Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

Analysing a Complex Agent-Based Model Using Data-Mining Techniques

Bruce Edmonds, Claire Little
Centre for Policy Modelling,
Manchester Metropolitan University, UK
Emails: bruce@edmonds.name,
clairelittle7@gmail.com

Laurence Lessard-Phillips, Ed Fieldhouse
Institute for Social Change,
University of Manchester ,UK
Emails: laurence.lessard-phillips@manchester.ac.uk,
ed.fieldhouse@manchester.ac.uk

***Abstract*—** A complex "Data Integration Model" of voter behaviour is described. However it is very complex and hard to analyse. For such a model "thin" samples of the outcomes using classic parameter sweeps are inadequate. In order to get a more holistic picture of its behaviour data-mining techniques are applied to the data generated by many runs of the model, each with randomised parameter values.

I. Introduction

The model that is discussed here is intended as a "data integration model" (Edmonds 2010b). That is a consistent, detailed and dynamic description, in the form of an agent-based simulation, of the available evidence concerning the question of why people bother to vote. This integrates a variety of kinds and qualities of evidence, from source data and statistics to more qualitative evidence in the form of interviews. The model was developed following a "KIDS" rather than a "KISS" methodology, that is, it aims to be more guided by the available evidence rather than simplicity (Edmonds & Moss 2005). A consequence of this is that the model is complicated, including many different, competing and interleaving mechanisms.

It is in the process of being validated. This validation will multi-dimensional, so that both the micro-evidence will be compared against available micro-level evidence as well as outputs compared against macro-level evidence in appropriate ways (Moss & Edmonds 2005). However this validation will necessarily be partial, that is, some aspects of the simulation will be compared against available evidence and some only against opinion. However by documenting as many of the assumptions as possible, these aspects will be amenable to criticism and correction in future versions of the model and hence play a part in the bootstrapping of knowledge (Edmonds 2010a).

In particular this model aims to enable the exploration of some social processes behind voter turnout, including demographic trends in household size and composition, social influence via the social networks the individuals are embedded within, wider social norms such as civic duty, personal habit and identity, as well as individual rationality. This structure was designed to allow the relative priority and interaction of many different context-dependent social processes to be explored.

However a consequence of this approach is that the model itself is too complex to fully understand.

This model is in the process of being abstracted in the form of simpler models. The aim of the more abstract models is that they will be analytically tractable, whilst also giving approximately the same outputs in terms of key outputs, such as the level and trends in voter turnout. In this way it is hoped that the set of models might obtain to both relevance and rigour, albeit in different parts of the 'model chain'.

However the model is very complex which poses a challenge when one tries to check, analyse and validate it. Here, we need to complement low-dimensional parameter sweeps and hypothesis driven experimentation to get a more holistic picture of simulation behaviour. One way of doing this this is described in this paper, that of using data-mining techniques to get insights into model dynamics and outcomes and then use this to direct more specific investigations.

The first section simply describes the model, roughly following the ODD format (Grimm et al. 2006, Polhill 2010). This is followed by the analysis of the model output, using data-mining techniques.

II. The Model

A. Model Layers

The model turned out to have a number of, "layers". Each of which (mostly) only depends on the "lower" layers.

1. *The demographics of households and individuals*. Individuals exist within households at locations within a 2D grid. Individuals are born, age, leave home, partner/split, have children and die based on statistics about these processes derived from the UK population.
2. *Membership of activities*. Individuals change their membership of households, neighbourhoods, schools, work, and activities over time, depending on their age as well as joining rates. Depending on the type of activity, the particular instance chosen to join (which workplace,

school, activity etc.) might be influenced by a number of factors.

3. *Dynamic social networks*. Individuals make friends through membership in these activities. A friend of a friend link creation mechanism also exists but only within each type of link. Links are dropped randomly at a certain rate.
4. *Social Influence over the network*. Messages representing political discussions/messages are sent over the (current) social networks, which are remembered by agents, leading (possibly) to changes in their characteristics.
5. *Voting behaviour*. An individual's characteristics, situation, household situation etc. results in a decision whether to vote, and then vote.

For the purposes of this paper we will only consider states 1, 2, 3, and 4. Thus, for reasons of space we will ignore other aspects of this complex model.

B. Entities, state variables, scales

The model is based around a 2D grid of locations, each of which may be a: household, place of work, school, activity (two kinds) or empty. Households consist of a number of agents which each represent a single person. Agents are born, age, partner, have children and die as the simulation progresses. Agents have a large number of characteristics, but these include: a memory of past events, a party affiliation (or none), a set of family relationships (children, partner, and/or parents) and social connections with other agents. It is over the network of social relationships that influence occurs in the form of events that represent communication about political or civic matters. The agents are influenced over time via these communications. When an election occurs agents decide whether to vote.

Places of work, schools and activities are placeholders. They do not change or move (unlike the households). Their only characteristic is their membership (who works there, which children go to school there, which are members of an activity). A household is simply a container for the agents who form that household.

Agents are the primary elements in the simulation and have the many characteristics, including: age, ethnicity, partner, children, parents, whether employed, immigrant-generation, class, memberships (those schools, places of work or activities that an agent belongs to), social links, and a memory of events (such as recent political conversations they have had).

C. Process overview, scheduling

The simulation is initialised at the start. Then the simulation proceeds in discrete time steps, one step usually representing each month in a year. Each time step the following stages are done.

1. External immigration – households moving into area from outside of the UK sampled from immigrants in BHPS sample, unless grid is full
2. Internal immigration – households moving into area from inside of the UK sampled from all BHPS sample, unless grid is full (remixed in terms of given majority/minority mix)
3. Emigration – households moving out of the area
4. Birth and death – births and deaths probabilistically using statistics
5. Forgetting – stuff being lost from the endorsements of agents at different rates, e.g. remembrance of conversations
6. Network-changes – social links to other agents and activities made and broken
7. Partnerships are formed, move to live together if possible
8. Partnerships dissolved, one partner moves out
9. Household might move within the simulation area
10. Have conversations – hold conversations over the social network, influencing others in the process, the frequency of this is adjusted using the influence-rate parameter
11. Updating agents' attributes in terms of: noticing politics, interest level, and civic duty
12. [Once a year] update: the party preference, party habit and generalised habit
13. Drift-process – shift of voters into and from each political party by a drift process: voters for ruling party (not very interested in politics) drift away to grey, some grey drift to a party
14. During an election tick agents decide whether to vote in a multi-stage process.
15. Updating various plots and statistics for output about what is happening in the simulation

For each of these stages agents are fired in a random order (newly random each time and process). In most of these processes the update for each agent has no immediate effect on any other agent, so these agent processes are effectively in parallel. Similarly most of these stages could be done in any order with very little impact on the outcome, the exception being the sub-stages of voting (14).

III. DESIGN

Basic principles. The starting point for the model design was a collection of "causal stories" about behaviour that might be relevant. Each such story traces a single causal thread through the complexity of social and cognitive processes whilst letting the context of these be implicit and whilst ignoring their possible myriad interactions. This "menu" of behaviours drove the architecture of the model as it was designed to allow most of these stories to be expressed by agents. When the simulation is run the local conditions of each agent separately define the context of that agent whilst also allowing the complex mixing of many different social and cognitive processes.

To fill in some of the cognitive and contextual "glue" evidence from many different sources has been included to motivate the assumptions and mechanisms of the model. Thus it is difficult to identify discrete "submodels" in this.

However, a post-hoc analysis of the structure that emerged suggests the following could be considered as submodels:

1. The main social unit is the household, a collection of individuals living within the same house. People who partner may form a new household, or people moving from outside the area may also do so. Many social processes occur within the household and others occur preferentially between members of the same household. Households occupy a place within the 2D grid.
2. Basic demographic processes specify how people enter the model (though moving into it from the UK or abroad), are born to partners, age, leave home, partner, separate, and die. These are based on some available statistics as to the probability of these events, depending upon whether the immediate situation of the agent makes these plausible. This demographic model includes a 5-category social class model using statistics to determine class mobility.
3. To this basic demographic is added a number of activities. These are schools, places of work, activity type1 and activity type2 (corresponding to things like: places of worship, sports clubs etc.). Agents between the age of 4 and 18 attend school; those 18-65 can go to a place of work, and join (or leave) activities. The activities take up a location but they have no characteristics except their current membership. All children are member of the nearest school; if in work adults are members of a random place of work; with a certain probability adults join an activity and, if so join the one whose other members are (on average) most similar to themselves.
4. A dynamic social network develops between agents. Each link represents a relationship that would allow for a conversation about politics and civic duty. The links are typed – the types are: partner, household, neighbourhood, work, school, activity1 and activity2. There are several ways that a new link can form: all people in the same household are linked with a household link, there is a chance that people in neighbouring households might link, people who go to the same school or parents of children who go to the same school might link; people who are members of the same activity might form a link. Further for each of these link types there is a chance of making a link with a "friend of a friend". Links can be dropped under certain circumstances and with certain situations.
5. Agents can have different levels of political interest, a party political leaning, a sense of civic duty to vote, a generalised habit to vote, a party identification, and a memory of whether past voting/not brought about their desired outcome.
6. A process of social influence occurs over this social network in the form of discrete (as opposed to continuous) political discussions. A political discussion occurs if: (a) there is a link between the two (b) the talker is at least interested in politics has at least a view on politics and (c) the receiver at least notices political discussions.
7. These political discussions have several effects on agents that are not described here.
8. When an election occurs, each individual decides whether to vote.
9. Voting statistics are then recorded.

The above are not the full details but a summary of their main features. Generally micro-causation in the model happens down the order above (from first to later), but there are some weaker and slower feedbacks that occur back upwards, for example the outcome of an election effects agents' perceptions of the experience of voting.

Emergence. Clearly in such a complicated model it is not possible to make an easy and clean distinction between results that emerge and those that are programmed into the model. Indeed, the model was designed with a view to integrate available evidence rather than produce or demonstrate emergent effects (or to be predictable). However it is not the case that all outcomes from the model are straightforwardly forced by the settings and programmed micro-processes.

The initialisation of the model (see below) has a complicated but predictable effect on the model, in that the kinds of household the model is seeded with will affect the tendencies that follow. Thus in the data set that these are selected (at random) from those from "invisible minorities" (Irish etc.) tend to be more politically involved and have a higher sense of civic duty.

The impact of many of the parameters is straightforward, for example: increasing the probability of holding a conversation increases the general level of political interest and hence the turnout; increasing the forgetting rate (the "forget-mult" parameter) means that people do not recall so many positive political messages and hence the level of interest in politics falls quicker. The immediate effect of mobilisation is fairly straightforward – the more people are mobilised the more vote – but how this effects the longer term is less obvious in that it seems to have greatest impact upon the levels of civic duty and general habit, than (for example) in terms of a cascade effect in brining yet others out to vote.

Adaptation. Agents generally do not seek to increase or optimise any measure of success nor do they reproduce behaviours that they perceive as successful. The exceptions are: (a) when agents weigh up their past experiences of voting as one factor in the decision of whether to vote again, (b) when moving to a new location within the model, the choice might be influenced in the sense of seeking a location with neighbours similar to themselves and (c) if choosing to join a type of activity agents will choose the instance of the activity whose membership is, on average, the most similar to themselves.

Learning. Agents do learn, adapting their traits depending on circumstances and history, including adapting their *social network*. Agents develop their social network in a number of ways over time: (a) they are automatically linked to other members of the same household, (b) they connect with a probability to those at the same school (or other parents with children at the same school), activity,

workplace or immediate neighbours (but preferentially to those more similar to themselves) and (c) for each kind of link (neighbourhood, school, activity1, activity2, workplace) they can make a link to some of those linked to those they are linked to (so called "friend of a friend". There is a fixed probability of dropping links at each time click, also if an agent moves they are almost certain to lose existing school, neighbourhood and household links (there is a small probability of retaining them).

Prediction. Agents do not do any prediction in this model. In particular, there is no tactical voting, nor expectations about whether it is worth voting based on predicted outcome.

Sensing. This is a social model, so that agents primarily sense other agents in three ways: (a) through their current links to other agents, (b) through indirect links to other agents, e.g. by being members of the same activity, having kids at the same school or being in neighbouring cells (c) through political discussions over the direct links. Thus all sensing is local in the sense of their links, memberships or neighbourhood (except that agents are aware of the result of elections).

Interaction. Agents interact with each other by having political "conversations", which may influence the recipient. Each "conversation" carries messages of political leaning and civic duty (depending on the characteristics of the converser). These are not strictly conversations since each one is one way, but over time these may go both ways between agents, reinforcing existing characteristics of leaning, political interest and sense of civic duty. If an agent moves location, it will bring its partner and children with it (as well as possibly orphaned children in the household). Agents form sexual partnerships, selecting from those in their social network, and can only have children when within such a partnership. Partnerships dissolve with a low random probability in which case one partner will move out leaving any children behind.

Stochasticity. Many processes in the model have a stochastic element in them once the conditions for their occurrence are locally met in an agent. This includes the processes of: moving location, emigrating, immigrating, getting a job, losing a job, making new social links or losing them, joining an activity or leaving one, having a political conversation, acquiring civic duty as a result of a conversation, dragging others to go and vote if they are, and mobilising voters. Other process have a probability of occurring but with the probability varying on the basis of some statistics, including: birth, death, moving out of the parental home, becoming ill, and children changing class later in life form that they were born with (which also depends on having a post-18 education).

The processes that determine the probability of someone voting are deterministic but somewhat complicated (see 8 in the section on design principles and 14 under the section on scheduling) . Many circumstances, such as having a sense of civic duty or being politically involved force a probability of voting at 1 (unless a confounding factor intervenes).

Processes that are entirely deterministic include: going to school or leaving it, retiring from work, the election result, changes in the habit of voting, or political identification.

A major stochastic impact on the model is in the initialisation of the households at the start of the simulation and the choice of new households that enter during the simulation due to immigration. In these processes entire households are selected at random from re-mixed sample of households form the 1992 wave of the BHPS. The "re-mixing" is done to achieve the user defined proportion of majority population as well as to ensure that out-of-UK immigration is selected form those recorded as immigrants in the BHPS sample. Thus the mix of initial households in each run of the simulation will be somewhat different, but on the whole, the balance of household characteristics will be representative for simulations with larger populations albeit with some stochastic variation.

Collectives. Some of the agent characteristics do influence how the agents make links and move. Which locations a household moves to is influenced by a bias towards moving next to households with similar characteristics; which instance of a kind of activity 1/2 are joined will be those whose existing members have (on average) the most similar characteristics as themselves; which person they make links with via an activity will be biased by a similar homophily formula. Thus over time agents will tend to have more links with those similar to themselves. However due to the presence of much stochasticity in the model this does not produce pronounced segregation, but rather a "softer" bias in terms of social links. The characteristics that are involved are: age, ethnicity, class and political leaning. At the moment there is a single dissimilarity measure used between two agents regardless of the context (in future versions this will be changed so that there are different measure for different circumstances, so (for example) a weaker one at work than for choosing which instance of an activity to join).

Political parties are not currently represented, except implicitly in terms of the mobilisation process. Individuals influence each other individually and not collectively in this model.

Observation. Many different statistics are collected from the simulation. Broadly the more complex a simulation, the more different aspects need to be validated in order to have any confidence that the model represents what one intends it to. Following the process of cross-validation (Moss & Edmonds 2005) broad evidence and statistics are used to inform the specification micro-level agent rules but then the results coming out of the model also checked, both statistically and in broader qualitative terms. Thus many graphs and histograms are provided, giving different "views" into what is happening in the complex simulation.

The simulation also monitors many statistics, including: the year, month, size of electorate, population size, number

of first generation immigrants, number of second generation immigrants, number of visible minority and invisible minorities, number of patches that are empty, average proportion of household links in which agents voted for the same party, average proportion of friendship links in which agents voted for the same party, average proportion of household links in which agents either voted or did not the same, average proportion of friendship links in which agents either voted or did not the same, the link density (proportion of all possible links that exist), and the average local clustering (proportion of linked to agents that are linked to each other).

In addition there is a trace, where the events that occur to a randomly chosen agent are logged. When this agent dies a new born agent is chosen and logged. This is to give a feel for the sort of life trajectories agents are going through.

Many statistics are (optionally) recorded in a ".csv" file for subsequent analysis, including:

- run-id: a unique integer assigned to the run of the simulations
- year: the year the simulation tick is in
- month: the month the simulation tick represents
- pop-size: the number of agents in the simulation
- electorate: the number of potential voters, i.e. those 18 and over
- av-age: average age of population
- num-voting: number who actually voted in last election
- turnout: proportion of electorate voting, i.e. num-voting / electorate
- av-adfriends: mean number of friends (adults only)
- sd-adfriends: standard-deviation of number of friends (adults)
- prop-maj: proportion of the population who are of the majority ethnicity
- prop-adult: proportion of the population who are of age 18 years and over
- prop-1stgen: proportion of the population who are 1st generation immigrants
- av-clust: average local clustering (the proportion of friends that are friends with each other) of adults
- link-dens: proportion of links from all possible links
- av-fr-samevote: average number of friendship links whose agents had voted for the same colour (grey if not)
- av-fr-whvoted: average number of friendship links whose agents had voted/not
- av-hh-samevote: average number of household links whose agents (at each end of link) had voted for same colour (grey if not)
- av-hh-whvoted: : average number of household links whose agents (at each end of link) had voted or not in the same way
- av-sim-hh: average similarity of individuals in a household
- av-sim-fr: average similarity of those linked
- ncvs-ac: number of conversations within 'activity' related links per month (rate-ncvs-ac is scaled by av population size)
- ncvs-sc: number of conversations within 'school' related links,per month (rate-ncvs-ac is this scaled by av population size)
- num-adult-involved: number of agents with "involved" level of political interest (prop-adults-involved is this scaled by av population size)

IV. Details

Initialization. The grid is initialised in the following manner:

1. The grid dimensions are set by the programmer
2. Set proportions of the grid are occupied with schools, work places, activity1 and activity2
3. A given proportion of patches that are left are populated with new households. These are selected as a complete household from a large sample of taken at random from the 1992 wave of the BHPS, but "remixed" to a set degree of majority population (by splitting the original file into majority/non-majority households and then probabilistically choosing at random from each part according to parameter settings). Some details about households have to be inferred from the data as this is not always unambiguous. Some initial agent characteristics are set using proxies from the data, e.g. civic duty is set for agents who are recorded as being a member of certain kinds of organisation
4. Links to household members and some random neighbours are made
5. To give the households an initial network the procedure to develop other network links is done 10 times for each household.
6. Appropriate activities are joined depending on those in the BHPS data.

Thus the exact composition of the grid varies in each run but are drawn from the same sample, so in a sufficiently large initial set of households (determined by the size of the grid and how much is left empty) one gets a similar mixture each time. Various other things are initialised including: shapes and colours for main display, election dates, and party labels.

Input Data. There are two sets of data that are used in the model:

1. A sample of the 1992 wave of the BHPS data as described above.
2. Various statistics concerning the underlying demographics, such as birth rate (depending on the age of parent), death probability (each age), probability of males and females leaving home. At the moment these are statistics from only roughly the appropriate time.

A. Submodels

The model has the following "foreground" parameters, including notably:

- birth-mult: a scaling parameter that changes the birth rates uniformly
- death-mult: a scaling parameter that changes the death rates uniformly
- move-prob-mult: a scaling parameter that changes the probability of moving
- drop-friend-prob: the probability a link is dropped in a year
- drop-activity-prob: the probability an activity membership is dropped each year
- influence-rate: a scaling parameter determining the maximum number of chances to influence others each agent has each year
- prob-partner: the probability of forming a sexual partnership if single per year
- density: the initial density of households in the 2D grid
- majority-prop: the proportion of the initial population from the majority group
- immigration-rate: percentage of population that immigrates from outside the UK into the model (and hence is randomly selected from the immigrants section of the BHPS file)
- int-immigration-rate: percentage of population that immigrates from inside the UK into the model (and hence is randomly selected from the re-mixed version of the BHPS file)
- emigration-rate: the rate (per year) that households leave the model.

V. Model Analysis

The approach adopted here is to do many runs of the model (in this case 3862 independent runs) with some of the parameters for each run set randomly. Thus, many different combinations of possible parameter values were tried. The idea is to sample a sufficient 'block' of possible parameter values in several dimensions (in this case 9 dimensions). Clearly the more parameters one varies (and hence the higher the dimension is the space of possibilities sampled) the broader a 'view' of the data one obtains. However one needs a sufficient 'density' of points, so the more dimensions the more runs need to be done.

The parameters and the uniform distributions used to select their values were:

- density: [0.65, 0.95]
- drop-activity-prob: [0.05, 0.15]
- drop-friend-prob: [0, 0.01]
- emmigration-rate: [0 ,0.03]
- immigration-rate: [0, 0.02]
- int-immigration-rate: [0, 0.02]
- majority-prop: [0.55, 1]
- prob-move-near: [0, 1]
- prob-partner: [0.01, 0.03]

The outputs of the model were a set of values measured at the end of the run (at end of year 100), including: pop-size, electorate, av-age, sd-age, av-hsize, sd-hsize, av-adfriends, sd-adfriends, prop-maj, prop-inv-min, prop-vis-min, prop-adult, prop-1stgen, prop-2ndgen, prop-nonempty-n, prop-sim-n, prop-sim-fr, link-dens, av-clust, av-sim-hh, av-sim-fr, ncvs-pt, ncvs-hh, ncvs-wm, ncvs-ac, ncvs-ne, ncvs-sc, num-adult-involved, num-adult-interested, num-adult-view-taking, num-adult-noticing, num-adult-not-noticing, num-with-0-friends, num-with-1-5-friends, num-with-6-10-friends, num-with-11+friends, num-short-campaign-messages, num-long-campaign-messages.

A. Clustering in lower dimensions

Many of the attributes are highly correlated, so here we concentrate on only 13 attributes:

pop.size, av.age, av.adfriends, prop.maj, prop.adult, prop.1stgen, link.dens, av.clust, av.sim.hh, av.sim.fr, ncvs.ac, ncvs.sc, num.adult.involved

For an initial exploration of the data, agglomerative hierarchical clustering was performed – this is a method whereby each record/simulation starts in its own cluster, and the algorithm iteratively joins the two nearest clusters until we reach the point where only one cluster remains. Euclidean distance was used to measure distance between pairs of simulations, and Ward's linkage criterion was used to calculate the dissimilarity between clusters.

Figure 1 displays a dendrogram of the clustering, which shows how the data was merged. In order to choose the number of clusters, the dendrogram is generally cut at the smallest height that has a large increase in within cluster variance – here three clusters are formed (as shown) by cutting at a height of around 1000.

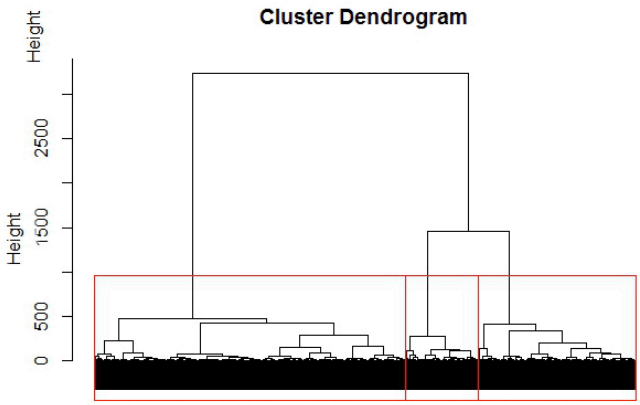

Figure 1. dendrogram of hierarchical clustering of simulations

Hierarchical clustering may be used to cluster the columns (attributes) as well as the rows (simulations) of data, and can be depicted using a heatmap as shown in Figure 2. A heatmap represents numbers as colours. We can see the

three clusters from Figure 1 representing the rows (but sideways this time), together with a clustering of the columns (the thirteen output attributes). The output attributes themselves appear to fit neatly into two clusters. This type of visualization can provide a good initial view of any patterns within the dataset.

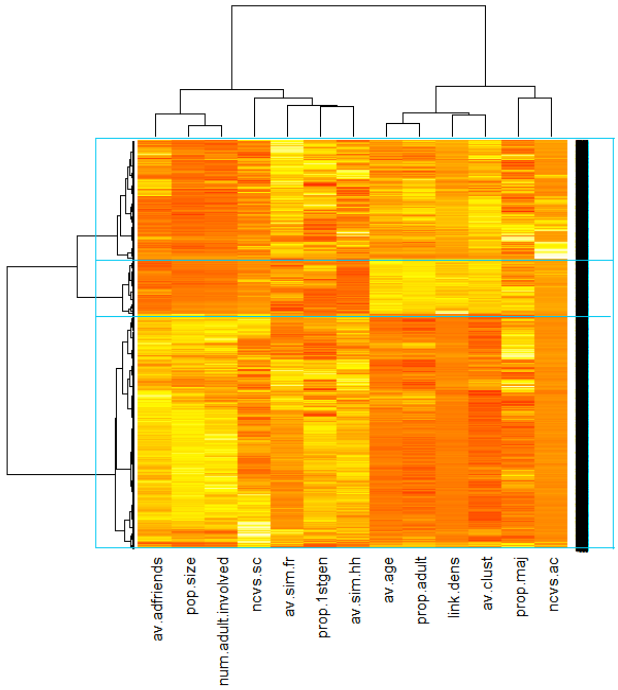

Figure 2. A heatmap of the hierarchical clustering

K-means clustering (Macqueen 1967) is a method of partitioning a data set into a number (k) of different clusters – it is a form of unsupervised machine learning, in that there are no pre-defined class labels for the data. The aim is simply to group together data items in such a way that items within a cluster are more similar to each other than to those in other clusters.

The k-means algorithm requires the number of clusters to be known beforehand – this may be determined either through expert knowledge and/or analysis of the data. K records are randomly chosen to represent the cluster centres, and every record within the data set is assigned to its nearest cluster centre (using a Euclidean distance measure for numerical data). Once all records are assigned to a cluster, the centre of the cluster is recalculated by taking the mean of all records contained in it. Any record that is now nearer to another cluster is reassigned, and the cluster centres are recalculated again. This process continues until no records change clusters (or we reach some pre-defined stopping criterion).

K-means has its drawbacks – it can be difficult to choose the optimal value for k, and the random nature of the initialisation (simply choosing k random records) means that we may not always find an optimal solution. However, it can provide a quick and efficient method for clustering numerical data.

The data was normalised and various experiments were performed to identify the optimal number of clusters**.** Figure 3 plots the within group sum of squares against the number of clusters for 10 randomly initialised runs of the k-means algorithm. The optimal number of clusters is generally thought to be the point at which there is an "elbow" or bend in the plot – this would seem to indicate that 3 clusters is optimal.

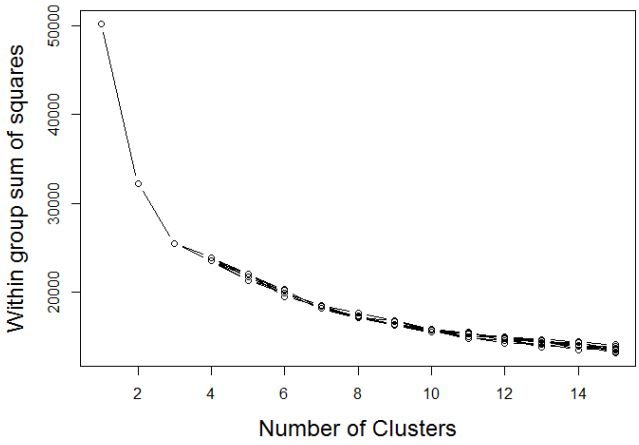

Figure 3. The within group sum of squares against the number of clusters for 10 randomly initialised runs using k-means

A clustergram (Schonlau 2004) is used to visualise cluster assignments as the number of clusters increases, using the k-means algorithm. The cluster mean is plotted for each k cluster, with the width of the lines on the graph representing how large the clusters are – therefore it is possible to visualise roughly how many records are in each cluster, and how the clusters split/join. The clusters are plotted proportional to size, by weighting the means against the first principal component of the data. The clustergram can be used as a visual aid in determining the optimal number of clusters (k) for a given data set.

Here (Figure 4) we can also see that 3 cluster centres would appear to be optimal. The data falls into three strands - even with the addition of further cluster centres those three strands still remain fairly stable. Further clustergrams were produced to check against random initialisations of the clustering (not shown for reasons of space), but from these it was concluded that three cluster centres would be optimal.

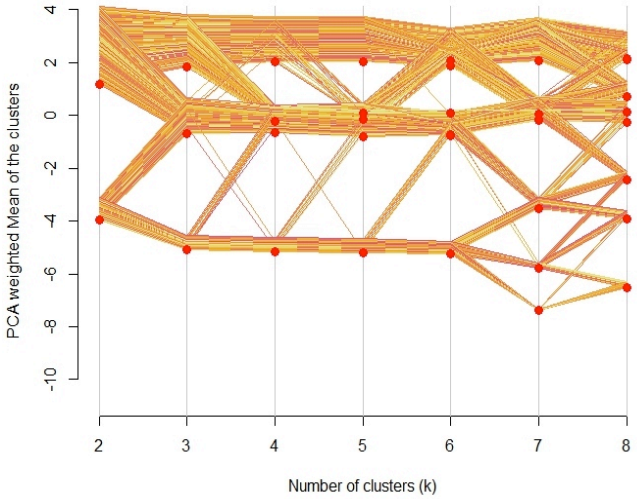

Figure 4. Clustergram of PCA-weighted mean of k-mean clusters vs. number of clusters

B. K-means clustering using 3 cluster centres

Clustering was performed on the 13 normalised attributes using 3 cluster centres. The R implementation of the k-means algorithm, which utilises the algorithm as defined by (Hartigan & Wong 1979) was used. Twenty randomly initialised runs were performed, with the best chosen. The table below (Table 1) contains the cluster means for each of the attributes in the three clusters in this clustering.

The goodness of the clustering, at 49.3% is not very high, indicating they are not very distinct. Of the three clusters, cluster 1 is fairly small (comprising 14% of the records), and the other two much larger (cluster 2 containing 35% of the records, and cluster 2 having 51%).

These clusters can be characterised as follows. **Cluster 1** is a sparsely populated outcome, with an older population, fewer friends on average, higher majority proportion, but more clustered. Relatively few adults are politically involved. **Cluster 2** and **Cluster 3** both have younger populations with a lower level of the majority ethnicity, and more immigrants, more similar friends and households, but higher levels of political involvement than **Cluster 1**. **Cluster 3** differs from **Cluster 2** in having a bigger population, lower clustering and a much lower rate of political conversation via school-related networks.

TABLE 1.
DETAILS OF THE CENTROIDS OF THE 3 K-MEANS CLUSTERS

Attribute	**Cluster 1 (543 records)**	**Cluster 2 (1333 records)**	**Cluster 3 (1986 records)**
Pop.size	100	557	1750
Av.age	76	58	55
Av.adfriends	0.73	1.36	1.82
Prop.maj	74%	67%	65%
Prop.adult	99%	94%	93.5%
Prop.1stgen	8%	13%	14%
av.clust	0.97	0.84	0.70
av.sim.hh	2.45	3.53	3.74
av.sim.fr	2.82	3.70	3.33
Rate ncvs.ac	1.3%	1.3%	0.0%
Rate ncvs.sc	0.45%	0.20%	0.13%
Prop. Adults involved	0.97%	1.6%	1.7%
Within cluster sum of squares	6748.243	11288.460	7407.591
Total sum of squares	50193		
Between SS/ Total SS	**49.3%**		

Figure 5 plots the clusters against the first two discriminants. In all, whilst the dividing line between different clusters is quite clean, it may be somewhat arbitrary just where to draw the line, as there are no clear gaps between clusters.

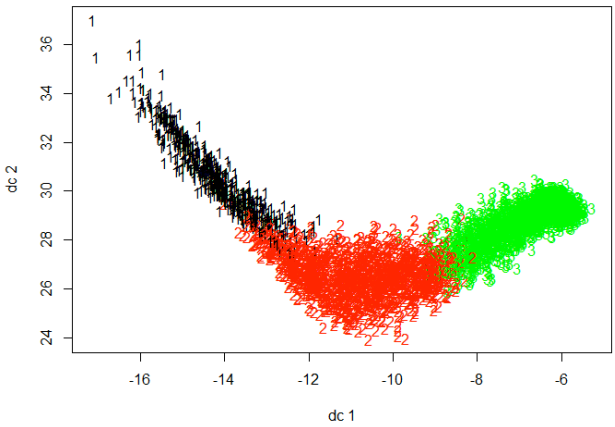

Figure 5. centroid plot against the first two discriminant functions showing the 3 clusters

If we go back to the previous hierarchical clustering and compare these three clusters with the three found using hierarchical clustering, 92% of the simulations fall into exactly the same clusters.

Figure 10 (at the end of the paper) is a pairs plot of the thirteen attributes coloured by their cluster assignments. It is difficult to read, but gives an idea of how the clusters are

distributed. Despite the fact that there are no neat lines separating the clusters in some of the dimensions, the plots do show that this categorisation spreads in meaningful patterns in each dimension. Here one can see that, for example the number of adult conversations across activity-related links is proportional to the extent of adult involvement, but only for **Cluster 2**, and that link-density only significantly varies for **Cluster 1**.

In an effort to understand how the varying input parameters might relate to the clustering of the simulation outputs, Figure 6 contains a pairs plot of the three of the varying input parameters coloured by their cluster assignments. This presents a noisier picture of how the inputs might lead to those cluster assignments (we have only shown the three relatively clear pair plots in Figure 6). Here we can see that **Cluster 1** does indeed tend to result from low immigration, high emigration parameter settings; **Cluster 3** from high immigration (either internal or international), low emigration parameter settings; and **Cluster 2** somewhere in between.

However, a method such as decision tree learning can provide a clearer view of how these inputs, in combination, produced the clusters. It may also provide a better understanding of how the model works, and allow a user to predict in advance which cluster a simulation will fall into. Figure 9 shows the pruned decision tree that defines the clusters.

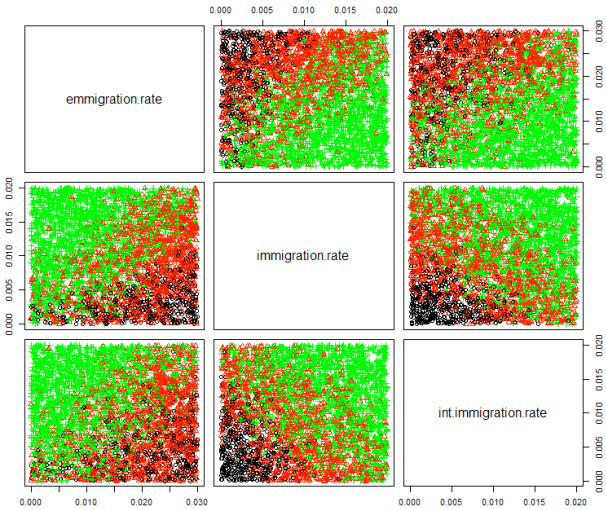

Figure 6. Shows the three clusters against the 3 parameters: emigration rate, immigration rate and internal immigration rate

C. Using varying input parameters to predict clusters

Decision tree learning is a method of predicting a target attribute (or classification), based on given input attributes. It is a form of supervised machine learning, in that the algorithm learns from labelled training data to produce a model which can then predict the value (or classification) of a target variable for new (unlabelled) data.

Decision trees are particularly popular since, in comparison to many other machine learning algorithms, their rules can be easier to understand and visualise in the form of a tree. Decision trees recursively partition data, using either the Gini coefficient or Information Gain at each step to determine the optimal input attribute to partition on. Given many input attributes, a decision tree will therefore select just those attributes that are important to predicting the target.

A tree model may over-fit data – i.e. learn the training data so well that it cannot generalise well on testing/unseen data – to avoid this, a tree is often grown overly large and then pruned back to an optimal level (using a complexity parameter).

The data was split into a training and testing data set (70:30 split) and a classification tree was built to predict the cluster assignment (derived previously from the outputs), using only the 9 varying input parameters as predictors. The rpart R package (Therneau & Atkinson 1997), which is an implementation of the Classification and Regression Tree (CART) algorithm (Breiman et al. 1983), was used to build the decision tree.

The resulting pruned tree (complexity parameter=0.0044) had **85.59% accuracy** on the testing data, and used only the attributes **emmigration.rate, immigration.rate and int.immigration.rate,** as predictors. This may therefore indicate that these particular attributes are more "important" to the data, at least in terms of predicting the previously found clustering of simulations. The tree is shown in Figure 9 in the appendix.

D. Comparison with Sensitivity Analysis

This is born out when a single parameter sweep varying just one of these dimensions is examined. So, for example, when runs of the model are done with only the immigration rate varied, and one plots the average similarity of friends for different rates one gets the graph show in Figure 7.

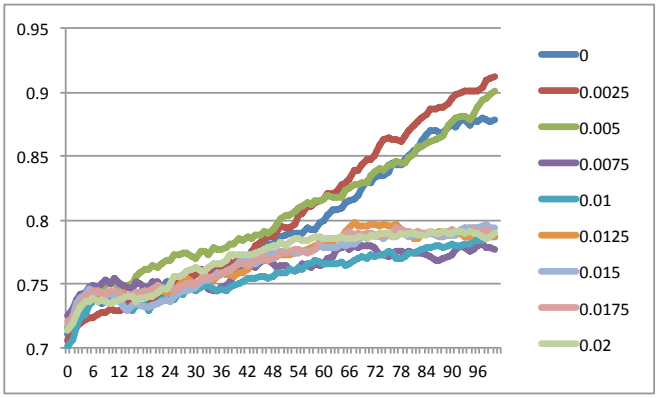

Figure 7. Average proportion of similar friends against time for different immigration rates

Here one sees a sharp division between runs with immigration rates of 0.5% and below and those above. The point of this paper is not to talk about why this happens (we hypothesise that for low rates households are able to segregate, whilst at higher rates the proportions of immigrants makes this impractical), but rather to get a broader picture of the overall behaviour of the simulation model and put particular selected results in that context.

In contrast, whilst the proportion of the original population that belonged to the majority population did impact upon the results, its influence diminishes over simulation time. Figure 8 shows how link density changes over the simulation for different initial proportions of the majority population.

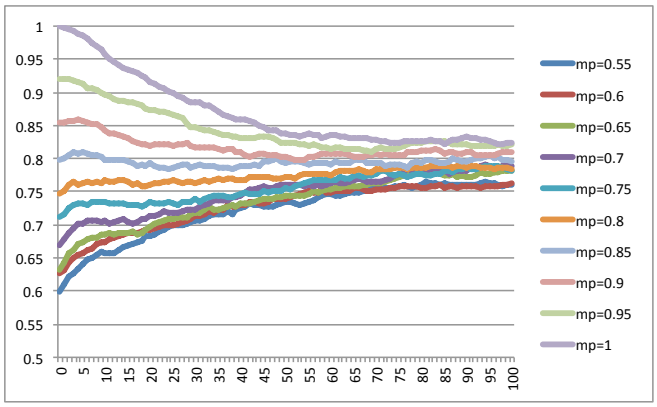

Figure 8. Average link density against time for different initial majority proportions

VI. Concluding Discussion

If complex models are inevitable, as has been argued elsewhere (Edmonds Edmonds and Moss 2005), then we will be faced with the problem of understanding them. Simple parameter sweeps and associated graphs may not be enough to characterise a complex simulation model, since they only give "thin" cross-sections of the overall behaviour. Applying Data-mining techniques to many runs of a model with randomised parameters may help to broaden the "view" of such a model, leading to a more holistic understanding. This broader view may allow a more complex understanding of the model behaviour, as well as help the researcher to focus in on which factors might influence the results more.

In general data-mining and knowledge discovery techniques have not been used much in conjunction with agent-based modelling, but this is a shame, since they both aim to understand complex data, in non-linear ways. They differ in the extent to which they are data-focussed or theory-driven – data mining being the former and ABM the latter. However both go beyond the simplistic assumptions and techniques of linear regressions models and their variants in showing some of the complexity that lies behind the data and in not hiding this within a linear fitted model.

Acknowledgment

The authors acknowledge support from the EPSRC, grant number EP/H02171X/1

References

[1] Breiman, L. et al., 1983. Classification and regression trees. Wadsworth, Belmont.

[2] Edmonds, B. (2010a) Bootstrapping Knowledge About Social Phenomena Using Simulation Models. Journal of Artificial Societies and Social Simulation 13(1)8. (http://jasss.soc.surrey.ac.uk/13/1/8.html)

[3] Edmonds, B. (2010b) Data-Integration Models (poster). European Conference on Complex Systems 2010 (ECCS), Lisbon, September 2010. (http://cfpm.org/cpmrep211.html)

[4] Edmonds, B. (2013) Complexity and Context-dependency. Foundations of Science, 18(4):745-755.

[5] Edmonds, B. and Moss, S. (2005) From KISS to KIDS –an 'anti-simplistic' modelling approach. In P. Davidsson et al. (Eds.): Multi Agent Based Simulation 2004. Springer, Lecture Notes in Artificial Intelligence, 3415:130–144.

[6] Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., Robbins, M. M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R. A., Vabø, R., Visser, U. and DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. Ecological Modelling 198 (1–2), 115-126.

[7] Polhill, J. Gary (2010) 'ODD Updated' Journal of Artificial Societies and Social Simulation 13(4):9. http://jasss.soc.surrey.ac.uk/13/4/9.html

[8] Hartigan, J.A. & Wong, M.A., 1979. A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100–108.

[9] Macqueen, J., 1967. Some Methods For Classification And Analysis of Multivariate Observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297.

[10] Moss, S. and Edmonds, B. (2005) Sociology and Simulation: - Statistical and Qualitative Cross-Validation, American Journal of Sociology, 110(4) 1095-1131.

[11] Schonlau, M., 2004. Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. Computational Statistics, 19(1), pp.95–111.

[12] Therneau, T.M. & Atkinson, E.J., 1997. An Introduction to Recursive Partitioning using the rpart Routine. Stats, 116(61), pp.1–52.

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

APPENDIX (FULL PAGE DIAGRAMS)

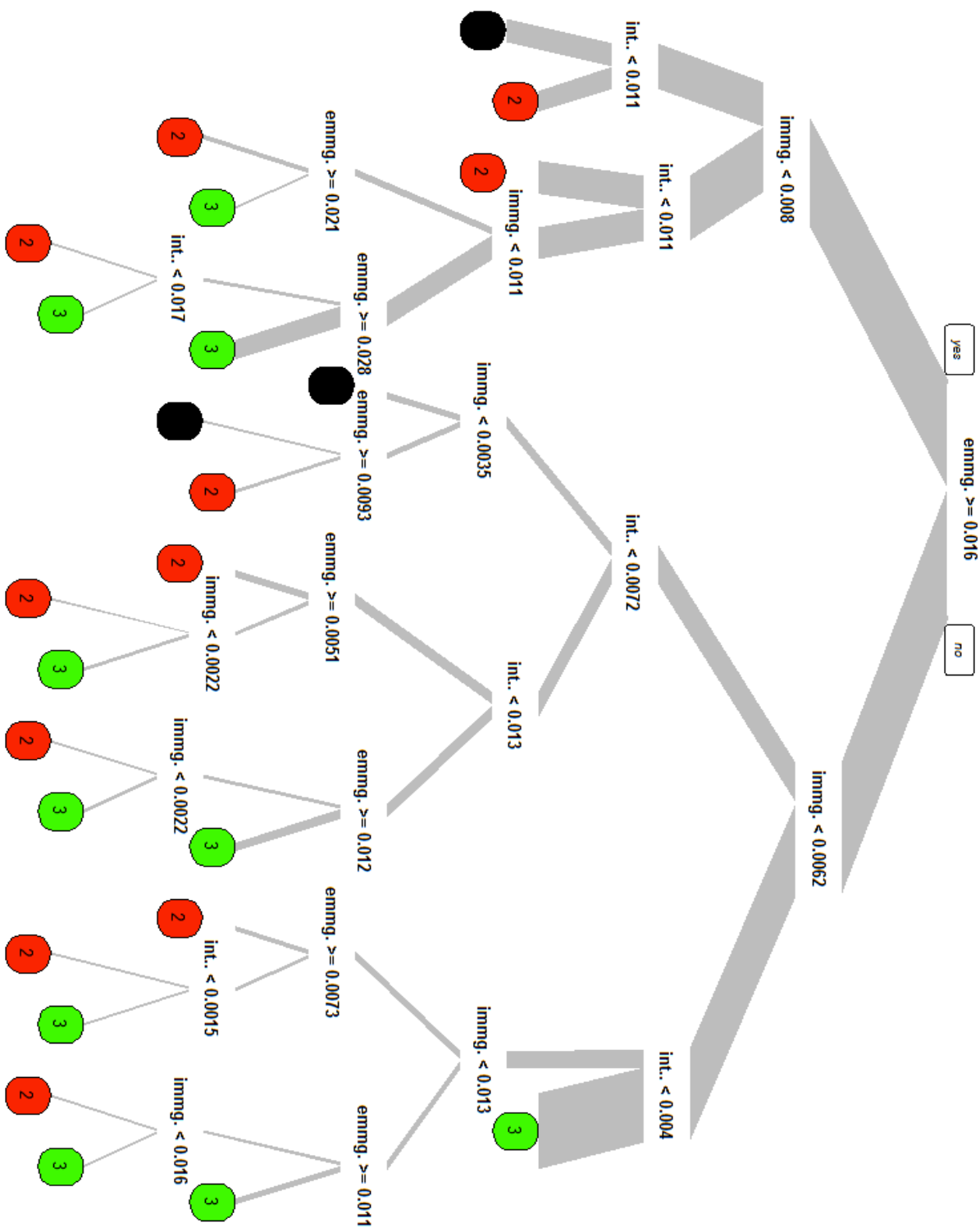

Figure 9. The induced decision tree that specifics the three clusters

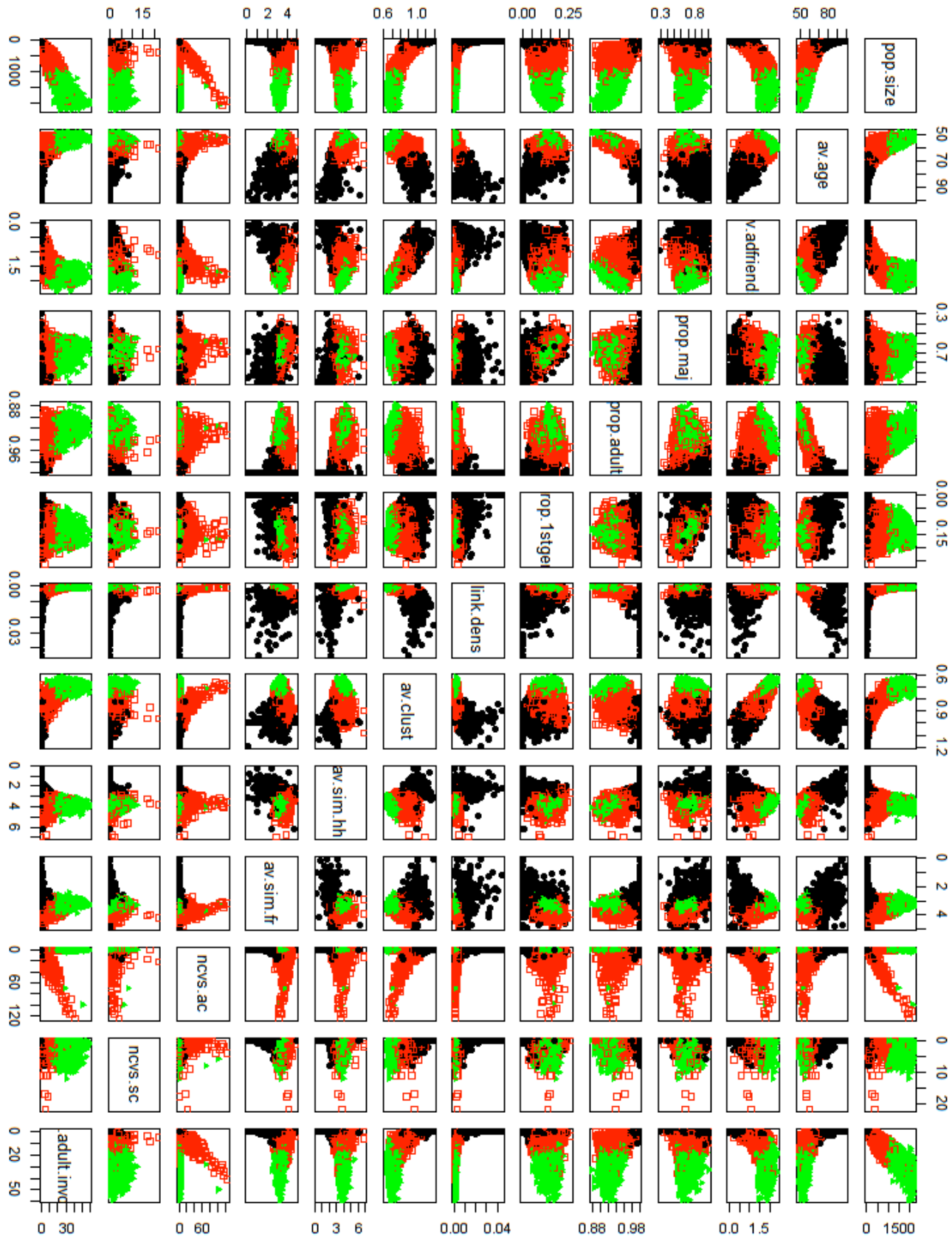

Figure 10. Scatter Plots of the Different Output Measures Against each other, with the three clusters coloured as above