Ecological Modelling 342 (2016) 113–122

Contents lists available at ScienceDirect

Ecological Modelling

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

Self-organizing maps for analysing pest profiles: Sensitivity analysis of weights and ranks

Mariona Roigé*, Matthew Parry, Craig Phillips, Susan Worner

BPRC, Lincoln University, New Zealand

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form
20 September 2016
Accepted 2 October 2016

Keywords:
Self-organizing maps
Pest profile analysis
Clustering
Prioritization
Invasive pest assemblages

ABSTRACT

Self organizing maps for pest profile analysis (SOM PPA) is a quantitative filtering tool aimed to assist pest risk analysis. The main SOM PPA outputs used by risk analysts are species weights and species ranks. We investigated the sensitivity of SOM PPA to changes in input data. Variations in SOM PPA species weights and ranks were examined by creating datasets of different sizes and running numerous SOM PPA analyses. The results showed that species ranks are much less influenced by variations in dataset size than species weights. The results showed SOM PPA should be suitable for studying small datasets restricted to only a few species. Also, the results indicated that minor data pre-processing is needed before analyses, which has the dual benefits of reducing analysis time and modeller-induced bias.

1. Introduction

Over recent decades there has been considerable research on biological invasions and their impacts (Barlow and Goldson, 2002; Blackburn et al., 2014; Hulme, 2003; McGeoch et al., 2006). Such interest has caused invasion ecology to become a multidisciplinary field, bringing together fundamental ecology, conservation, environmental management, border control and biosecurity (Kolar and Lodge, 2001; Perrings et al., 2005; Vitousek, 1990). Despite its diversity, there is consensus about the need to develop proactive invasion prevention strategies rather than reactive pest management programmes.

An important tool for preventing invasions is pest risk analysis, which draws together several sub-disciplines of quantitative and qualitative science. In most developed countries, biosecurity and quarantine agencies use pest risk analysis to help make decisions about which species and entry pathways to regulate (EPPO, 2004; FAO, 2006; Leung et al., 2012).

Self-organizing maps for pest profile analysis (SOM PPA) is a quantitative method intended to assist pest risk analysis, which was first described by Worner and Gevrey (2006). A pest profile is the assemblage of insect pest species in a region, and a SOM is an artificial neural network algorithm that performs unsupervised classification (Kohonen, 1982). In SOM PPA, pest profiles for all geopolitical regions of the world are collected and their similarity is analysed. Regional profiles clustered together are assumed to share similar biotic and abiotic conditions that have allowed their respective species assemblages to become established. The output of SOM PPA is a list of species ranked according to the level of the risk they present to the region under consideration. A species that is present in many of the regions which cluster with the target region but is absent for the target region, could establish in the target region if introduced. The level of risk is indicated by SOM species weights, which are explained below.

Due to the algorithmic nature of SOM, the validity of its output depends on the quality of the input data. Species occurrence databases that contain records at a global scale inevitably include errors, which may invalidate the SOM PPA. Previous research has investigated the sensitivity of the method to certain data problems: first, Paini et al. (2010a) measured the method's sensitivity to data errors (presences recorded as absences and vice versa) and demonstrated that SOM PPA is insensitive to errors in the data up to 20%. Paini et al. (2010b) showed the predictive value of SOM PPA when applied to a simulated dataset.

Nevertheless, issues about using SOM PPA remain (Worner et al., 2013). SOM PPA uses weights as a proxy for species risk of establishment, but directly comparing SOM weights for the same species between studies is invalid because weight values change whenever different input data are used. This variability casts doubt upon the capability of SOM species weights to be used as indicators of species establishment risk. Weights change because they are m-dimensional coordinates in the m-dimensional space (where m is the number of species) created by the SOM algorithm. Thus, when

* Corresponding author.

http://dx.doi.org/10.1016/j.ecolmodel.2016.10.003
0304-3800/© 2016 Elsevier B.V. All rights reserved.

input datasets contain different species, the *m*-dimensional spaces and coordinates will also differ, and the same species will receive different weights for the same target region. An alternative is to use species' relative ranks to generate the output risk lists (Paini et al., 2010b). However, it remains uncertain if relative ranks generally show more stability between input datasets than species weights.

An example can help explain the weights variability problem. In Worner and Gevrey (2006), the highest ranked species (rank 1) was *Planococcus citri*, which received a SOM weight of 0.93. The second ranked species (rank 2) was *Icera purchase* which had weight 0.92. When the analysis was run with updated data from 2014 (unpublished data), the global distributions of some species had changed, and *Planococcus citri* obtained a weight of 0.82 and *Icera purchase* obtained 0.71. Nevertheless, their ranks remained first and second.

Another issue is how regions with few species, and species that are present in very few countries, impact SOM PPA results. In the simulated data test of Paini et al. (2011), SOM PPA had difficulty distinguishing species that could establish in regions with few species from those which could not. Thus, they suggested that species-poor regional pest profiles should be excluded from the analysis. Similarly, Singh et al. (2013) found that species which were present in few regions had significantly lower weights than widespread species, which suggested that weight (and rank) could be correlated with species' worldwide prevalence. This, however, is controversial since Watts and Worner (2009) showed otherwise.

A third issue is that species occurrence datasets are highly dimensional, which puts SOM PPA at risk of the 'curse of dimensionality' (Breiman, 2001). Each new species in the input dataset represents a new dimension for the algorithm to account for, but also provides more information for the algorithm to learn from. Thus, there may be trade-offs between number of species and the accuracy of SOM weights and ranks. Knight et al. (2011) tentatively explored the effects of data dimensionality (number of species) on SOM PPA results and obtained contradictory results.

The overall aim of our study was to investigate the sensitivity of the SOM PPA outputs to changes in input data. Specific objectives were to assess: the relationship between weight variability and number of species in the dataset, the relative stability of weights and ranks, and the relationship between weight, rank and global species prevalence. We created datasets of different dimensionality and studied changes in weights and ranks of each species.

2. Methods

2.1. Terminology

SOM PPA terminology is sometimes confusing. In Table 1 we aggregated model nomenclature used across the different studies cited in this paper, and chose one name for each feature.

2.2. The self-organizing map algorithm

A SOM is an artificial neural network first described by Kohonen (1982). It is a machine learning algorithm suitable for analysing non-linear highly dimensional data that converts relationships amongst a set of variables to two dimensional maps of clusters. It consists of two layers of neurons. The input neurons are the variables in the input matrix. When the sample units (rows) are presented to the algorithm, SOM captures the similarities between them through a machine learning process, and places similar sample units close together on an output map (Kohonen, 2013). The output map is also composed of neurons (output neurons). The number of neurons of the output map is smaller than in the input matrix because multiple individuals are mapped onto fewer number of output neurons, which creates clusters.

Table 1
SOM PPA terminology

Unified SOM PPA nomenclature		
Name	Short description	Other names
Input matrix	Matrix of regions and species to classify	Input layer, occurrence matrix, pest profiles matrix, input dataset
Pest profile	Each row of the input matrix that defines the presence/absence of all the pests in a region	Input neuron, regional profile, regional pest profile, input vector
Output map	Two dimensional representation of SOM classification results composed of *n* output neurons	Output layer, SOM map
Output neuron	Smaller constituent unit of output map	Neuron, cluster, unit, cell
Weight vector	Vector of coordinates for each pest profile in the output neuron to which is classified	Weights
Species weight	Each component of the weight vector that corresponds to each species of the input matrix	SOM index, species risk, risk of establishment, risk index
Species rank	High (1) to low order of species weights for a target region	Rank

2.3. The SOM PPA

In SOM PPA, rows of the occurrence matrix are regional pest profiles. In the final output map classification, two pest profiles mapped to nearby neurons are more similar than two pest profiles allocated to neurons that are far apart. Input and output neurons are linked through a parameter called the weight vector.

Weights describe the position in the output map of each of the regional profiles of the input matrix. They are coordinates of each pest profile in *m*-dimensional output space where *m* is the number of species of the input matrix (Gevrey et al., 2006).

Ecologically, weights are interpreted as the degree of association between a species and a particular regional profile. Thus, the higher the weight for a species, the more closely associated the species is with that regional profile, and consequently, with all the regional profiles clustered nearby. When modelling binary presence/absence data, weights range between 0 and 1.

Fig. 1 outlines the SOM PPA process. The first step is to identify the neuron to which the target region has been allocated. Then the weight vector for that neuron is extracted. Each component of the weight vector corresponds to one species weight and represents the degree of association between that species and the target region. Species are then ranked by weight, and the species with the highest weight is given rank 1.

2.4. Data

We used the occurrence data extracted by Worner and Gevrey (2006) from the Plant Quarantine Data Retrieval System (PQR) and CABI Crop Protection Compendium (CABI, 2007). It comprised the global distributions of 873 insect pest species for 460 regions. We then subsetted the occurrence matrix (*dataset A*) into 10 occurrence matrices, one for each of the 10 most common crops worldwide; apples, bananas, cotton, grapes, maize, mangoes, potato, rice, tomatoes and wheat (FAO, 2006).

We named these crop restricted matrices *datasetsB$_i$*, where *i* = *crop* (Fig. 2). The species present in each data set varied according to whether they were associated with the crop and associations were determined using the information in PQR. The range in