# Chapter 11
# A Brief Introduction to the Use of Machine Learning Techniques in the Analysis of Agent-Based Models

**María Pereda, José Ignacio Santos and José Manuel Galán**

**Abstract** In this paper, we give a succinct introduction to some basic concepts imported from the fields of Machine and Statistical Learning that can be useful in the analysis of complex agent-based models (ABM). The paper presents some guidelines in the design of experiments. It then focuses on considering an ABM simulation as a computational experiment relating parameters with a response variable of interest, i.e. a statistic obtained from the simulation. This perspective gives the opportunity of using a supervised learning algorithm to fit the response with the parameters. The fitted model can be used to better interpret and understand the relation between the parameters of the ABM and the results in the simulation.

**Keywords** Agent based modelling · Machine learning · Simulation · Permutation test · Statistical learning

## 1 Agent-Based Modelling

Agent-based modelling (ABM) is currently one of the most active modelling paradigms in many scientific disciplines ranging from Sociology (Macy and Willer 2002) to Industrial Organization (Chang 2011) or Economics (Hernández et al. 2014). The gist of the approach lies on the particular process used to build the abstraction from the target system that is being studied. In an ABM model each

M. Pereda (✉) · J.I. Santos · J.M. Galán
Grupo INSISOC, Área de Organización de Empresas,
Dpto. de Ingeniería Civil, Escuela Politécnica Superior,
Universidad de Burgos, C/Villadiego S/N, 09001 Burgos, Spain
e-mail: mpereda@ubu.es

J.I. Santos
e-mail: jisantos@ubu.es

J.M. Galán
e-mail: jmgalan@ubu.es



C. Hernández (ed.), *Advances in Management Engineering*, Lecture Notes
in Management and Industrial Engineering, DOI 10.1007/978-3-319-55889-9_11

entity identified in the target system is explicitly and individually represented as an agent, and the different interactions among the agents and the environment are also explicitly represented in the model. This direct correspondence provides the modeller with several interesting features, most of which are a consequence of making the abstraction process easier (Galán et al. 2009). In an ABM model it is almost straightforward to remove simplifying assumptions often used in other modelling paradigms, and consider the effect of heterogeneity, spatial influence, finite populations or bounded rationality, just to mention some examples.

This advantage at the modelling stage often increases the difficulty of analysis of the model, which is sometimes so complicated that it is not easy to understand the combined effects of all its assumptions. In the case of models used to illustrate a general mechanism or an emergent stylized fact, this circumstance is usually circumvented, if possible, obtaining closed analytical solutions, or exploring the complete range of parameter combinations by simulation. However, in models trying to reproduce specific and detailed situations, occasionally it is not easy to decide a priori those assumptions that should be simplified and those that are key elements to keep the model descriptive in terms of the analysed target. Analysts then face models with a number of parameters so large that a prohibitive quantity of computational resources is required to fully explore them, and so complicated that general analytical solutions are difficult to obtain.

The aim of this paper is to discuss a set of concepts and activities used in machine and statistical learning that can help to understand the behaviour of complicated (and not only complex) agent-based models. The rest of this work is structured as follows: The following section succinctly discusses how to sample a model in an efficient manner. We then explain why it can be useful to think about the results of a model as a classification or regression problem, and present possible avenues that an analyst can follow to adjust and interpret the model. Subsequently, a common way to analyse variable significance is discussed, and finally, conclusions are presented in the last section.

## 2 Design of Experiments with Space Filling Properties

An experiment is a procedure in which the input variables of a model (i.e. the system under study) are changed in order to analyse the reasons for the change in the *response variable* (a.k.a. output variable). The conduction of formal planned experimentation, i.e. Design of Experiments (DOE), is a crucial step that consists in getting the maximum amount of information from the model with the minimum amount of resources (the more samples, the more CPU time). This is carried out by properly choosing samples in the *design space* (part of the parameter space under study).

Given a parameter space in an ABM model, there are different ways of selecting samples and design experiments (Lee et al. 2015). Some of them have a particular configuration, such as Factorial Design, Central Composite, Taguchi, among others,

consisting in discretising the parameter ranges in levels and sampling these values. Other approaches are the so-called space filling techniques, which aim to cover the design space uniformly and are based on statistical sampling.

When exploring a stochastic model, deciding whether to spend resources on exploring more diverse parameter combinations or replicating several times a given combination of parameters to reduce the uncertainty about the expected output value implies a trade-off that should be balanced and that is case-dependent.

The most obvious sampling technique is the Monte Carlo random sampling, which consists in sampling each parameter range randomly. The problem with this technique is that the design space is not covered evenly, but there can appear clusters of samples and empty spaces by chance. Another common issue while planning a DOE for computer models is that, sometimes, large design spaces need to be explored.

A space filling sampling technique that enjoys great popularity in computer simulation is Latin Hypercube Sampling (LHS) (McKay et al. 1979), since it provides an even sample set that is representative of the sampled space. Its popularity is explained by the fact that a DOE with a desired number of samples can be created, and because of its flexibility (dimensions can be dropped out from the DOE and still have a LHS, because the samples are non-collapsing (Viana 2013). The main drawback of LHS is that it suffers from the curse of dimensionality, where large LHS designs can have inter-variables' correlations (Viana 2013) and space filling properties become questionable. There have appeared some methods to avoid these drawbacks, such as orthogonal arrays and orthogonal LHS, but at the cost of complex optimization algorithms (Viana 2013).

In LHS, for an $N$-dimensional design space, each parameter range is divided in $p$ uniformly spaced levels, thus producing $S = p \cdot N$ subspaces. Each level is uniformly sampled only once, ensuring that the full space is sampled, and the resultant number of samples is $p$.

## 3 Results as Classification and Regression Problems

Whether the purpose of an ABM model is to provide precise quantitative predictions or simply a better understanding of the logical implications of the model hypotheses, the task of analysing the relation between the model output and parameters is usually not simple or easy. The difficulty is greater as the amount of parameters is larger, complicating the use of the traditional graphical techniques to draw inferences. In general, but particularly in models with high dimensional parameter spaces, machine learning techniques can be usefully applied to analyse ABM models.

In order to understand a model, it is absolutely necessary to understand the relationship between the model parameters and the model output. In general, we usually define a statistic representative of the behaviour of the model and analyse the values that it reaches after a number of time steps—the probability function over

the set of values—(Izquierdo et al. 2009). In some cases our interest focuses on the asymptotic behaviour of the model for which the statistic is determined by their absorbing states or stochastically stable states (Izquierdo et al. 2009); in others, we want to figure out the state distribution at a time of special interest for the research case study. In any case, the sort of inferences about the statistic, i.e. output variable $Y$, can be described as a function $f$ of the model parameters, i.e. input variables $X = (X_1,..., X_p)$ plus an error term of mean zero and independent of $X$ (see Eq. 1). This last assumption may be problematic. In such cases, different strategies can be applied depending on the dependence.

$$\mathrm{Y} = \mathrm{f}(\mathrm{X}) + \text{ error} \tag{1}$$

We talk about a regression problem if the statistic takes on numerical values (quantitative values); otherwise (qualitative values) we talk about a classification problem. Even with quantitative variables, sometimes the state distribution is composed of a reduced set of states for which the values of the statistic can be grouped in classes, thus turning the problem into a classification problem too. In these cases, classification methods become very useful to explain the relationship between the statistic and the parameters.

Regardless of the type of statistic variable, we can discuss some important issues. The essence of the problem is to estimate the unknown function $f$ based on simulation data set. Machine Learning provides a set of parametric and non-parametric methods to solve this supervised learning problem. The selection of a specific learning method usually determines the form of the function $\hat{f}$, used to estimate $f$, so this choice conditions the interpretability of the results. Sometimes $f$ takes a linear form, making it easy to understand the influence of the parameters on the statistic; other times $f$ is more complex, making the inference more challenging.

The expected test error of a learning method, i.e. the expected error when the estimated function $\hat{f}$ is evaluated on new data not used in the training, can be decomposed in the sum of three terms: the bias error, the variance error and the irreducible error (Hastie et al. 2009). In simple words, the bias error is due to using a function $\hat{f}$ that is not flexible enough to fit the unknown function $f$. The variance error represents the expected change of the estimated function $\hat{f}$ when using different training data. Finally, the irreducible error gathers the natural noise of data, which is the variance of the error term in Eq. (1).

When we seek an estimate $\hat{f}$ of the function $f$ we always face a bias-variance trade-off. The goal is choosing a method with the smallest test error, meaning low bias and low variance simultaneously. This issue is related to the flexibility of the learning method, i.e. the degrees of freedom of the function $\hat{f}$. More flexible methods have less bias error, but may overfit data and present higher variance errors than less flexible methods. Frequently the flexibility of a learning method is inversely related with its interpretability, i.e. the ease to explain the relationships between the model output and its parameters. For example, linear regressions are

**Table 1** Some popular supervised learning techniques sorted in increasing order of flexibility

| Technique |
|---|
| Linear and logistic regression with regularization (Ridge regression, the lasso) |
| Linear and logistic regression |
| Kernel smoothing methods |
| Trees |
| Boosting methods |
| Neural networks |
| Random forests |
| Support vector machines |

more interpretable than kernel smoothing methods (see Table 1), something which is desirable in terms of inference; however, kernel smoothing methods may offer greater flexibility at the expense of overfitting the data losing interpretability.

Assuming that we have selected a function $\hat{f}$ cross validation (CV) techniques (Hastie et al. 2009) can be applied to provide an accurate estimate of the test error. In particular, the k-fold CV divides randomly the data into k subsets of approximately equal size, called folds, and use k-1 of these subsets as training set and the remaining subset as test set. The process is repeated k times, using different subset combination and averaging the results over the test sets (Hastie et al. 2009). If we have several estimating functions to choose from, the same technique can be used to select the best one. Normally, this result is enough to proceed with the inference about the model. However, the estimated test error might not be an unbiased estimate of the performance of the selected function. If an unbiased estimation of the test error of the chosen function is needed, we could apply other refined techniques such as nested CV (Varma and Simon 2006), which develops the essence of CV using two nested loops, an inner loop for function selection and an outer loop for estimating test error.

Table 1 gathers the most important machine learning techniques ordered by increasing flexibility. In general, the election of more flexible methods can be useful when we conduct a preliminary analysis, to get an overall insight to the interactions between all model parameters, while less flexible methods are better for detailed inferences about particular relationships between parameters (Santos et al. 2015). Obviously, the particular research interest always drives the election of the learning method.

## 4 Variable Importance Analysis

The process of fitting a function $f$ using the parameters as predictors is not only relevant to predict the value of the response variable of interest. Once a regression or a classification model has been adjusted, it is also possible to identify the relevant parameters with the greatest impact on the results. The rationale to conduct this

analysis in the context of ABM comes from two related features: (i) the function $f$ can be simplified and hence could be more amenable for interpretation and (ii) the computational effort in subsequent experiments can be focused on those parameters with high influence.

This problem is known in the machine learning and statistics community as feature or variable selection (James et al. 2013). There are several classes of methods to address this problem, e.g., shrinkage or regularization, dimension reduction or subset selection. Some of the subset selection approaches are based on estimating the variable importance for each possible predictor using the fitted function $\hat{f}$. The concept "importance" tries to capture the contribution of each variable to the function $f$.

In recent years, random forests (Breiman 2001), an ensemble learning method that employs trees as weak learners, have become one of the most popular and widely used techniques in many scientific disciplines. This popularity not only comes from the good predictive performance in classification and regression problems (even in high dimensional and non-linear problems). Random forests are also appealing since they provide a very natural way to find out the importance of each predictor (Criminisi et al. 2011).

A first measure of variable importance in classification problems with ensembles of trees is the decrease in the node impurity measure used to train the trees in each splitting criterion. However a more sophisticated variable importance measure is the "permutation accuracy importance" measure (Strobl et al. 2007). The idea is as follows: given a set of predictor variables $X_j$ used to predict a response variable $Y$ using a random forest, in order to measure the importance of the $X$th variable after training, the values of this variable are permuted among the training data and the (out of bag) error is computed and compared before and after the permutation over all trees. This score is sometimes normalized using the standard deviation. Variables with large values are those with higher influence in the response and are therefore considered more important. The underlying assumption of this permutation test is that, by randomly permuting the variable under study, its original association with the response is broken. If that association was originally relevant the prediction accuracy of the forest will decrease, and the higher the decrease the higher the variable importance (Strobl et al. 2007).

Although other variable importance methods are available using different machine learning algorithms (Altmann et al. 2010), the simplicity and interpretability of the permutation test used in random forests and their suitability for being used in complicated situations have made them very popular as a method to gain insight into the significance of the different variables. Notwithstanding, recent research (Strobl et al. 2007, 2008; Wei et al. 2015) has pointed out different sources of bias that can affect this importance measure in random forests. The standard method using CART trees can be hindered in cases where variables vary in their scale level or in their number of categories, or in situations in which there are correlated predictor variables. In those cases, conditional permutation schemes and unbiased conditional inference trees are recommended to reflect the true relevance of the different potential features (Strobl et al. 2007, 2008).

## 5 Concluding Remarks

The aim of this paper has been to briefly identify some of the concepts from the Machine and Statistical Learning fields that can be useful to the ABM community. We have discussed that this approach can help to improve the understanding of complicated models with many parameters that can be difficult to fully explore. Given that such models can have a huge parameter space and the limited computational resources available, we have discussed some alternatives to explore the parameter space in a more efficient manner.

Considering an ABM model as a function that relates its parameters with its results, one can usefully employ different supervised learning mechanisms to fit such a function. The constructed estimate of the function is an approximation of the agent-based model that can be useful for many purposes. In some cases the approximated fitted function can be more amenable to interpretation and understanding; in other cases it can be useful for generalization and visualization; and very often it can serve to assess the individual impact of the parameters on the results. We have discussed this last analysis process in the case of random forests, but this work is far from being an exhaustive review of all the methods—each one with their advantages and disadvantages—that can be used in the analysis of ABM models.

The insights obtained using these techniques can guide subsequent steps of the modelling process, such as simplifying the model or focusing on a finer grain analysis of the most influential parameters. But they can also be useful from an empirical and a policy-making perspective, leading the efforts of calibration or control into the most relevant variables in the target system.

**Acknowledgements** The authors would like to thank Dr. L.R. Izquierdo for some advice and comments on this paper. The authors acknowledge support from the Spanish MICINN Project CSD2010-00034 (SimulPast CONSOLIDER-INGENIO 2010) and by the Junta de Castilla y León GREX251-2009.

## References

Altmann, A., Tolosi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347. doi:10.1093/bioinformatics/btq134

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi:10.1023/A:1010933404324

Chang, M. H. (2011). Agent-based modeling and computational experiments in industrial organization: Growing firms and industries in silico. *Eastern Economic Journal, 37*(1), 28.

Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Computer Vision, 7*(2–3), 81–227. doi:10.1561/0600000035

Galán, J. M., Izquierdo, L. R., Izquierdo, S. S., Santos, J. I., del Olmo, R., López-Paredes, A., & Edmonds, B. (2009). Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation*, *12*(1), 1. http://jasss.soc.surrey.ac.uk/12/1/1.html

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hernández, C., Galán, J. M., López-Paredes, A., & del Olmo, R. (2014). Economía Artificial: Métodos de inspiración social en la resolución de problemas complejos. *Revista Española de Física, 28*(3), 23–30.

Izquierdo, L. R., Izquierdo, S. S., Galán, J. M., & Santos, J. I. (2009). Techniques to understand computer simulations: Markov chain analysis. *Journal of Artificial Societies and Social Simulation*, *12*(1), 6. http://jasss.soc.surrey.ac.uk/12/1/6.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Lee, J. S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., Voinov, A., Polhill, G., Sun, Z., & Parker, D. C. (2015). The complexities of agent-based modeling output analysis. *Journal of Artificial Societies and Social Simulation 18*, 4. http://jasss.soc.surrey.ac.uk/18/4/4.html

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology, 28,* 143–166.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics, 21*(2), 239–245. doi:10.2307/1268522

Santos, J. I., Pereda, M., Zurro, D., Álvarez, M., Caro, J., Galán, J. M., & Briz i Godino, I. (2015). Effect of resource spatial correlation and Hunter-Fisher-Gatherer mobility on social cooperation in Tierra del Fuego. *PLoS ONE, 10*(4), e0121888. doi:10.1371/journal.pone.0121888

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307. doi:10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25. doi:10.1186/1471-2105-8-25

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(1), 91. doi:10.1186/1471-2105-7-91

Viana, F. A. C. (2013). Things you wanted to know about the Latin hypercube design and were afraid to ask. In *10th World Congress on Structural and Multidisciplinary Optimization*, Orlando, Florida, USA.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety, 142,* 399–432. doi:10.1016/j.ress.2015.05.018