The Importance of Input Uncertainty Quantification in Social Science Simulation

Bhakti Stephan Onggo and Lucy E. Morgan

Statistics and Operational Research Centre for Doctoral Training in Partnership with Industry
Lancaster University
Lancaster, LA1 4YR, UK

Abstract. Input uncertainty is a consequence of not knowing the true input distributions that drive a simulation model. It is often ignored when simulation outputs are reported. This paper argues that input uncertainty quantification should become a common practice in social science simulation, especially for models that will be used to support decision making.

Keywords. Agent-based simulation. Input uncertainty. Decision support

1 Introduction

Many stochastic simulation models require input distributions that are fitted using samples of real-world data. Since the number of samples is finite, a fitted distribution is almost surely not a perfect representation of the reality. The misspecification of an input distribution in a simulation model affects the quality of its output. Researchers have proposed methods to quantify the effect of input uncertainty on simulation outputs. These methods encompass: direct and bootstrap resampling; approximations based on Taylor's theorem; mean-variance meta-model approximations; Bayesian model averaging and a fully Bayesian approach. For a review of the existing methods see Barton (2012).

Arguably, collecting good quality data for a simulation model in social science is hard. We often need to collect difficult-to-measure data (e.g. trust and risk perception), difficult-to-find representative samples (especially when it is related to human behaviour) and subjective data (e.g. levels in a Likert scale may mean differently to different people). Hence, we argue that input uncertainty quantification should become a common practice among social simulation modellers. The need for input uncertainty quantification is even more important for models that will be used to support decision making.

This paper demonstrates the importance of input uncertainty quantification in social science simulation. As an example, we choose SugarScape (Epstein and Axtell 1996) because it is a widely known social science simulation model. We use the method proposed by Song and Nelson (2015) to quantify the input uncertainty because it is relatively easy to implement in the simulation tool that we use, i.e. Repast (North et al. 2013).

2 Experiments

In the SugarScape model, agents live in a two-dimensional world that is divided into $n \times n$ grids. Sugar grows in each grid. The initial amount of sugar varies between grids but is set deterministically at the start. The sugar grows at constant rate (1 unit per simulation step) until it reaches a pre-determined maximum amount. When agents are created, they are endowed with different amounts of sugar in their possession. They have varying levels of vision quality and metabolism rate. In each simulation step, every agent will move to a grid with the highest amount of sugar. Since an agent can only move to any unoccupied grid within the range of its vision, if there is more than one agent competing for the same grid the winner is chosen at random. Once an agent has moved to a new grid, it will harvest all sugar in the grid. In each simulation step, every agent also consumes some of its sugar depending on its metabolism rate. If the amount of sugar in its possession is not enough, the agent will die. Every agent will eventually die when it reaches its maximum age which is predetermined at its creation. When an agent dies, a new agent is created somewhere in the world to keep the figure of population constant but the sugar will not be passed on to the new agent. Hence, the SugarScape model has four input distributions: vision range, metabolism rate, initial sugar and maximum age. We are interested in estimating three measures of the population at the end of the simulation: average amount of sugar; mean vision and mean metabolism. The fist measure is an indicator of the population's wealth and the other two measures are an indicator of population's evolution (survival of the fittest). The simulation is run for 1,000 steps (approximately 14 generations).

Suppose there are 10,000 agents in the real SugarScape world and the correct distributions for vision range, metabolism rate, initial sugar and maximum age are 1+Poisson(2), 1+Poisson(1), 5+Poisson(6) and Poisson(70), respectively. In practice, we do not know what the true input distributions are. Hence, following the usual simulation modelling process, we first collect data from a sample from the population (suppose the sample size is 100). Next, we fit the data and run a simulation experiment to produce the confidence intervals as shown in Table 1. Song and Nelson (2015) refer to this as the nominal experiment. Table 1 illustrates a common way to report simulation outputs, by providing confidence intervals that quantify simulation sampling errors. But this ignores the fact that the input models are estimated from real world data and are subject to misspecification. Many papers have demonstrated that the error due to input uncertainty may overwhelm the simulation sampling error (see Song and Nelson (2015)).

Table 1. Outputs from nominal experiment (100 replications)

Output measures	Mean Sugar	Mean Vision	Mean Metabolism
Confidence interval	20.04 ± 0.1444	3.03 ± 0.0137	1.40 ± 0.0069

To measure the effect of input uncertainty, we use the method proposed in Song and Nelson (2015). This method expresses the impact of input uncertainty on the overall variance in the simulation output with the help of a mean-variance meta-

model approximation. The meta-model relates the mean of a simulation output to the means and variances of its input distributions. The meta-model is fitted using repeated bootstrap samples from the real-world data to simulate a number of possible samples that could be used to fit the input distributions in the model. A detailed explanation of the method is presented in Song and Nelson (2015).

Table 2 shows the contribution of each input distribution to the overall variance of each simulation output. The initial metabolism has the highest contribution to mean amount of sugar owned by the population and the average metabolism rate of the living population at the end of the simulation. The initial vision has the highest contribution to the average vision of the living population at the end of the simulation. This result shows that if we can put an extra effort in data collection, we should collect more data on these two inputs depending on which output that we are more interested in.

Table 2. Estimated contributions of inputs to outputs

Output	Initial vision	Initial Metabolism	Initial sugar	Max age	Overall input uncertainty	$\hat{\gamma}$
Mean Sugar	0.0980	0.2142	0.0125	0.0286	0.3533	0.0651
Mean Vision	0.0193	2.47×10^{-5}	1.11×10^{-6}	2.70×10^{-6}	0.0193	0.3922
Mean Metabolism	9.04×10^{-6}	0.0016	2.81×10^{-5}	8.33×10^{-6}	0.0016	0.1291

The overall input uncertainty in Table 2 measures the total contributions of all input distributions which can be expressed in the unit of the simulation sampling error as a ratio $\hat{\gamma}$. For example, the input uncertainty for mean vision is 39% of the simulation sampling error. The confidence intervals in Table 1 can be adjusted by $\sqrt{(1+\hat{\gamma}^2)}$ to include the input uncertainty as shown in Table 3. In this instance, the level of input uncertainty may be considered acceptable. But, suppose the level of input uncertainty was unacceptable; the results in Table 2 could then guide us as to which input distributions are most sensitive and where further data collection would be most beneficial.

Table 3. Adjusted confidence intervals for simulation outputs

Output measures	Mean Sugar	Mean Vision	Mean Metabolism
Confidence interval	20.04 ± 0.1447	3.03 ± 0.0147	1.40 ± 0.0070

Since we know the true distributions and their parameters, we can compute the contribution of every distribution to the input uncertainty in each simulation output using the following algorithm (from Song and Nelson 2014).

```
1. For each input distribution l {
2.   For (b=0; b<B; b++) {
3.     Generate m samples from correct distribution l
```

```
4.      Use the m samples to fit distribution F(l)
5.      Run R replications using fitted F(l) and correct
          distributions for other input distributions
6.    }
7.    Contribution(l) = sample variance of simulation output
8. }
```

Table 4 shows the contributions of all four input distributions to all three simulation outputs. This result can be used to validate the result in Table 2 The dominant sources of input uncertainty identified in Table 2 are consistent with the result in Table 4.

Table 4. Estimated contributions of inputs to outputs

Output	Initial vision	Initial Metabolism	Initial sugar	Max age
Mean Sugar	0.0241	0.0867	0.0400	0.0067
Mean Vision	0.0197	9.99×10^{6}	1.10×10^{5}	1.99×10^{5}
Mean Metabolism	7.06×10^{6}	0.0039	1.67×10^{5}	4.10×10^{6}

3 Conclusion

In this paper, we have shown how input uncertainty in the SugarScape model can be quantified with the help of a metamodel. It further shows how input uncertainty is used to adjust the confidence intervals of simulation outputs. In the given example, the level of input uncertainty is low. However, if the level of input uncertainty is high, decision makers can be at risk of making incorrect decisions. Hence, this paper argues that input uncertainty quantification should become a common practice, especially in social simulation where good quality input data is an issue.

Acknowledgement

We gratefully acknowledge the support of the EPSRC funded EP/L015692/1 STOR-i Centre for Doctoral Training. We also thank Barry Nelson for his helpful discussion.

References

Barton, R. R. (2012). Tutorial: Input uncertainty in outout analysis. In Simulation Conference (WSC), Proceedings of the 2012 Winter, pages 1-12. IEEE.

Epstein, J.M. and Axtell, R. (1996) Growing artificial societies: social science from the bottom up. Washington, D.C.: Brookings Institution Press.

North, M. J., N. T. Collier, J. Ozik, E. R. Tatara, C. M. Macal, M. Bragen, and P. Sydelko. (2013) Complex Adaptive Systems Modeling with Repast Simphony. Complex Adaptive Systems Modeling, 1(1): 3. doi:10.1186/2194-3206-1-3

Song, E. and Nelson, B.L. (2015) Quickly Assessing Contributions to Input Uncertainty. IIE Transactions, 47:1–17.