; however, for higher values of the *love-radar* parameter (i.e., 2 and 3), the output was further influenced by *new-link-chance* and in other instances, the parameters related to agents' dating in the simulation.

K-Means clustering on all 13 DITCH output variables We now turn towards the K-means clustering algorithm to find clusters in the generated dataset. We performed the cluster analysis on the 13 output variables of the DITCH model that were recorded from our simulation experiments. We chose the data from Phase-II, which involved five varied parameters for each sample area (a UK city) with 9720 runs altogether. Our purpose of applying this technique was to group the output instances that were similar in nature into clusters. All output variables were first normalized before proceeding to the next step of finding the optimal number of clusters (k). We then followed the technique used by Edmonds et al. (2014), in which within group sum of squares were calculated against the number of clusters for multiple random initialized runs. The optimal value of clusters in the plot could then be identified as the point at which there is a bend or an elbow like curve. Figure 7 (left) suggests the optimal number of clusters to be around 3 or 4 where a bend is observed.

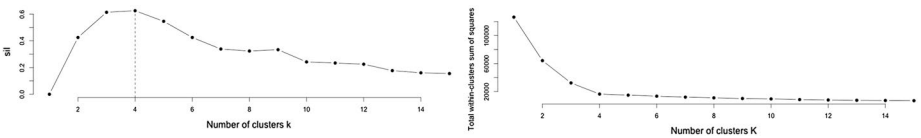

Fig. 7 Finding the optimal number of clusters for the K-means clustering algorithm. Left: using the within group sum of squares technique. Right: using the silhouette analysis

The silhouette analysis[8] shown in Fig. 7 (right) also shows that the optimal value for k is around 3 or 4 in this case. Here the plot displays a measure of similarity between the instances in each cluster and thus provides a way to assess parameters like the number of optimal clusters (Rousseeuw 1987). The results from this analysis confirms that the optimal number of clusters should be around 4. Hence, we ran the K-means clustering algorithm for all the thirteen outputs; the centroids of the four K-means clusters are given in Table 3. The partitioning of data into the four clusters gives a good split across parameters explored.

As Table 3 shows, the goodness of fit is high (~ 87%) indicating that the clusters are distinct, with an almost equal number of instances across all the four clusters. The mean percentage of cross-ethnic marriages was highest in Cluster 2 (19.78%) and lowest in Cluster 4 (1.25%); while Clusters 1 and 3 were found to be closer in terms of the average cross-ethnic marriages. These are results we expect as they present quite an accurate picture of the population distribution of ethnicities in the four UK cities (Newham, Dover, Bradford and Birmingham). We can check the distribution of the input parameter *eth-proportions* across these four clusters and the resulting matrix in Table 4 shows that each region is quite accurately labeled in each cluster. The dominant ethnicity in which most of the cross-ethnic marriages occur like in Cluster 1 representing the sample area Birmingham has *ethnicity-z* (Black/Black British: Caribbean(B/BBC)), which is 5.64% of the total population showing the most cross-ethnic marriages, while in Cluster 2 ethnicity-y(Asian/Asian British: Indian A/ABI) which is 6.57% of the total population, in Cluster 3 *ethnicity-x* ((Asian/Asian British: Indian A/ABI)), which is 17.9%, and in Cluster 4 *ethnicity-x* (White: Other (WO)), which is 1.83% of total populations are representing the highest cross-ethnic marriages.

Figure 8 (top) shows the 2D representation of all the data points of the four clusters. As discussed earlier, Clusters 1 and 3 have some overlapping points while Clusters 2 and 4 were distinct and separate. Finally, Fig. 8 (below) shows the variability in terms of the average cross-ethnic marriages across the four clusters.

Conclusions and outlook

As agent-based models of social phenomenon become more complex, with many model parameters and endogenous processes, exploring and analyzing the generated data gets even more difficult. We need a whole suite of analyses to look into the data that such agent-based models generate, incorporating traditional or dynamic social network

[8] https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/silhouette.html.

Table 4 Centroids of the four K-means clusters for all the thirteen output variables in the DITCH model based on the data generated through simulations in Phase-II (between_SS/total_SS = 87.1%)

Output variables in DITCH	Cluster-1 (2431 instances)	Cluster 2 (2430 instances)	Cluster 3 (2429 instances)	Cluster 4 (2430 instances)
num-turtles-w	2265.52	1494.93	2410	2945.4
num-turtles-x	367.94	536.23	47.77	54.58
num-turtles-y	196.92	389.81	84	0
num-turtles-z	169.6	579	458.19	0
married-turtles-w	1043.43	672	1113.87	1377
married-turtles-x	150.3	225.5	17.29	18.93
married-turtles-y	77.22	160.33	30.58	0
married-turtles-z	65.88	245.82	189	0
out-percent-w (%)	6.63	13.13	5	0.6
out-percent-x (%)	29.19	26.73	57.37	47.54
out-percent-y (%)	38.68	29.8	50.95	0
out-percent-z (%)	42.45	25.28	24.68	0
cross-ethnic (%)	12.75	19.78	9.49	1.25
Within cluster sum of squares	3278.95	3981.93	5186.85	3899.76

analysis, spatio-temporal analysis, machine learning or more recent ones such as deep learning algorithms. There is a growing number of social simulation researchers who are employing different data mining and machine learning techniques to explore agent-based simulations.

The techniques discussed in this paper are by no means exhaustive and the exploration of useful analysis techniques for complex agent-based simulations is an active area of research. Lee et al. (2015), for example, examined multiple approaches in understanding ABM outputs including both statistical and visualization techniques. The authors proposed methods to determine a minimum sample size followed by an exploration model parameters using sensitivity analysis. Finally, the authors discussed focused on the transient dynamics by using spatio-temporal methods to gain insights on how the model evolves over a time period.

In this paper, we propose a simple step-by-step approach to combine three different analysis techniques. For illustration, we selected an existing evidence-driven agent-based model by Meyer et al. (2014, 2016), called the 'DITCH' model. As a starting point, we recommend the use of exploratory data analysis (EDA) techniques for analyzing agent-based models. EDA provide simple, yet an effective set of techniques to analyze relationship between a model's input and output variables. These techniques are useful to spot patterns and trends in a model's output across varying input parameter(s) and to gain insight into the distribution of data that is generated. Sensitivity analysis (SA) techniques follow the exploratory space and are useful, e.g., to rank input parameters in terms of their contribution towards a particular model output. SA techniques are not only useful in identifying those parameters but also quantify the variability of the effect these input

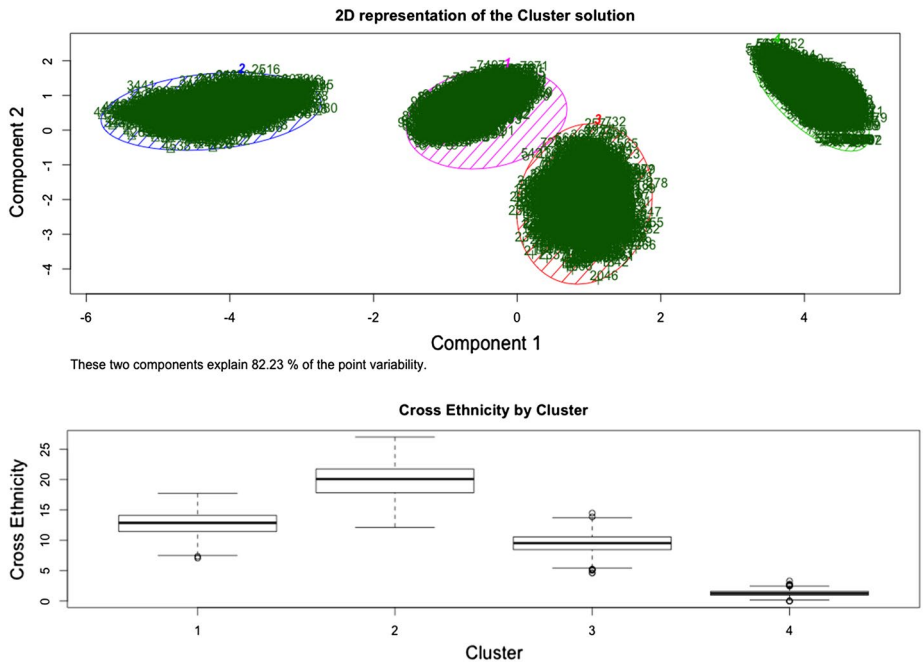

Fig. 8 Top: 2D plot of the four clusters found using the K-means clustering algorithm for the data generated in Phase-II of the simulation experiments. Bottom: box plot showing the distribution of data points against the output variable *crossethnic*

parameters may have on different model output variables. The application of datamining (DM) techniques to analyze agent-based social simulations is relatively new. While traditional techniques such as EDA or SA (or other statistical techniques) are useful, they may fail to fully capture a complex, multidimensional output that may result from agent-based simulations. DM can be useful in providing a better and holistic understanding of the role parameters and processes in generating such output.

Additional file

Additional file 1. Additional figures and table.

Authors' contributions
HP, MA and SJA drafted the manuscript; SJA and MS designed the study; HP and MA generated the data; HP, MA, SJA and MS analyzed and interpreted the data. All authors read and approved the final manuscript.

Author details
[1] Department of Computer Science, University of Karachi, Karachi 75270, Pakistan. [2] School of Science & Engineering, Habib University, Karachi 75290, Pakistan.

Acknowledgements
We are thankful to the anonymous reviewers for their useful feedback and also to the reviewers of the Social Simulation 2017 conference where an earlier version of this paper was presented.

Consent for publication
Not applicable.

Ethical approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Funding
No funding received.

Publisher's Note

Received: 14 October 2017 Accepted: 30 December 2017
Published online: 22 January 2018

References

Arroyo J, Hassan S, Gutiérrez C, Pavón J (2010) Re-thinking simulation: a methodological approach for the application of data mining in agent-based modelling. Comput Math Org Theory 16:416–435
Blower SM, Dowlatabadi H (1994) Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model as an example. Int Stat Rev Revue 62:229–243
Broeke GT, van Voorn G, Ligtenberg A (2016) Which sensitivity analysis method should i use for my agent-based model? J Artif Soc Soc Simul 19:1. http://jasss.soc.surrey.ac.uk/19/1/5.html
Edmonds B, Little C, Lessard-Phillips L, Fieldhouse E (2014) Analysing a complex agent-based model using data-mining techniques. In: Social Simulation Conference 2014. http://ddd.uab.cat/record/125597
Friedman J, Hastie J, Tibshirani R (2009) The elements of statistical learning, 2nd edn. Springer, New York
Hall M, Witten I, Frank E (2011) Data Mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
Janert PK (2010) Data analysis with open source tools: a hands-on guide for programmers and data scientists. Newton, O'Reilly
Brownlee J (2016) Master Machine Learning Algorithms: discover how they work and implement them from scratch. https://machinelearningmastery.com/master-machine-learning-algorithms/. Accessed 1 Dec 2017
Lee JS, Filatova T, Ligmann-Zielinska A, Hassani-Mahmooei B, Stonedahl F, Lorscheid I, Voinov A, Polhill JG, Sun Z, Parker DC (2015) The complexities of agent-based modeling output analysis. J Artif Soc Soc Simul 18:4. http://jasss.soc.surrey.ac.uk/18/4/4.html
Meyer R, Lessard-Phillips L, Vasey H (2014) DITCH: a model of inter-ethnic partnership formation. In: Social simulation conference 2014. http://fawlty.uab.cat/SSC2014/ESSA/socialsimulation2014_037.pdf
Meyer R, Lessard-Phillips L, Vasey H (2016) DITCH—a model of inter-ethnic partner-ship formation (version 2). CoMSES computational model library. https://www.openabm.org/model/4411/version/
Morino S, Hoque IB, Ray CJ, Kirschner DE (2008) A methodology for performing global uncertainty and sensitivity analysis in systems biology. J Theor Biol 254(1):178–196
Remondino M, Correndo G (2006) MABS validation through repeated execution and data mining analysis. Int J Simul 7:6
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
Russell S, Norvig P (2009) Artificial intelligence: a modern approach. Pearson, Upper Saddle River
Seltman HJ (2012) Experimental design and analysis. Carnegie Mellon University, Pittsburgh, p 428
Tukey JW (1977) Exploratory data analysis. Pearson, Upper Saddle River
Villa-Vialaneix N, Sibertin-Blanc C, Roggero P (2014) Statistical exploratory analysis of agent-based simulations in a social context. Case Stud Bus Ind Govern Stat 5:132–149