Ecological Modelling 220 (2009) 2594–2597

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Short communication

# The power of simulating experiments

Katrin M. Meyer[a,*], Wolf M. Mooij[b], Matthijs Vos[c,1], W.H. Gera Hol[d], Wim H. van der Putten[a,e]

[a] Netherlands Institute of Ecology NIOO-KNAW, Centre for Terrestrial Ecology, Department of Multitrophic Interactions, Boterhoeksestraat 48, 6666 GA Heteren, The Netherlands
[b] Netherlands Institute of Ecology NIOO-KNAW, Centre for Limnology, Department of Food Web Studies, Rijkstraatweg 6, 3631 AC Nieuwersluis, The Netherlands
[c] Netherlands Institute of Ecology NIOO-KNAW, Centre for Estuarine and Marine Ecology, Department of Ecosystem Studies, Korringaweg 7, 4401 NT Yerseke, The Netherlands
[d] Netherlands Institute of Ecology NIOO-KNAW, Centre for Terrestrial Ecology, Department of Terrestrial Microbial Ecology, Boterhoeksestraat 48, 6666 GA Heteren, The Netherlands
[e] Wageningen University, Laboratory of Nematology, Binnenhaven 5, 6709 PD Wageningen, The Netherlands

ARTICLE INFO

*Article history:*
Received 10 April 2009
Received in revised form 2 June 2009
Accepted 4 June 2009
Available online 3 July 2009

*Keywords:*
Sample size
Replication
Individual-based simulation model
Experimental design
Statistical power
Stochasticity

ABSTRACT

Addressing complex ecological research questions often requires complex empirical experiments. However, due to the logistic constraints of empirical studies there is a trade-off between the complexity of experimental designs and sample size. Here, we explore if the simulation of complex ecological experiments including stochasticity-induced variation can aid in alleviating the sample size limitation of empirical studies. One area where sample size limitations constrain empirical approaches is in studies of the above- and belowground controls of trophic structure. Based on a rule- and individual-based simulation model on the effect of above- and belowground herbivores and their enemies on plant biomass, we evaluate the reliability of biomass estimates, the probability of experimental failure in terms of missing values, and the statistical power of biomass comparisons for a range of sample sizes. As expected, we observed superior performance of setups with sample sizes typical of simulations ($n = 1000$) as compared to empirical experiments ($n = 10$). At low sample sizes, simulated standard errors were smaller than expected from statistical theory, indicating that stochastic simulation models may be required in those cases where it is not possible to perform pilot studies for determining sample sizes. To avoid experimental failure, a sample size of $n = 30$ was required. In conclusion, we propose that the standard tool box of any ecologist should comprise a combination of simulation and empirical approaches to benefit from the realism of empirical experiments as well as the statistical power of simulations.

## 1. Introduction

In order to test hypotheses and theories, ecology fundamentally relies on manipulative experiments ranging from controlled conditions, such as in the lab or in greenhouses, to conditions as uncontrolled as in the field (Scheiner, 2001). However, the increasing complexity of ecological questions over the past decades, e.g. in the areas of multitrophic interactions and above–belowground ecology, has challenged the logistic constraints of empirical experiments (Peck, 2008). These constraints usually result in a trade-off between experimental complexity in terms of number of treatment combinations and sample size at the treatment level that empiricists can deal with. Simulating experiments in the computer could solve at least a part of this dilemma (Peck, 2004, 2008). Simulations are also subject to constraints due to computing power limitations, but these allow much greater complexity-sample size combinations than the logistic constraints of empirical experiments. The basis for simulated experiments is rule- or code-based models that recreate the relevant features of an ecological system (Peck, 2008), often considering individuals, time and/or space explicitly (Grimm et al., 2005). In principle, the experimental design of simulated experiments parallels that of empirical experiments, e.g. ANOVA-type designs. Yet, assuming that all model parameters can be estimated from empirical data or expert knowledge, simulated experiments can deal with a greater number of factorial treatment combinations while retaining greater numbers of replicates.

The major aim of the present study was to explore whether the effort of developing a simulation model is justified when at the same time an empirical experiment of the same complexity could be performed, albeit with smaller sample size. One derived aim was to investigate the relationship between sample size and experimental failure, referring to those cases where all plants had died at the end of all experimental replicates generating only missing values. The second derived aim was to determine how the relationship between simulated standard errors and sample size deviates

[*] Current address: University of Göttingen, Ecosystem Modelling, Büsgenweg 4, 37077 Göttingen, Germany. Tel.: +49 0551393795.
*E-mail address:* kmeyer5@uni-goettingen.de (K.M. Meyer).
[1] Corresponding author. Current address: University of Potsdam, Institute of Biochemistry and Biology, Department of Ecology & Ecosystem Modeling, Am Neuen Palais 10, 14469 Potsdam, Germany.

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.06.001

from statistical expectations. To reach these aims, we compared the output reliability and the statistical power of simulated experiments with numbers of replicates typical of computer simulations and empirical setups, respectively. This comparison was entirely based on simulation results from the rule-based ABove–BElowground interactions model ABBE (Meyer et al., 2009). ABBE has been developed to assess the relative impacts of herbivores and higher trophic levels on plant shoot and root biomass in above- and belowground food webs. ABBE includes environmental, demographic, and individual stochasticity (Table 1). Therefore, output values vary among replicates reflecting natural variation. The magnitude of this stochasticity-induced variation of biomass values in the model matched observed levels of variation (see formal validation in Meyer et al., 2009). This variation gave us the opportunity to compare the impact of different sample sizes on the reliability of plant biomass estimates derived from the model output. By focusing on measures of variability in model output, our approach goes beyond many model analyses that consider only averages.

## 2. Methods

### 2.1. Simulating experiments with ABBE

ABBE considers aboveground trophic levels at the individual level and belowground trophic levels at the population level, using empirical data from Soler et al. (2005) for parameterization and successful validation (Meyer et al., 2009). The greenhouse experiment of Soler et al. (2005) consisted of potted plants with and without root and shoot herbivores as well as parasitoids and hyperparasitoids aboveground in a crossed design. The level of replication was ten plants per treatment combination. In our simulation model, we also designed a factorial experiment with presence and absence of all trophic levels as treatments and shoot biomass as the response. Understanding the determinants of aboveground plant biomass is of great importance in both fundamental ecology (e.g. diversity-function relationships) and more applied agro-ecology (e.g. biological control). All ecologically relevant treatment combinations were considered, excluding combinations where higher trophic levels were present without the underlying trophic level(s) present, yielding 24 combinations in total. We used non-parametric Mann–Whitney *U*-tests for the analysis of treatment effects since the assumptions of ANOVA were violated.

**Table 1**
Sources of stochasticity in the simulation model ABBE.

| Model parameters |
|---|
| *Environmental stochasticity*[a] |
| Extractable proportion of nutrients |
| *Demographic stochasticity*[b] |
| Mortalities of shoot herbivores, parasitoids, hyperparasitoids, root herbivores and antagonists |
| Parasitism and hyperparasitism success probabilities |
| Reproduction probabilities of shoot herbivores and root herbivore antagonists |
| Proportion of female shoot herbivores |
| Egg viability of root herbivores |
| *Individual stochasticity*[a] |
| Initial body mass of shoot herbivores |
| Initial body mass of parasitoids |
| Initial body mass of hyperparasitoids |

[a] Implemented by drawing a random number from a Gaussian distribution with mean and standard deviation given by the parameter values.

[b] Implemented by comparing a random number drawn from a uniform distribution with the corresponding parameter value; for parameter values see Meyer et al. (2009).

### 2.2. Replication analysis

We first determined the standard error of biomass values and the proportion of successful experiments at different levels of replication. Then, the statistical power of our biomass comparisons at different levels of replication was analysed. We focused on numbers of replicates typical of greenhouse experiments as compared to simulated experiments (about 10 and 1000, respectively). To determine the standard error of biomass values, we recorded shoot biomass at the end of each of a target number of replicate runs, ranging from 2 to 50,000 runs. Then, we calculated the average shoot biomass over the respective number of target replicate runs. We repeated this procedure 100 times for each target number of replicate runs and determined the standard deviation of the averages over these 100 iterations. We dealt with missing values caused by plant death in all target replicate runs by excluding the respective iteration(s). The standard deviation of averages defines the standard error of the mean $\mathrm{SE} = s * n^{-0.5}$, where $s$ is the standard deviation of one set of replicates and $n$ is the number of replicates (Sokal and Rohlf, 1995). We fitted this relationship to our data to identify what our stochastic model can add to a statistical calculation of standard errors that is solely based on the envisaged number of replicates and standard deviation of a study. This also enabled us to determine the approximate minimum number of replicates that keeps the standard error as low as possible. To evaluate experimental failure, we determined its inverse, the proportion of successful experiments, where the plant survived until the end of the experiment in at least one of the replicates, for different numbers of replicates ranging from 2 to 50,000.

Then, we used the treatments with and without aboveground parasitoids as an example to analyze the statistical power of the Mann–Whitney *U*-test of biomass differences at different levels of replication. We used the parasitoid comparison to obtain a balanced design with equal numbers of treatment combinations in both groups. The power calculation included the number of replicate runs, the corresponding standard deviation of shoot biomass, the significance level of 0.05, and the desired effect size $\delta$ to be detected. The effect size represented the difference in average shoot biomass with and without parasitoids. We evaluated numbers of replicate runs ranging from 2 to 10,000, where a power of 1 was reached. We tested effect sizes between 0.5 and 4.0, including the effect size of 2.43 g (shoot biomass) emerging from the 1000 replicate runs in Meyer et al. (2009). We also determined those combinations of effect size and numbers of replicates that are required for a power of more than 0.8, which is the generally promoted threshold for sufficient power (Crawley, 2007). Due to different plant mortality with and without parasitoids, there were different numbers of replicates for presence and absence of parasitoids. We always used the smaller number of replicates to obtain a conservative power estimate. We applied the power analysis procedure for *t*-tests and adapted it to the Mann–Whitney *U*-test used here by dividing the number of replicates by 0.955 (Lehmann, 1975).

## 3. Results

The modelled relationship between standard error and the number of replicates followed the statistically expected negatively asymptotic relationship with an $R^2$ of 0.996 when the first nine data points were excluded (solid points and dashed line in Fig. 1). The first nine data points were lower than expected from statistical theory. For 10 replicates, the standard error was about 1.5 g, which is close to 10% of the average shoot biomass, and 10% of the experiments failed (Fig. 1). Up to 100 replicates, a small increase in number of replicates yielded a great reduction in standard error, asymptotically approaching 0 g at about 1000 replicates (Fig. 1).

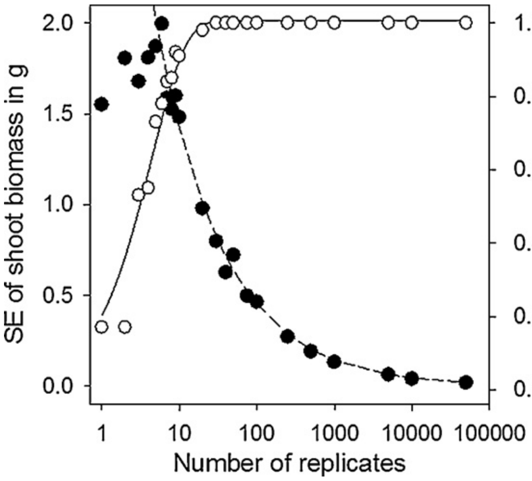

**Fig. 1.** The relationship between the standard error of shoot biomass (SE, calculated as the standard deviation over 100 iterations of the averages over different numbers of replicate simulation runs) and the number of replicates $n$ and relationship for the respective proportion of successful experiments (open circles). Experimental success was defined as survival of the plant until the end of the experiment in at least one of the replicates. Shoot biomass averages did not include zeros. Note the logarithmic scale of the $x$-axis. Regression equations: $\mathrm{SD} = 4.5098 * n^{-0.5}$, $R^2 = 0.996$ (excluding the first nine $n$ values; dashed line); Success probability = $1.0056 * (1 - e^{-0.2241 * n})$, $R^2 = 0.96$ (solid line).

When using 30 or more replicates, 100% of the experiments were successful (Fig. 1).

The relationship between the statistical power of the comparison of shoot mass with and without aboveground parasitoids and effect size and number of replicate runs was positive, as expected (Fig. 2). To exceed the threshold of 0.8 for statistical power at any tested effect size, 1000 replicates were sufficient. With 10 replicates, only an effect size greater than 3.79 g (or about 27% of the average shoot biomass) can be detected with a power of at least 0.8. The effect size in Meyer et al. (2009) of 2.43 g (about 18%) would have been detected with a power of 1 with 1000 replicates and with a power of 0.43 with 10 replicates.

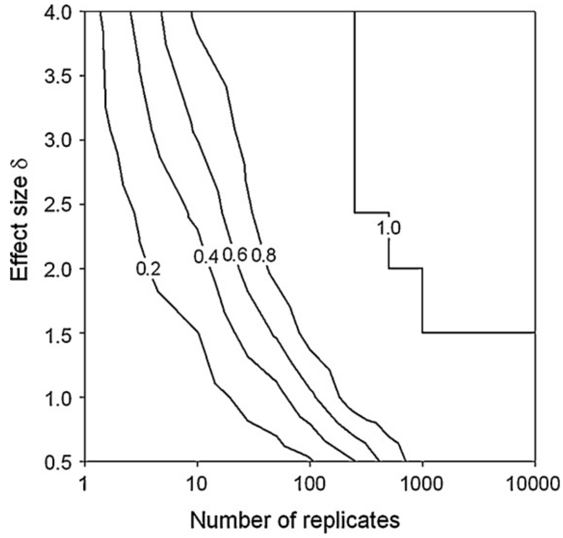

**Fig. 2.** Relationship between the statistical power of a Mann–Whitney $U$-test (contour labels), the difference in mean shoot biomass with and without aboveground parasitoids (effect size $\delta$), and the number of simulation runs of the model ABBE per treatment combination (number of replicates). The power was calculated for the significance level $\alpha = 0.05$ and the standard deviation of shoot biomass over all replicates within every simulated level of replication (1, 10, 50, 100, 250, 500, 1000, 10,000). There were different number of replicates for presence and absence of parasitoids because only simulation runs with surviving plants were included in the biomass record. We used the respective smaller number of replicates at a level of replication in the power calculation. We present here the original number of simulation runs on the $x$-axis because selecting the appropriate number of replicates must also take plant mortality in the course of the experiment into account. At a level of replication of 1000, the power is always greater than 0.8, while with 10 replicates, only an effect size greater than 3.79 can be detected with a power of at least 0.8. Note the logarithmic scale of the $x$-axis.

## 4. Discussion

We show that simulation of experiments presents a powerful research approach. It offers the opportunity to raise the sample size beyond the logistic limits of greenhouse experiments while maintaining natural levels of stochasticity. The complexity (24 treatment combinations) and replication (1000 runs) of a typical simulation experiment correspond to an experiment with 24,000 pots in the greenhouse. Replication at the level of $n = 10$ is typical (if not overestimated) for greenhouse experiments, but 10% of the cases failed and the power was far from the generally accepted minimum threshold of 0.8 (Crawley, 2007). With 30 replicates or more, there were no experimental failures anymore, with 100 replicates or more, the standard error was asymptotically approaching 0, and with more than 1000 replicates, experimental power was always greater than 0.8. While 30 replicates may be feasible in practical experiments by reducing the number of treatments, 1000 replicates will not be realistic.

Part of our conclusion could have been reached from applying statistical theory alone as the almost perfect fit of the standard error relationship for sample sizes greater than ten exemplified (Fig. 1). However, the standard error of simulated data was much lower than expected from statistics for small replicate numbers, highlighting the crucial role of model-inherent stochasticity (cf. Table 1). Hence, at small sample sizes, it may not be sufficient to estimate standard deviation from expert knowledge and apply statistics to derive the minimum sample size. Rather, empirical pilot studies or stochastic simulation models should be used in these cases to obtain realistic estimates of required sample sizes. Especially for complex experimental designs, simulated experiments have much to offer to ecologists. These allow exploration of the complete design and can identify crucial treatment combinations on which less complex empirical experiments can focus subsequently.

Simulated experiments will always depend on empirical works for conceptualization, parameterization, and validation (Peck, 2004). This is particularly true for the realistic estimation of the variance of parameter values, which ultimately determines statistical power and reliability. Therefore, we do not advocate the replacement of the established practice of empirical experimentation in ecology by a less constrained simulation approach. Rather, in line with Peck (2008) and van der Putten et al. (2009), our results encourage a close coupling between empirical and simulated experiments as an integral part of the standard tool box of teams of ecologists. Probably the most promising way to efficiently couple models and empirical work is pattern-oriented modeling (Grimm et al., 2005), where model output is validated against multiple empirically observed patterns simultaneously.

We see major implications for the generality of conclusions drawn from simulations. Simulated systems will produce results of at least the same generality as attributed to the greenhouse experiment they are derived from. This can only be exceeded when the parameter space of the model is explored beyond the standard parameterization, for instance during a sensitivity analysis or when applying the model to a new system. We conclude that simulated and empirical experiments should more routinely be combined in ecological studies to benefit from the best of both worlds: the inherent realism of working with natural organisms and the statistical power of simulations.

## Acknowledgements

We thank the members of the Multitrophic Interactions Department for fruitful discussions during model conceptualization and the ALW-biodiversity program of the NWO (Project ALWPB/05-02) for funding. This is NIOO publication No.4570.

## References

Crawley, M.J., 2007. The R Book. Wiley.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. Science 310, 987–991.

Lehmann, E.L., 1975. Nonparametrics. Statistical Methods Based on Ranks. Holden-Day.

Meyer, K.M., Vos, M., Mooij, W.M., Hol, W.H.G., Termorshuizen, A.J., Vet, L.E.M., van der Putten, W.H., 2009. Quantifying the impact of above- and belowground higher trophic levels on plant and herbivore performance by modeling. Oikos 118, 981–990.

Peck, S.L., 2004. Simulation as experiment: a philosophical reassessment for biological modeling. Trends Ecol. Evol. 19, 530–534.

Peck, S.L., 2008. The hermeneutics of ecological simulation. Biol. Philos. 23, 383–402.

Scheiner, S.M., 2001. Theories, hypotheses, and statistics. In: Scheiner, S.M., Gurevitch, J. (Eds.), Design and Analysis of Ecological Experiments. Oxford University Press, pp. 3–13.

Sokal, R.R., Rohlf, F.J., 1995. Biometry: The Principles and Practice of Statistics in Biological Research. W.H. Freeman and Co.

Soler, R., Bezemer, T.M., Van der Putten, W.H., Vet, L.E.M., Harvey, J.A., 2005. Root herbivore effects on above-ground herbivore, parasitoid and hyperparasitoid performance via changes in plant quality. J. Anim. Ecol. 74, 1121–1130.

van der Putten, W.H., Bardgett, R.D., de Ruiter, P.C., Hol, W.H.G., Meyer, K.M., Bezemer, T.M., Bradford, M.A., Christensen, S., Eppinga, M.B., Fukami, T., Hemerik, L., Molofsky, J., Schädler, M., Scherber, C., Strauss, S.Y., Vos, M., Wardle, D.A., 2009.Empirical and theoretical challenges in aboveground-belowground ecology. Oecologia.doi 10.1007/s00442-009-1351-8.