Ecological Modelling xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Ecological Modelling

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

Machine learning meets individual-based modelling: Self-organising feature maps for the analysis of below-ground competition among plants

Ronny Peters[a,*], Yue Lin[b], Uta Berger[a]

[a] Institute of Forest Growth and Computer Science, TU Dresden, P.O. 1117, 01735 Tharandt, Germany
[b] School of Life Science, Northwest University, 710069 Xi'an, China

A R T I C L E I N F O

Article history:
Received 25 February 2015
Received in revised form 15 October 2015
Accepted 20 October 2015
Available online xxx

Keywords:
Pattern recognition
Pi model
Plant–plant interactions
Pattern-oriented modelling
Inverse modelling
Artificial neural networks

A B S T R A C T

Individual-based models (IBM) simulate populations and communities whose dynamics are shaped by the properties, interactions and behaviour of the constituent organisms as well as the corresponding abiotic boundary conditions. Structurally realistic IBM can provide insights into the functioning of such systems and predict the effects of variable scenarios. We suggest complementing IBM with machine learning (ML) methods in order (i) to visualise correlation patterns between model inputs and model outputs, (ii) to provide simulation-based decision tools for non-modellers, and (iii) to derive information about factors difficult to obtain in the field on the basis of data that are more readily measurable. On top of this, ML methods can complement the established pattern-oriented modelling approach used to analyse the behaviour of IBM and to detect model uncertainties. As an example to demonstrate the strength of an IBM-ML connection, we combined the individual-based Plant Interaction Model (*Pi* model) with self-organising feature maps (SOM) – a special type of ML. Based on simulation experiments with complete knowledge of the simulated system, the SOM was trained and used to visualise the nonlinear relationship between two IBM inputs (namely the mode of below-ground competition and below-ground resource limitation) and two model outputs (the mortality rate and the Clark Evans Index of the spatial distribution of plants). Our study also highlights an application of the SOM to infer the modes of below-ground competition (either symmetric or asymmetric) from the remaining measurable variables (resource limitation, mortality rate and Clark Evans Index). This procedure was successful in 92% of cases, revealing its great potential as a means to assess parameters difficult to measure in nature. This example shows that SOM are powerful tools to revert the hierarchy of variables and to generalise dependencies of parameters in individual based modelling.

1. Introduction

In individual-based models, complex interactions between organisms occurring over multiple time steps reveal emergent properties of populations and communities that cannot be predicted based on the averaged characteristics of the organisms. The determination of properties across different hierarchical levels cannot be achieved by a simple reverse calculation. In experimental ecology, however, the population properties are often easier to measure than the whole set of model assumptions on the level of the individual. The pattern-oriented modelling (POM) approach was developed to overcome this problem (e.g., Wiegand et al., 2003; Piou et al., 2009; Grimm and Railsback, 2011 and references within). Rigorously applied in all steps of the modelling cycle (Grimm and Railsback, 2005), POM links patterns occurring at individual level with patterns observed at the system level, and vice versa (see the techniques of inverse modelling introduced to IBM by Wiegand et al., 2003; Grimm et al., 2005 and others).

POM supports the definition of the right entities and scales of an IBM, the selection of the most reasonable sub-models, and the calibration of the parameters (e.g., Wiegand et al., 2003; Piou et al., 2009; Grimm and Railsback, 2011; Jakoby et al., 2014). Patterns are used (i) to determine which scales, entities and variables a model needs, (ii) to test contrasting theories about the functioning of a system by selecting the most reasonable sub-models representing key processes, and (iii) to find suitable parameter sets through calibration. A strict adherence to all modelling steps leads to 'structurally realistic' models with predictive power even for changing

* Corresponding author. Tel.: +49 351 46331625; fax: +49 351 46331632.
E-mail address: ronny.peters@tu-dresden.de (R. Peters).

http://dx.doi.org/10.1016/j.ecolmodel.2015.10.014

environmental conditions (Berger et al., 2008; Lin et al., 2012, 2013). Structural realism is achieved when the objects, variables and processes described in the model correspond to the internal organisation of the real system so that changes in the environment result in realistic responses at the corresponding hierarchical levels.

The contribution of IBM to applied ecology is particularly strong where predictions obtained through simulation experiments can directly guide management (e.g., Railsback et al., 2002; Stillman et al., 2015) or where driving factors that are unsuited to direct measurement in the field can be quantified (e.g., Berger et al., 2004, 2006; Meyer et al., 2014). The former is limited by the fact that separate IBM cannot be developed for each particular scenario and location. The latter on the other hand requires long-term experience in developing and using IBM, limiting the group of ecologists who can take advantage of this approach.

In IBM, the emerging, higher-level response variables result from the low-level parameters on the individual scale. To map the reverse relationship so as to draw conclusions regarding the causes of observed higher-level patterns, a direct functional shortcut between model input and model output is necessary. This possibility would be valuable to further strengthen the link between IBM research and applied ecology. In the following, we use the term 'reverse modelling' to denote this approach and to distinguish from the 'inverse modelling' used in the POM context, for example, by Wiegand et al. (2003). Inverse modelling, unlike reverse modelling, is a largely iterative procedure. We suggest that methods of machine learning could prove useful as they provide several properties suitable for this purpose, as described in the following.

Artificial neural networks (ANN) were first developed by neuroscientists to describe human neural reactions. Rosenblatt (1958) developed the Perceptron network (also known as multilayer feed-forward network) for pattern recognition. In recent decades, the Perceptron network has become the most widely employed in applied sciences. Most of the applications cited below refer to multilayer Perceptron networks. A completely new type of ANN – the self-organising map (SOM) or Kohonen network – was introduced by Teuvo Kohonen (1982). In addition to these milestones in the development of ANN, several other such models exist, as presented in Hagan et al. (1996).

Since the broad entry of computers and simulation models into the applied sciences, ANN have been employed in a wide range of applications as flexible empirical modelling tools. One of the main features of ANN – function approximation – made them quite useful for various purposes. Data-driven approaches help to detect functional relations between measured input and output variables (e.g., Simpson et al., 1992, for plankton identification).

The application of ANN is most powerful when they are used as surrogate models; that is, ANN trained using datasets obtained by mechanistic models. The advantage of such ANN is that they can memorise system behaviour beyond the observations made in the field, because they benefit from the structural realism in the mechanistic model. By reducing the complexity to only the variables relevant to the researcher and simplifying their linkage to just a single functional relation, they are able to learn a wide range of model behaviour and at the same time reduce the computation costs of application. ANN have found widespread implementation as surrogate models used to substitute sophisticated mechanistic models based on differential equations for nonlinear temporal processes; for example, in water quality modelling for standing water bodies (e.g., Petzoldt et al., 2003), rainfall-runoff models (e.g., Cullmann, 2007), hydrodynamic flow models (e.g., Solomatine and Avila Torres, 1996; Peters et al., 2006) and as ecological simulation models used, for instance, to predict the propagation of green algae in the Mediterranean Sea (Aussem and Hill, 1999).

Lek and Guégan (1999) provided an introduction to the use of the multilayer Perceptron networks and Kohonen networks (or SOM) as tools in ecological modelling. Kalteh et al. (2008) provided a review of modelling applications with SOM in water resources. Zhang (2010) presented a good overview of ANN focussing on both a theoretical explanation of ANN types and on ecological applications. More recent case studies demonstrated the broad potential of such methods for ecological research; for example, Watts et al. (2011) used ANN to predict the probability of reef occurrence as a function of bathymetric and slope variables; Millie et al. (2012) trained ANN to predict micro-algae abundance; Kulhanek et al. (2011) set up ANN to predict carp distribution from limnological and climatic variables for lakes in Minnesota.

In some approaches, multilayer Perceptrons were used as sub-modules in IBM to control the movement of agents. Huse and Giske (1998) applied ANN to model fish migration on a daily basis in an artificial environment. Dreyfus-Leon (1999) implemented ANN as sub-modules within a spatially explicit IBM to mimic the search behaviour of fishermen.

However, to the best of our knowledge, the potential offered by ANN as tools to carry out reverse modelling is not being utilised in ecology as yet. ANN may complement IBM in a new way, helping to find a link from the emerging patterns back to invisible (and only assumed) properties of the individuals or their interactions. ANN can be used as a reverse surrogate-models for IBM in situations where we can measure and simulate variables that result from a particular process (high level parameters), but where the underlying (low-level) parameters are not available. Multiple runs with the IBM would provide a comprehensive database containing all possible reasons (model inputs) for combinations of model outputs. The trained SOM then memorises the relationships between the variables for the whole parameter space. A prediction of the underlying low-level properties is subsequently possible with a single step. In this article, we will demonstrate this for a particular example.

There is a multitude of statistical (e.g., randomForest, Pearson et al., 2014) and ML approaches (see above) capable of describing functional relations and predicting variables. We selected self-organising feature maps for the purposes of this study. SOM is not necessarily superior to other modelling techniques, but we made this choice because SOM combines some useful properties for the exploration of functional links and prediction purposes:

- SOM is trained unsupervised (this will be explained in more detail in the methods section); i.e., there is no predefined response variable;
- a visualisation of the SOM provides an easy to read summary of a multidimensional correlation;
- SOM can predict any response variable even with incomplete input datasets.

Giraudel and Lek (2001) compared SOM to a selection of statistical approaches to predict abundances of tree species. They stated that SOM provides a visual way to find structures in ecological communities and so facilitates the discovery of unexpected structures within datasets. The focus of this paper is not a comparison with other techniques derived from either ML or classical statistics. The aim was restricted to the following specific research goals:

(1) To develop a SOM to visualise the complex correlation between the input and output variables of an IBM describing plant population dynamics.
(2) To test the potential of the trained SOM to identify unknown modes of below-ground competition based on variables that are easier to measure in real systems, such as below-ground resource limitation, which we use as a proxy variable for nutrient or water availability (e.g., soil moisture), or that are