2020

A Framework for Validating and Testing Agent-based Models: a Case Study from Infectious Diseases Modelling.

Elizabeth Hunter
Technological University Dublin, elizabeth.hunter@tudublin.ie

John Kelleher
Technological University Dublin, john.kelleher@tudublin.ie

Recommended Citation

Hunter, E. & Kelleher, J. (2020). A framework for validating and testing agent-based models: a case study from infectious disease modelling. *34th. Annual European Simulation and Modelling Conference*Toulouse, France, 21-23 October, 2020. doi:10.21427/2xjb-cq79

A FRAMEWORK FOR VALIDATING AND TESTING AGENT-BASED MODELS: A CASE STUDY FROM INFECTIOUS DISEASES MODELLING

Elizabeth Hunter and John D. Kelleher
Technological University Dublin
email: elizabeth.hunter@tudublin.ie

KEYWORDS
Agent-based model, Discrete Simulation, Validation, Verification, Infectious Disease Model

ABSTRACT

In this paper we present a framework for validating and testing an agent-based model that includes determining the appropriate number of runs to account for variation in model runs, validating the model and showing that the model can be used to learn about the system. To demonstrate the framework we use a case study of an agent-based model for the spread of infectious diseases.

INTRODUCTION

Agent-based models are a type of simulation that are made up of agents that can interact with other agents and with their environment. The actions of these agents are determined by a set of probabilistic rules. Agent-based models are typically stochastic due to the agent's abilities to make decisions and react to their environment and the state of the model. Because of this stochasticity, each model run can produce different results. Additional variation in model results is often caused by different initial conditions for each model run. Thus agent-based model results are a distribution of possible outcomes of the system being modeled. This distribution of results can lead to challenges in presenting the model results and proving their validity and usefulness. Agent-based models can have a number of different purposes including prediction, description, and theoretical exploration (Edmonds et al. 2019) and the different purposes might necessitate changes in how validation and testing a model is done. It is also necessary to tailor the methods used for validation so that they fits with the requirements for the results of the model. For example looking at an average of a single statistic calculated over the model runs versus looking at changes in agent behaviour or looking at a distribution of statistics.

Here we present a framework for validating and testing an agent-based model that is not limited to a specific model purpose but is instead designed for a model where the results are presented with a single statistic such as an average, a percent of runs where an event occurs or a threshold has been reached, or a single statistic measuring the change in a model output. We start the paper by discussing the existing literature on validating and testing agent-based models and then present our framework and illustrate each step of the framework via an accompanying case study of validating and testing agent-based models for infectious diseases.

BACKGROUND

A major advantage of using agent-based models is that the approach allows for flexibility in terms of the simulation design. However, once a model is created it does not mean that the model is valid and producing accurate results. There is no broadly accepted method to validate agent-based model. This poses a major challenge to the field of agent-based modelling because if a model is not properly validated, any results, especially surprising results, that come from the model cannot be trusted (Richiardi et al. 2006).

The stochasticity in an agent-based model leads to a distribution of results across model runs; consequently, a prerequisite to model validation and testing is to decide on the representation of the distribution. Often an aggregate statistic calculated across model runs is used, for example an average across the runs or a percentage of runs where a certain threshold has been reached.

In a review of agent-based models that calculated the statistical power of the models (the power of a statistical test is the probability that it correctly rejects a false null hypothesis), Secchi and Seri (2017) find that the majority of papers do not perform any estimations of the robustness of the proposed model and most studies did not have enough runs to account for the stochasticity in the model. This could have an impact on the validity of theses agent-based models. In developing our framework we have included a process for selecting the number of models runs necessary to calculate a statistic.

The ideal scenario for validating a model, is comparing the model results to real data, but data is not always available. Thus other methods of validation need to be used. Within the literature we have found two other validation methods that are commonly used: (i) cross-validation or comparing the results of the model to a previously validated model (Rakowski et al. 2010a, Skvortsov et al. 2007); and, (ii) showing that appropriate decisions were made in creating the model and that

small change in inputs do not disproportionately affect the model (Apolloni et al. 2009, Xia et al. 2013). As a result our model validation process covers all three of these approaches; cross-validation, a sensitivity analysis, and comparing the model results to real data.

The final stage in our framework is model testing. In this paper we make a distinction between model validation and model testing. Validation shows that the model results are trustworthy and that the model is accurately simulating a system whereas testing shows that the model can test a hypothesis or can add to the existing knowledge about the system being modelled. Both are important in determining the usefulness of a model. The following sections describe each step in the framework in greater detail and discuss the application of each step using a case study of an agent-based model for the spread of measles in a town. For each model run, a different outbreak scenario (in terms of the population and the initial infection vector being stochastically sampled) is created, and to aggregate the data the main statistic that we look at is the percent of runs that result in a measles outbreak occurring in the town. Note that model creation, validation, and testing is a detailed process that requires a large amount of work and thus this case study draws on research from a number of previous papers: Hunter et al. (2018) and Hunter et al. (2020) discuss validation, and Hunter and Kelleher (2020) discusses testing. However, this case study extends beyond this prior work in terms of explaining the process used to determine the appropriate number of runs necessary for the model, and also situates the work in these papers within the overarching validation and testing framework proposed in this paper.

Stochasticity and Confidence Intervals: Selection of Number of Runs

We propose using the statistical concept of a confidence interval as the basis for determining the number of runs necessary to ensure the calculated statistic is representative of the distribution of outcomes the model is sampling from. Confidence intervals relate the uncertainty about the true value of a statistic (i.e., variability across a number of samples the statistic is calculated over) with the sample size used in the calculation of the value. The relationship between these elements is that in general as sample size increases the uncertainty relating to a statistical value decreases and this is reflected in the confidence interval around the statistical value becoming narrower. Confidence intervals provide an ideal basis for determining the number of runs necessary for an agent-based model: in this context, the sample size is number of times the model has been run and so confidence intervals relate the uncertainty relating to the value of a statistic calculated over the model runs with the number of model runs.

For a given model, our goal in using confidence intervals is to find out how many runs of the model is typically required in order for us to have a high confidence that the statistic calculated over those runs is representative of the true parameter of the distribution the model is sampling from. The assumption we make in doing this analysis is that if we determine the necessary number of runs of a model for one model scenario (e.g., for one population of agents, environmental, or intervention scenario) then we can use the same number of runs of the model to calculate the statistic in a different (but similar) scenario.

To do this we run the model on a scenario a pre-set number of times: more than what would be expected to account for the variability. Once the model is run a given number of times, we take samples of increasing sizes, the first sample contains the first 5 runs, the second sample contains the first 10 runs, the third sample contains the first 15 runs, until the final sample contains all runs. For each sample we look at one main output statistic and the size of the 95% confidence interval around that statistic. By definition a confidence interval is determined by both the sample size and the variability in the sample: as the number of runs in the sample increases the size of the confidence interval will decrease. The formula for a confidence interval varies based on the parameter being estimated: mean, proportion, difference in mean or difference in proportion, and the appropriate formula should be selected for the model. In the Case Study section we give the formula for the confidence interval for a single proportion and demonstrate how it can be used to determine the appropriate number of runs for a model.

The desired size of the confidence interval will vary based on what is being measured but it should be sufficiently small enough to provide enough confidence that the aggregated statistic is close to the true population parameter and that our results would not change significantly if the model was run additional times. An important benefit of using confidence intervals in this way is that they naturally account for the complexity of the distribution of the model, or in other words the calculation of number of runs is dependent on the variation of outcomes across the runs: the larger the variation in model outcomes the more runs are required to meet the confidence interval criterion.

We see determining the number of runs as an important step in showing that the results of the initial model account for the stochasticity in the model. Additionally, once the calculation for the number of runs for a given model has been determined it is not necessary to do the calculation again for each different scenario the model is run on. For example, if a base model is created and the model is run 500 times to determine that 200 runs are needed to account for stochasticity, then when interventions are added to the model or the model is ported to an new society, each additional scenario will only need to be run 200 times.

Case Study: The design goal of our town agent-based model was to simulate the spread of measles in a number of different Irish towns. In order to find the number of runs necessary to find a stable result for the model three towns with varying characteristics were initially modelled. The three initial towns were Schull (population 987, Area $17.03km^2$), Tramore (population 9,548, Area $16.60km^2$) and Kenmare (population 2,912, Area $55.61km^2$). For each town we ran the model 1,000 times keeping all factors besides those related to the town population, such as age specific vaccination rates, area or town layout, constant.

In this case study, the target statistic we were examining with the model was the percent of runs that led to a measles outbreak occurring in the town. To define the outbreak condition for a run we used the World Health Organization's definition of a measles outbreak of two or more linked cases of measles. As the outbreaks in the model are affected by the actions of the infected agents not all runs lead to an outbreak and the percent of runs that did lead to an outbreak was found to vary between towns in the model.

In order to find the appropriate number of runs for this model we divided the runs for each town into different samples, the first sample containing the first 5 runs, the second sample containing the first 10 runs, and so on until all the runs are included in at least one sample. Then for each sample we calculated the target statistic and the 95% confidence interval for the statistic given the sample. As the statistic was a percent or proportion, the formula used is that for the confidence interval around a single proportion. The equation is presented in Equation (1) where $\hat{p}$ is the proportion, n is the sample size and z is the z statistic for the appropriate confidence level. For our agent-based model, $\hat{p}$ is the proportion of runs that lead to an outbreak, n is the number of runs and z is 1.96 which creates a 95% confidence interval.

$$\hat{p} \pm z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (1)$$

Table 1 shows selected values of the statistics and confidence intervals for different sample sizes from the 1,000 model runs. From Table 1 it can be seen that the size of the confidence interval decreases as the size of the sample increases.

Looking at the size of the confidence intervals, the length of the 95% confidence interval for all three towns is approximately 10% when the sample size is 300 runs.

The confidence interval for 300 runs of Kenmare has a length of 9.5%; the confidence interval for 300 runs of Schull has a length of 10.8% and the confidence interval for Tramore has a length of 10.2%. Therefore, we determined the 300 runs was an appropriate number of runs for this town model when calculating this statistic. Using 1,000 runs instead of 300 we cut the length of the confidence interval down to 5%, however, we decided that decreasing the size of the confidence intervals by 5% with an extra 700 runs was not worth the time and computing power required.

Table 1: Percent of Runs Leading to an Outbreak and Confidence Intervals

Town	Sample Size	Percent	Confidence Interval
Kenmare	5	80.0	(44.9 115.1)
	300	77.6	(72.9 82.4)
	1000	77.4	(74.8 80.0)
Schull	5	100.0	(100.0 100.0)
	300	65.0	(59.6 70.4)
	1000	66.6	(63.7 69.5)
Tramore	5	40.0	(-2.9 82.9)
	300	72.0	(66.9 77.1)
	1000	70.0	(67.2 72.8)

While a 95% confidence interval is a standard selection for confidence intervals and can be used for any model, the confidence interval length of 10% is model specific. It is selected here due to the use of percents as a statistic and the magnitudes of the percents, however, for other models or different statistics a smaller or larger confidence interval might be better suited. The advantage of using confidence intervals in this way is that they can be applied to any agent-based model to determine the appropriate number of runs for that particular model. That is why we believe that this method is a useful approach to make informed decisions on the number of runs to use for an agent-based model in a given domain.

Model Validation

The next step in our framework is the validation step. Once we are sure that our results are statistically robust, its necessary check that the model is producing the expected results and capturing the desired dynamics.

There are three parts to our validation framework: cross validation, sensitivity analysis and finally comparing to real data. While we think that all three are important in validating a model, in some scenarios all three might not be necessary or possible. For example, real data to compare the model results to might not be available. The following sections describe each part of the validation framework in more detail.

Cross Validation

We first consider cross validation: using the results of another previously validated model as a baseline for the agent-based model results. Here the outputs of the agent-based model are compared against the outputs of the previously validated model. If this comparison shows that the outputs of the two models are similar then this cross-validation provides evidence that the agent-based model is producing the overall general dynamics that would be expected in the type of system being modelled. This is important to show especially

in the case of a model simulating something such as an infectious disease outbreak where the dynamics are well known. If the new model does not produce the expected dynamics then the model should not be validated. If the model does produce the expected dynamics then the modeler can move on to the next step in the validation framework.

Case Study: In infectious disease modelling the most commonly accepted and used model type are SEIR (susceptible, exposed, infected, recovered) models which have been shown to capture the macro level dynamics of an outbreak. Consequently, for our work an SEIR model was a natural choice to cross-validate the agent-based model. If a simple version of the agent-based model was able to emulate the general dynamics of an SEIR model of the same domain it could be concluded that the basic dynamics of the model are working correctly and that complexity can be added to the model. In the town model, to compare the agent-based model to the SEIR model both models were run with an entirely susceptible population. Figure 1 shows the results for the SEIR model and 10 runs of the agent-based model. Although they do not produce exactly the same results, the figure shows that the two models follow the same dynamics: for example, in both cases the peak of the exposed is before the peak of the infected and the exposed peak is higher. Our agent-based model of infectious diseases

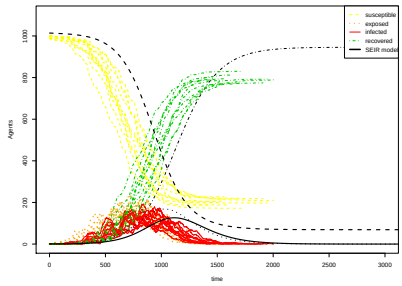

Figure 1: Cross Validation Comparing a SEIR Model to an Agent-Based Model (Black Plots are the SEIR Model)

was used to analyse disease dynamics at two geographic scales: at the individual town level, and at region level. Although the town level version of our model was cross-validated against an SEIR model, to cross-validate the region level model we used the town agent-based model. To do this, outbreaks in the individual isolated towns were compared to the outbreaks in the towns where commuting to other towns was allowed but no agents who lived outside the town were included and also to simulated outbreaks in the town when the towns were connected in a larger network of towns that make up a region. Table 2 shows results previously presented in Hunter et al. (2020) that were used to cross validate the county model with the town model. While the results for the isolated town and the town within a county were not identical, the results for the county model are between the isolated and commuting town model results which we interpret as showing that the outbreaks within a town were somewhat stable when the town becomes part of a network but that commuting and interacting with other agents outside of the town had an effect on the results.

Table 2: Cross Validation from Hunter et al. (2020)

Model	Manorhamilton	Kinlough
Town Model	66.3 (61.0, 71.7)	48.0 (42.3, 53.7)
Town Commuting	41.0 (35.4, 46.6)	32.0 (26.7, 27.3)
County Model	52.7 (47.0, 58.3)	42.5 (36.1, 47.2)

The fact that the town agent-based model is cross-validated with an SEIR model, and that the region agent-based model is cross-validated with the town agent-based model, demonstrates the potential flexibility of the cross-validation approach in terms of defining the baseline model that validation is carried out against.

Sensitivity Analysis

After cross validation is completed it may be necessary to consider a sensitivity analysis. Although sensitivity analysis is often done to analyze how stochasticity affects the variability of results (Raimbault et al. 2019), here we use a sensitivity analysis as a method to determine if the response of the model to changes in its parameters is appropriate. To do this it is important to determine what parameters should be investigated and what range of those parameters should be used in the analysis. Determining which parameters to vary and what range of values to use during sensitivity analysis is done on the basis of domain theory. For example, in the field of infectious diseases we would expect that as the vaccination rate for a disease in a population drops the number of infected people in a simulation of an outbreak within that population should generally increase. As we will discuss below in the case study, this type of domain information provides a basis for designing sensitivity analysis tests for a model. If as the values of the parameters are changed something unexpected occurs in the model results, the results should be investigated to determine why the model is not behaving as the system being modeled is expected to. In doing a sensitivity analysis it is important to test the extremes that might not be likely to occur but where the model might break down. While in most cases a sensitivity analysis is important in validating the model, the requirement for sensitivity analysis of a model does not always hold, for example if a model is being scaled up from a previously validated model and no new parameters are added.

Case Study: The parameter selected to show in the case study for sensitivity analysis is the vaccination rates of the agents. The parameter was selected as it is a parameters that there are strong expectation with regards to how the model results should change as the parameter changed. Table 3 shows results for three vaccination scenarios which shows that the average number of infected agents across model runs and the number of runs leading to an outbreak decrease as the immunity level in the population increases. As the expected patterns were observed when changing vaccination rates this helps to show that the model is valid.

Table 3: Sensitivity Analysis from Hunter et al. (2018)

Immunity Level	Average Infected	Percent Outbreaks
No Immunity	726.7	94
All Immune	1.0	0
Herd Immunity	3.3	52

Comparison to Data

The last step in validation process is comparing the results of the model to real data. When comparing the results of the model to data it is essential to understand that when run multiple times each run of an agent-based model simulates a different possible scenario and the majority of runs will not predict the exact data. For the case of an infectious disease model, the type of data typically used for this comparison are previous outbreaks of a disease in a population. The course of an infectious disease outbreak depends on a number of factors, including who the initially infected individual was and the decisions of that individual about staying home when sick or going out to potentially infect others, and the attributes of the other agents the infected agent comes into contact with such as their vaccination status. Thus when comparing to real data, we want to make sure that the distribution of outcomes generated by the model is such that the real data could be considered a likely sample from the distribution.

Case Study: The actual steps in the comparison to real data will be determined by the domain of the model and the data available but here we compared the results from a simulated measles outbreak in the town of Schull in Ireland, to a real outbreak in Schull in 2012. While in the other analyses for the town model the percent of runs that lead to an outbreak was the main statistic that was studied, when comparing to a real outbreak we needed to consider other ways to compare our model results to the real outbreak data.
One measure to look at was the total number infected in the outbreak which was compared to the number of infected agents in the model. As the agent-based model produces a distribution of runs each driven by a different outbreak scenario, we did not expect to have every model run produce the exact number of infected in the real outbreak. Instead we expected the real number to be in the range of model results, to determine if this was the case for the town model, we checked if the real data was within the interquartile range of the model result. As it was, we showed that the real outbreak is one of the possible scenarios simulated in the model.

Model Testing

The final step in the model validation and testing process is testing the model. A model is only useful if it can help us learn something that we did not already know about the system. Thus it is important as a final step of the modelling process to determine if the model is able to do that. Similar to the comparison to data the model testing can be domain specific but should show some aspect of how a model can be ported to different scenarios, for example different populations, or how changing different inputs to the model or agent actions might influence the system.

Case Study: The county model was tested by looking at the role of the centrality of a town in the spread of an outbreak (Hunter et al. 2020) and also through using the knowledge gained from the centrality analysis to implement a school closure intervention. School closure strategies are often questioned in their effectiveness in mitigating an outbreak especially with their high social and economic cost. The county model tests a school closure strategy that used the knowledge of a town's place in a network of towns to decide which schools to close based on distance to the town where the outbreak starts and the centrality of the towns in the county's commuting network. The study showed that closing the schools in the town where the outbreak starts and the highest centrality town in the commuting network had a mitigating effect on the outbreak (Hunter and Kelleher 2020). This allowed us to conclude that not only is the model working as we expect it should but that it is a useful model that can help us to learn about the system being modelled which can aid in planning for future outbreaks.

CONCLUSION

A framework for model validation and testing is important when creating a model to show that the results can be used for accurate predictions. If agent-based models are to be routinely used as policy tools a consistent validation method should be determined. Without such a method it may be difficult to distinguish a model that will provide accurate results for a given population from a model that will not. However, there is no set framework in the literature. In this paper we presented a framework that can be applied to agent-based models,

and used an agent-based model for the spread of infectious diseases as a case study. While the framework was developed in validating the model for infectious disease spread it contains methods that will be useful to models in other domains. The framework is multi-step and involves determining the number of runs that are necessary to make sure that the model results are representative, validating the model through cross-validation, sensitivity analysis and comparing the model to real data if possible, and finally testing that the model can provide useful results.

What we have proposed is a validation and testing framework that can be adjusted to better suit a given model. The framework was created for an agent-based model for the spread of infectious diseases the primary purpose of which was explanation or description, and the framework is most applicable to other models designed for similar purposes. We believe, however, that our framework is a useful starting point for an agent-based modeller to help determine the appropriate process for model validation even if the model does not fit into the same categories as the models used as an example here. Determining the number of runs that should be done taking into account the variability in the model is essential in providing results that can be trusted and the method using the size of the confidence interval and the stability of the model results presented here can be used for many other types of agent-based models. While it might not be possible to present a multi-step validation process for every model, as we do here, using at least one of the steps should help provide some proof of model validation. Finally, a model should only be created if we can learn something from it to increase our knowledge of the system which is why the final step in our framework is testing the model. For instance, if the model is designed to predict a phenomenon first it should be shown that the predictions are valid and then the final step should show that we are able to learn more about the system using the model.

References

Apolloni A.; Kumar V.A.; Marathe M.V.; and Swarup S., 2009. *Computational Epidemiology in a Connected World. IEEE Computer Society*, 97–100. 10.1109/MC.2009.386. URL `http://staff.vbi.vt.edu/swarup/papers/computational-epidemiology.pdf`.

Edmonds B.; Le Page C.; Bithell M.; Chattoe-Brown E.; Grimm V.; Meyer R.; Montañola Sales C.; Ormerod P.; Root H.; and Squazzoni F., 2019. *Different Modelling Purposes. Journal of Artificial Societies and Social Simulation*, 22, no. 3, 6. ISSN 1460-7425. 10.18564/jasss.3993. URL `http://jasss.soc.surrey.ac.uk/22/3/6.html`.

Hunter E. and Kelleher J., 2020. *Using a Hybrid Agent-Based and Equation Based Model to Test School Closure Policies*. 10.21203/rs.3.rs-20889/v1. Preprint.

Hunter E.; Mac Namee B.; and Kelleher J., 2018. *An open-data-driven agent-based model to simulate infectious disease outbreaks. PLOS ONE*, 13, no. 12, 1–35. 10.1371/journal.pone.0208775. URL `https://doi.org/10.1371/journal.pone.0208775`.

Hunter E.; Mac Namee B.; and Kelleher J.D., 2020. *A Model for the Spread of Infectious Diseases in a Region. International Journal of Environmental Research and Public Health*, 17, no. 9, 3119. URL `https://doi.org/10.3390/ijerph17093119`.

Raimbault J.; Cottineau C.; Le Texier M.; Le Nechet F.; and Reuillon R., 2019. *Space Matters: Extending Sensitivity Analysis to Initial Spatial Conditions in Geosimulation Models. Journal of Artificial Societies and Social Simulation*, 22, no. 4, 10. ISSN 1460-7425. 10.18564/jasss.4136. URL `http://jasss.soc.surrey.ac.uk/22/4/10.html`.

Rakowski F.; Gruziel M.; Bieniasz−Krzywiec L.; and Radomski J.P., 2010a. *Influenza epidemic spread simulation for Poland — a large scale, individual based model study. Physica A: Statistical Mechanics and its Applications*, 389, no. 16, 3149 – 3165. 10.1016/j.physa.2010.04.029. URL `http://www.sciencedirect.com/science/article/pii/S0378437110003687`.

Richiardi M.; Leombruni R.; Saam N.J.; and Sonnessa M., 2006. *A Common Protocol for Agent-Based Social Simulation. Journal of Artificial Societies and Social Simulation*, 9, no. 1, 15. ISSN 1460-7425. URL `http://jasss.soc.surrey.ac.uk/9/1/15.html`.

Secchi D. and Seri R., 2017. *Controlling for 'false negatives' in agent-based models. A review of power analysis in organizational research. Computational and Mathematical Organization Theory*, 23, 94–121. 10.1007/s10588-016-9218-0.

Skvortsov A.T.; Connell R.B.; Dawson P.D.; and Gailis R.M., 2007. *Epidemic Modelling: Validation of Agent-based Simulation by Using Simple Mathematical Models. International Congress on Simulation and Modelling*, 657–662. URL `http://mssanz.org.au.previewdns.com/MODSIM07/papers/13_s20/EpidemicModeling_s20_Skvortsov_.pdf`.

Xia H.; Barrett C.; Chen J.; and Marathe M.V., 2013. *Computational Methods for Testing Adequacy and Quality of Massive Synthetic Proximity Social Networks. CSE '13 Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering*, 1113–1120. URL `http://staff.vbi.vt.edu/chenj/pub/NDSSL-TR-13-153.pdf`.