24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Sensitivity Analysis for Dimensionality Reduction in Agent-Based Modeling

Bianca Granato[1] and **Nicole Y.K. Li-Jessen**[2]

Abstract. Agent-Based Models (ABMs) can be used to numerically simulate highly non-linear phenomena that emerge from local interactions of multiple, independent entities. ABMs are often used to understand dynamic processes such as animal migration in ecology and pathogenesis in biomedicine, in which group-level patterns and space-time constraints are of interest. However, high fidelity ABMs, especially those in biomedical applications, usually have a large number of parameters, creating substantial uncertainty and high dimensionality at both local and global levels. Uncertainty analysis (output variance estimation) and sensitivity analysis are essential steps in investigating model robustness. In addition to allocating uncertainty to each parameter, sensitivity analysis can also be used to reduce ABM dimensionality for effective model calibration and optimization. We review common sensitivity analysis methods that have been used to decrease the number of parameters in complex ABMs, highlighting Garg et al. (2019)'s paper, where random forests – a non-parametric ensemble algorithm – are used as a sensitivity analysis method for ranking parameters in a biomedical ABM.

1 INTRODUCTION

Agent-Based Models (ABMs) are a computational approach that allows large scale interactions between individual objects or creatures to be simulated with a bottom-up design. ABMs consist of three elements: agents, rules, and world. *Agents* are individual entities that can move around and interact with other agents. Agents can behave stochastically, representing the probabilistic nature of the system of interest. *Rules* are usually encoded into ABM algorithms as mathematical equations or sequences of conditional statements. These rules describe, for example, how quickly or how often behaviours happen, what takes place during a behaviour, or agents' physical limitations. Finally, the *environment* acts as a boundary that represents the landscape of the virtual world. ABMs are suitable for characterizing dynamical systems where individual- or group-level responses are non-linear and can generate emergent behaviours. Agents in these systems may influence and be influenced by the environment, adding another level of complexity to the model [2]. Given their intricate nature, ABMs often have a great number of parameters with non-linear interactions between them, making uncertainty and sensitivity analyses key challenges for modellers in the field [1, 11].

For the purpose of this paper, *uncertainty analysis* is defined as the process through which the output variance and its confidence bounds are estimated. *Sensitivity analysis*, in turn, is the process of systematically changing model parameters to estimate their contribution to the overall model variance [8], which allows uncertainty to be reduced. By quantifying each parameter's contribution, it is plausible to improve the model's agreement with empirical data either by more carefully tuning specific parameter values or by guiding future experimental research on the simulated system. Uncertainty allocation can also give clues regarding causal mechanisms (see [9] for a discussion). Another major goal for sensitivity analysis, amd the primary focus of this paper, is dimensionality reduction. ABMs, especially those developed for high-fidelity biomedical applications, often have hundreds of parameters and non-convex solutions, making their calibration computationally expensive. As such, dimensionality reduction is imperative to reducing computational costs. Through sensitivity analysis, each parameter's contribution to the output can be assessed and the parameter set can be reduced to those with the highest impact on the overall result.

2 METHODS FOR SENSITIVITY ANALYSIS

Most common methods for sensitivity analysis are often not appropriate for ABMs due to their underlying assumptions and scope. *Local sensitivity analysis* informs about model behaviour at a specific baseline value, usually the point of best-fit. Common local sensitivity analysis tools are One Factor At a Time (OFAT), Fourier amplitude sensitivity testing (FAST), and the Morris method. In these methods, the value of each parameter is varied and they are ranked based on the effect on output variance. These methods are often easy to implement, and may be able to detect interactions and non-linearities in the modeled systems [8, 11]. Local sensitivity analysis can be applied to linear systems and can complement global sensitivity analyses, but can generate incomplete or incorrect results if applied to non-linear systems where parameter interactions are prevalent - such as ABMs.

Regression-based methods can be used for local or global sensitivity analysis. Once the model's uncertainty is calculated, it can be regressed on the parameters of interest, allowing the magnitude and direction of each factor's contribution to be quantified. Regression-based methods can also detect whether systems are non-additive or have non-linear interactions, though they operate under the assumption that the model is linear and that residuals are normally distributed, which is often not the case with ABMs [4, 7].

Most complex ABMs are non-linear and thus model-free methods for sensitivity analysis are often required. The Sobol method of variance decomposition has been a popular choice for non-linear, non-monotonic systems. In this method, the contribution of each parameter to the total model uncertainty is estimated through its first order sensitivity and total sensitivity indices. *First order sensitivity* is the variance obtained by keeping a parameter constant and varying all others, i.e. the main effect of a parameter. The *total sensitivity index*,

[1] McGill University, Canada, email: bianca.granato@mail.mcgill.ca
[2] McGill University, Canada, email: nicole.li@mcgill.ca

on the other hand, is the variance obtained by varying a parameter and keeping all others fixed, i.e. the effect of a parameter and its interactions with other parameters. The Sobol method is model-free and as such more flexible than linear methods. However, the assumption that the underlying parameters are independent would need to be met [11]. The high number of parameters in high-fidelity ABMs also make variance-based methods such as Sobol computationally expensive. As such, a non-linear method capable of handling large number of parameters, often times unknown, is highly desirable. One such method is Random Forests, which is the focus of the next section.

3 RANDOM FORESTS FOR SENSITIVITY ANALYSIS

Random forest is a non-linear method commonly used for classification and regression tasks [6]. In this method, decision trees are ensembled together to make an overall model that is more stable and robust than individual tree models. Each decision tree is applied to random subsets of the data obtained through bootstrapping. Aggregation also smooths the decision boundary, increasing the overall accuracy. To further increase model robustness, decision trees are built using an arbitrary subset of features which can be optimized with variable importance methods such as the Gini criterion. The Gini criterion can be thought of as a measure of importance of each parameter, similar to Sobol' variance decomposition [6, 10].

A disadvantage of using random forest-based sensitivity analysis is that the method overestimates the importance of correlated variables to the overall model. Strobl et al. (2008) demonstrate that this preference arises from early variable sub-setting, which only takes into account the marginal distribution of the variables and the data. When calculating variable importance, the Gini criterion uses unconditional permutation, further biasing importance measures in favour of correlated variables. However, with a large number of underlying decision trees and by carefully choosing the number of pre-selected variables for sub-setting [10], bias can be reduced and results are interpretable. Additionally, random forests are capable of handling large amounts of data and many unknown parameters, making it an ideal method for analyzing highly complex ABMs.

4 HIGHLIGHT: GARG ET AL. (2019)

Vocal folds are the voice organ in the human body which vibrate when we speak or sing. Vocal pathologies such as nodules and polyps may require surgery and, in some patients, iatrogenic scarring may arise following treatment. To predict the risk of vocal fold scarring, a series of computer models has been developed, namely, the Vocal Fold-ABMs (VF-ABMs) [5]. The most current VF-ABM operates at physiological level, with billions of agents at any given point. It is run on compute nodes two NVIDIA GPUs and 32 Intel CPUs using parallelization and high performance computing techniques [3]. The VF-ABM contains over 200 parameters, many of which whose values are unknown. To deal with the non-linearity, high dimensionality and large number of unknown parameters in the VF-ABM, Garg et al. (2019) implemented random forests and the Gini criterion to reduce the dimensionality of the VF-ABM.

Random forests were used to perform independent sensitivity analyses for the first four time points. Then, using the Gini criterion, the top three parameters for each cell at these time points was determined, reducing the number of parameters to be calibrated from 214 to 24. According to the authors, it would have taken seven years to run the 6 million iterations required for sensitivity analysis with traditional methods. By using random forests, however, they were able to reduce the number of iterations to only 5,000, making sensitivity analysis for the VF-ABM feasible. The reduced VF-ABM successfully estimated the distribution of all cells with reasonable accuracy, demonstrating how random forests are an effective and reliable method to analyze complex ABM.

5 CONCLUSION

This paper reviewed common methods for sensitivity analysis for ABM with a focus on biomedical applications, discussing the strengths and limitations of each method. Sensitivity analysis continues to be an open challenge for ABM developers given the high dimensionality, local nonlinearities, and global emergent behaviour inherent in this modeling approach. Garg et al. (2019) demonstrated how random forests can be used to rank the importance of parameters in a complex, physiological-scale ABM. Random forests have the advantage of being non-parametric and non-linear, require less computation time than many sensitivity analysis methods and can be used with sparse datasets where the value of many features is unknown. These characteristics make random forests good candidates for dimensionality reduction of complex ABM in biomedicine.

ACKNOWLEDGEMENTS

We would like to thank the reviewers for their constructive feedback and Joseph Szymborski for helpful discussions. This work was supported by the National Sciences and Engineering Research Council of Canada [RGPIN-2018-03843], the Canadian Institutes of Health Research [388583], *NSERC Canada Graduate Scholarship - Master's* and *FRQNT - Bourse de maîtrise en recherche*.

REFERENCES

[1] Mark Alber et al., 'Integrating machine learning and multi-scale modeling: Perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences', arXiv preprint arXiv:1910.01258, (2019).

[2] Eric Bonabeau, 'Agent-based modeling: Methods and techniques for simulating human systems', Proceedings of the National Academy of Sciences, **99**(suppl 3), 7280, (2002).

[3] Aman Garg et al., 'Towards a physiological scale of vocal fold agent-based models of surgical injury and repair: Sensitivity analysis, calibration and verification', Applied Sciences, **9**(15), 2974, (2019).

[4] Bertrand Iooss and Paul Lemaître, A review on global sensitivity analysis methods, 101–122, Springer, 2015.

[5] Nicole Yee-Key Li, Biosimulation of Vocal Fold Inflammation and Healing, Doctoral dissertation, University of Pittsburgh, 2009.

[6] Bjoern H. Menze et al., 'A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data', BMC Bioinformatics, **10**(1), 213, (2009).

[7] Saman Razavi and Hoshin V Gupta, 'What do we mean by sensitivity analysis? the need for comprehensive characterization of "global" sensitivity in e arth and e nvironmental systems models', Water Resources Research, **51**(5), 3070–3092, (2015).

[8] Andrea Saltelli and Paola Annoni, 'How to avoid a perfunctory sensitivity analysis', Environmental Modelling & Software, **25**(12), 1508–1517, (2010).

[9] Andrea Saltelli et al., 'Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices', Environmental Modelling & Software, **114**, 29–39, (2019).

[10] Carolin Strobl et al., 'Conditional variable importance for random forests', BMC Bioinformatics, **9**(1), 307, (2008).

[11] Guus ten Broeke, George van Voorn, and Arend Ligtenberg, 'Which sensitivity analysis method should i use for my agent-based model?', Journal of Artificial Societies and Social Simulation, **19**(1), 5, (2016).