The structure of global invasive species assemblages and their relationship to regional habitat variables: converting scientifically relevant data into decision relevant information.

A thesis
submitted in partial fulfilment
of the requirements for the Degree of
Doctor of Philosophy
at
Lincoln University

by

Mariona Roigé Valiente

Lincoln University, New Zealand
2017

A la Nut, per la seva espera < 3

Abstract

Quantitative methods for pest risk assessment combine sound statistical tools with sound ecological theory to convert scientifically relevant data into decision-relevant information. This thesis investigated a quantitative method for pest risk assessment called pest profile analysis (PPA). PPA is a new methodology that is based on the premise that the risk of invasion by crop pests into new areas can be predicted by analysing regional insect pest assemblages (also known as pest profiles). Regional pest assemblages comprise the presence or absence of recognised pest species in each region of the world. The analysis involves clustering these regions based on similarities between their pest profiles. PPA assumes that co-occurrence of pest species in a region is the outcome of a non-random structured process driven by biotic and abiotic characteristics of the region. The most commonly used clustering technique for grouping regional pest assemblages is a self-organizing map (SOM), which is an artificial neural network algorithm. Two other clustering methods that have also been used for PPA are hierarchical clustering (HC) and k-means. The main aim of this thesis was to perform a thorough validation test of the PPA approach. To do so, I first analysed the sensitivity of SOM PPA to changes in the number of species used as input data. The results showed that SOM PPA outputs (weight values that are interpreted as risk indices) were quite sensitive to changes in the input data. However, when the risk indices were transformed into ranked lists of species, the ranks were significantly less sensitive and hence potentially more useful for pest risk assessment. I assessed the validity of the groups (clusters) of regions obtained from a SOM PPA by applying an external validation measure, the ζ diversity metric. The ζ metric was used to quantify similarities between pest profiles within clusters. The results showed it can be used for assessing the uncertainty associated with PPA outputs. I also conducted a temporal study of distributional changes of crop pests worldwide to measure the degree of biotic homogenization that had occurred in the regional pest profiles over 10 years. The major findings were that homogenization is certainly occurring, but it is in an inceptive stage and pest assemblages still remain strongly regionalized. I made a detailed comparison between the SOM, HC and k-means clustering methods to identify the one that produced the most accurate predictions. Unexpectedly, HC performed best. This appears to contradict the main hypothesis behind clustering world's regions according to their pest profiles because the expectations were that since SOM and k-means create a higher number of highly similar clusters, they would provide better predictions. The results of this research showed that PPA can help to prioritise risks of invasion by insect pests. It provided a new measure of uncertainty to improve communication of model results to decision makers. The results highlighted the urgent need for research to identify the determinants of insect pest species' distributions around the globe, and to implement that knowledge into PPA and biosecurity decision making.

Acknowledgments

I feel the need to acknowledge and thank everyone along the way. And that is going to be long because I am submitting this thesis after an enjoyable and well-lived education path that lasted 30 years. I had an early start and took it easy. I started school at kindergarten Montisbel when I was still not one year old. Thanks Senyoreta Montse and Senyoreta Tina for teaching me how to read, how to write and how to make those amazing penne pasta necklaces that I wore with so much pride.

Thanks to my primary school, La Salle Palamos, for its big sandy playground, for the amazing stone stairs where we took the end-of-the-school-year pictures and where so many other life-changing things happened for the first time. And thanks Rodri, for being the 4th grade teacher every 9 years old in the world deserves to have in their life.

Special thanks go to my secondary high school, IES Palamos, where I learned that the world was a much bigger place than what I thought before, where I met Javi, who, by inscrutable means, is still a part of my life today. Thanks Nuria Corredor, before writing in a public forum, I always think about you, and about how embarrassed I would be to disappoint you by misspelling a Catalan word.

Then, University of Girona was the place where I graduated as an Economist and then obtained my masters in Environmental Sciences. Without the thoughtful mentoring and advice from Gemma Renart, Diego Varga, Marc Saez, and Andreas Kyriacou that would never have happened. Seriously, never!

And here we are... at the final stage, what could have been better to finish with a life-time of schooling experience than my supervisor, the unbeatable Associate Professor Susan P.Worner? Thanks Sue, thanks for believing in me, or at least for pretending so well! I will miss you.

I extend abundant and sincere thanks to the rest of my co-supervision team, Dr. Craig Phillips for being the best language editor that a non-English speaker can ever dream of, among many other qualities. Also thanks so much Dr.Matthew Parry for the patience and your ability to put extremely complex concepts in understandable words that even I could grasp. You have been very useful and I appreciate it so much. I would also like to thank my collaborators in Monash University, Associate Professor Melodie A.McGeoch and Professor Cang Hui from Stellenbosch University for allowing me to work at a top-level project hands on with them.

Thanks to all the Ecological Informatics department, to the ones that stayed and the ones who went on with their amazing careers elsewhere. Dr.Hossein Ali Narouei Khandan, Jack Linton (my favourite summer scholar ever), Marona Rovira Capdevila, Ursula Torres and Dr.Senait Senay. And very special *merci beaucoup* to my very special Dr. Audrey Lustig, whose last name stands for 'merry' or 'cheerful' in English, and she has taught me more than what she will ever realize.

To the friends I left in the northern hemisphere, most of whom are today spread throughout the world, I miss you, think you and thank you. Nuria Abad, Marjorie Aznar, Deli Moreno (and Fletxa), Dani Sanchez (where are you now?), Marc Melus, Pau Mora, Joan Romans, Estibaliz Herranz (and Trap and Fion), Nahia Canibe, and my two little ones (with their two little ones), Sonia Cabezon and Natalia Abascal, and Nahia and Ona.

To the friends I gained in the southern hemisphere. New Zealand without you would just have been a bit less great. Martona ('hemos sabido volver a encontranos'), Gabriel (whatever I write here, it won't be special enough), Ursu ('the best you can is good enough'), SimonLimón , Cucaracha, Laura, Marjon, Damià, Dede, Sam, Andy, Mark, Bruno, Marona, Euge, Fede, Cesco, Aimee, Pri, Majo, Richard, Ruth, Pavi, Mauricio, Fanny, Cian, Nia, Yann and all those that are still to come.

And a final acknowledgement and all my gratitude to my family. Families are complicated because they are made of people. Thank you, Avi, Mama, Papa, Josep, Cristina, Rosa, Mar, Pau, Iaia, and to the rest of my extended family. Sometimes though, families are chosen, thanks Ash, Nut and Drake, I could not have chosen better.

Contents

List of Figures

List of Tables

Chapter 1

General introduction

1.1 Biological invasions and Biosecurity

The book 'The ecology of invasions of animals and plants' by Charles S. Elton in 1958 is considered to have set the basis of the Invasion Biology field. It described the global distributions of seven invasive species and analysed, for the first time, the relationships between the species' populations and the habitat patterns that were disrupted by them (Arnold B . Erickson, 1960).

Since then, invasive species have been considered by many sub-disciplines in ecology, biology and biogeography and, due to this multidisciplinarity (Hulme, 2011), they have also been the source of extensive debate. There have been two main groups of invasion biologists divided by the taxa they study; those concerned with animal invasions and those concerned with plant invasions (Blackburn *et al.*, 2011). Each of those groups has adopted different invasion model frameworks, resulting in controversy about the adequateness of the terminology and the limits of the concepts and definitions. Basically, animal ecologists traditionally have followed Williamson's invasion framework (see (Williamson, 1989; Williamson & Fitter, 1996; Williamson, 2006)) while plant ecologists have followed Richardson's classifications (Richardson *et al.*, 2000, 2011). Effort has recently been made (Blackburn *et al.*, 2011) to bring together the animal and plant invasions ecology traditions to create a general invasion ecology framework. Another notable dispute in invasion biology has concerned the political and ethical standpoints towards the invasive species problematic (see Warren (2007) for a critique of the language and practice in the field, then Richardson, D.M., Pyšek, P., Simberloff, D., Rejmánek, M., Mader (2008) for the response and then Warren (2008) for the counter-response).

In this thesis I adopt the terminology of Blackburn's (2011) proposal for an unified framework for biological invasions and define an ***alien species*** as a *non native species that has been transported beyond the limits of its native range by human mediated dispersal.* The arrival of an alien species can result in different degrees of colonization. One widely accepted system to characterize the potential colonization success is the naturalization-invasion framework (Pyšek *et al.*, 2004; Richardson & Pyšek, 2012) that excludes previous binary definitions of invasiveness. Instead, each alien species is placed somewhere along a continuum from casual invader to successful invader depending on the invasion stage (or barriers) that they have overcome. Nevertheless, it is also informative to define degrees of invasiveness according to colonization success. (See Table 1.1.)

Table 1.1: Definitions of the levels of invasiveness used in this thesis. Adapted from Blackburn *et al.* (2011)

Level of invasiveness	**Definition**
Casual	Alien species transported out of its native range either in captivity or quarantine. When released into the novel environment it is incapable of surviving for a significant period.
Naturalized/Established	Alien species with individuals surviving in the wild in the location where they were introduced, either reproducing or not. A wild population might be sustained.
Invasive	Alien species with self-sustaining populations at multiple sites in the wild, with individuals spreading significantly from the original point of introduction.

Many ecological factors have been shown to influence the success of alien invasions (Williamson, 2006). They have been summarized in Pimentel *et al.* (2005) as: 1) lack of natural enemies in the new habitat, 2) development of new host or parasite associations 3) presence or absence of other predators 4) degree of disturbance of the newly colonized habitat, and 5) degree of adaptability of the species. Much attention has also been paid to invasion consequences or impacts. An invasive species is likely to cause some impact that will alter the established order of the invaded ecosystem (Vitousek *et al.*, 1996), and potentially stimulate some physical and structural changes in those ecosystems where it is introduced (Mack *et al.*, 2000; Simberloff, 2011).

Invasive species may or may not be considered pests, which depends on the level of impact they cause. It seems that there are no particular traits that can describe a species as a pest; 'each pest is a pest for its own reasons' (Williamson & Fitter, 1996).

According to the International Union for Conservation of Nature (IUCN), a species is a pest if it causes either other species to decline or the structure and function of natural and productive ecosystems to be altered, resulting in economic impacts and/or decreases in biodiversity (IUCN/SSC, 2000; Worner, 1991). Therefore, in this thesis I follow IUCN criteria and define a ***pest*** as *an invasive species that can either cause economic damage or a decrease in biodiversity in the environment where it has been introduced.*

Evaluating ecological impacts is complex and subject to strong biases (see Blackburn *et al.* (2014) for one of the latest and more comprehensive developed frameworks for impact classification). Consequently, the true ecological impact of a pest remains rather imprecise (Roques, 2012; Vilà, 2013). In contrast, economic impacts have often been more precisely quantified. Even though most estimates are approximate and are often underestimates (Pimentel *et al.*, 2001; Bradshaw *et al.*, 2016), they need to be taken very seriously. In the United States, introduced species are estimated to cause up to USD$120 billion per year of environmental damage and loss of production (Pimentel *et al.*, 2005). An example of how extraordinarily costly one single species can be, is the red imported fire ant (*Solenopsis invicta*), which is native to South America and has become widespread in southern USA and Caribean (CABI, 2007). Pimentel's (2005) report approximates the damage caused by this ant as USD$1 billion per year in the US. For New Zealand, the overall economic damage due to invasive species was estimated in NZD$2 billion per annum in Barlow & Goldson (2002) and Clout (2002). More recently, the total economic cost of pests to New Zealand including the downstream economic impacts from pest-related output losses has been calculated as NZD$2.128 billion, or 1.20% of New Zealand's gross domestic product (Giera *et al.*, 2009). Analogously, in Europe, the estimate for invasive species impact is of USD$13 billion per year, but this figure is probably an underestimate, as potential economic and environmental impacts are unknown for almost 90% of the alien species in Europe (Hulme, 2009). Ultimately, the economic impacts of invasive species are especially expensive for developing economies (Vitousek *et al.*, 1996), which are disproportionally vulnerable to invasions by agricultural pests (Paini *et al.*, 2016).

1.1.1 Biosecurity: Policies, Agencies and Pest Risk Analysis

Modern societies depend on productive economies which rely on trade, market access and tourism. Together, global movements of people and goods facilitate the spread of pest species. Human assisted dispersal of pest species occurs through the direct imports and exports of commodities, and also as pests hitch-hiking in containers, other conveyances, and even on humans themselves. Regardless of entry pathway, the arrival of a pest causes a change to the recipient nation's economic well-being, due to both its arrival and to the efforts taken to mitigate its impact (Perrings *et al.*, 2005; Warziniack *et al.*, 2013).

Most trading nations have recognized the invasive species problem and have established regulated trade procedures to mitigate it. The United Nations' Food and Agriculture Organization (FAO) and the World Trade Organization (WTO) have produced several worldwide agreements and standards, such as the International Plant Protection Convention (IPPC). The IPPC is an international agreement on plant health, currently with 180 adhering countries, which aims to protect cultivated and wild plants by preventing the introduction and spread of plant pests. To support the IPPC, there are currently nine regional plant protection organizations (RPPOs) around the world. Some of these are the European and Mediterranean Plant Protection Organization (EPPO), the Inter-African Phytosanitary Council and the North American Plant Protection Organization (NAPPO) (all are listed in https://www.ippc.int/en/partners/regional-plant-protection-organizations/). Some countries also have their own national plant protection organisations (NPPOs). Examples are New Zealand's biosecurity agency regulators within the New Zealand Ministry for Primary Industries and Australia Biosecurity within the Australian Department of Agriculture, Fisheries and Forestry. Those two agencies are commonly used as an example of what would be desirable elsewhere, mainly due to their rigorous management of phytosanitary risks from international trade (Bacon *et al.*, 2012).

The term 'biosecurity' refers to a coordinated approach, generally led by a particular governmental authority or network of authorities, to understand and manage natural and human-caused threats to a range of biological resources (Quinlan *et al.*, 2015). The main aim of biosecurity authorities is to make informed decisions based on trade-offs between preventing species invasions while sustaining or facilitating trade levels and tourism. To do so, agency managers have to make important choices. Given the large number of items that cross the border each day (in New Zealand that number is estimated at 170,000 a day (Silcock & Guy, 2013)) biosecurity programs cannot target all alien species. Instead, border controls, policies and quarantine procedures have to be efficiently prioritised and implemented.

To help achieve such efficiencies, the IPPC published the International Standards for Phytosanitary measures (ISPMs), which are part of the FAO global programme of policy and technical assistance in plant quarantine. This programme makes available to FAO members and other interested parties, standards, guidelines and recommendations to achieve international harmonization of phytosanitary measures. Some interesting tools provided by the ISPMs are the general Phytosanitary principles for the protection of plants and the application of phytonsanitary measures in international trade (ISPM 01) (International Plant Protection Convention (IPPC), 2006a), the Framework for pest risk analysis (ISPM 02) International Plant Protection Convention (IPPC) (2007) and the pest risk analysis for quarantine pests (ISPM 11) (International Plant Protection Convention (IPPC), 2004). All the ISPMs were republished in 2016 and are publicly available at the FAOs website (https://www.ippc.int/en/core-activities/standards-setting/ispms/).

1.1.1.1 Pest Risk Analysis

Pest risk analysis (PRA) is divided into three stages: initialization, assessment of the risk, and management. During initialization, the pests and pathways that require PRA are identified. Then, pest risk assessment determines the status of each species as a pest, its associated entry pathways, and characterizes its likelihood of entry, establishment, spread and economic importance. The third stage; pest risk management, involves the development, evaluation, comparison and selection of strategies for reducing risk. The geographical area to which a PRA applies is usually a country, but can also be an area within a country, or a more extensive area covering all or parts of several countries.

Pest risk analysts conduct PRA. They investigate the presence of the pest in different places on the biosecurity continuum, including pre-border, at-border and post-border. They have to seek answers to questions of where the pest is currently present, where could it get into the future, what are its possible entry pathways and the time it might take to spread (Low-Choy, 2015). Therefore, risk analysts gather a lot of complex and uncertain data from global databases (see McGeoch *et al.* (2012) for a review of uncertainty present in invasive species listings) and interpret it, within restricted time frames, to extract the information needed for decision making. PRA are science-based evaluations that usually involve reviewing and interpreting articles and databases but not scientific experimentation. Usually risk analysts draw on scientific literature to make inferences regarding the components of the overall risk(McLeod, 2015). Many quantitative and qualitative tools have potential to assist pest risk analysts with this complex process.

While academic research has focused on refining quantitative predictive models for risk assessment (some examples are (Liu *et al.*, 2011; Singh *et al.*, 2015; Wattenbach *et al.*, 2006)), policy improvements have centred upon better elicitation of expert knowledge based on risk-scoring methods (Leung *et al.*, 2012). Qualitative methods generally consist of subjective statements (often verbal classifications) regarding elements contributing to risk before providing a conclusion on the overall risk (McLeod, 2015). However, expert-based risk assessment is known to be highly biased by the experts morals, beliefs and by their self-perceived objectivity (Burgman, 2005). On the other hand, quantitative methods aim to obtain numerical estimations of the risk, using deductive statistical approaches. However, quantitative models for PRA are often perceived as too complex and uncertain by pest risk analysts and can also be biased by subjective knowledge when data for risk factors are unavailable or uncertain (McLeod, 2015). Both quantitative and qualitative PRAs and their combinations can be improved, although it is unrealistic to expect a completely automatated quantitative method that makes expert input redundant and can be used under any circumstance for any pest (Sutherst & Bourne, 2008; Sutherst, 2014).

1.1.1.2 Quantitative approaches to PRA

At what scale biosecurity research should focus remains difficult to decide (McNeely *et al.*, 2001; Hulme, 2003) since the understanding of what scale ecological processes such as invasion, spread and species production (speciation) actually happen is still limited (Ricklefs, 2004; Lewinsohn *et al.*, 2005; Lawton & Gaston, 1989). However, in a globalized world where humans have removed major biogeographic barriers to species disperal, the methods developed to aid PRA need to have a global scope. Similarly to the global perspective required in invasion biology (Lewinsohn *et al.*, 2005; Chown, 2015), there is a biogeographic approach to pest risk assessment.

In general terms, quantitative biogeographical approaches to pest risk assessment (also known as 'pest risk mapping' (Venette *et al.*, 2010)) aim to prioritize geographic domains suitable for the establishment of pest species (insects, weeds or diseases). Pest risk mapping produces models, lists and maps that can help answer questions of what species are of concern, where these species occur, how they may spread if they were introduced and what their potential impacts might be. These models, maps and tools should have strong fundamental ecological and geographical foundations. Some reviews of biogeographical approaches to pest risk assessment can be found in Sutherst (2014) and Leung *et al.* (2012), along with model comparisons in Sutherst & Bourne (2008) and also a list of the identified flaws and directions for improvement in Venette *et al.* (2010).

In this thesis I am interested in the subset of the quantitative biogeographical approaches to pest risk assessment, specifically the creation of lists of pest prioritization based on the study of global invasive pest assemblages.

1.2 Global invasive assemblages analyses

An assemblage of species is the group of species occupying a particular site. In the context of agricultural plant protection, the species of interest are crop pests, either insects, weeds or other pests and pathogens such as fungi, virus and bacteria. Therefore, the ***pest assemblage*** *of a region is the group of pest species that co-occur in that region.*

1.2.1 Pest profile analysis (PPA)

The study of pest assemblages to inform biosecurity decisions commenced with the publication of Worner & Gevrey (2006) and Gevrey *et al.* (2006). In these two papers, the available data on global insect pest assemblages was used to infer the risk of establishment of a list of approximately 800 insect pest species in a region of interest, which was New Zealand. While Worner & Gevrey (2006) focused on providing a list of species ranked by their potential threat to New Zealand, Gevrey *et al.* (2006) presented the methodology in a generalizable form and suggested its potential application to other regions of interest, and potentially to other taxa. The authors used the terminology 'pest profile' of a region as an exact synonym of 'pest assemblage' of a region, and it has been since then used interchangeably across all related literature (Table of terminology and synonyms in Chapter 2, section 2.1).

The rationale behind pest profile analysis is that a pest assemblage implicitly contains information about a region's biotic and abiotic conditions. Biotic conditions that can influence the composition of the assemblage inlcude, for example, the agricultural crops grown in the region and influential abiotic conditions include the region's climatic characteristics (Eyre *et al.*, 2012). Other regional characteristics that could influence assemblages include the tectonic activity, historical trade paths (Ricklefs, 2004), structure and quantity of trade, and the biosecurity measures applied. It is assumed that regions with similar pest profiles will also have similar biotic and abiotic characteristics that allow the species to establish. Therefore, comparing pest assemblages between regions can provide insights about which regions may exchange species with high probabilities of establishment.

Worner & Gevrey (2006) and Gevrey *et al.* (2006) analysed pest assemblages using a clustering method known as self-organizing map (SOM) (Kohonen, 1982, 1990). Section 1.2.2.1 contains a more detailed introduction to the SOM algorithm. The SOM method is especially useful for clustering highly dimensional data and has been be applied to many fields of scientific research (see Oja *et al.* (2003) for a detailed review of applications), such as finance (Deboeck & Kohonen, 1998), genomics (Törönen *et al.*, 1999), natural language processing (Honkela, 1997) and numerous areas of ecological sciences (Chon, 2011) where it has successfully described environmental or species spaces by clustering sets of environmental variables.

In this thesis I will refer to the approach of *clustering pest profiles as a method for inferring risk of invasion as* ***pest profile analysis (PPA)***. PPA has been used by many studies since Worner & Gevrey (2006) and (Gevrey *et al.*, 2006). Paini *et al.* (2010a) used PPA to raise the alarm about biosecurity risks from internal trade within the US, and Paini *et al.* (2011), tested the PPA approach on fungal pathogens for first time, using artificially simulated data.Watts & Worner (2012) studied assemblages of bacterial crop diseases and Vänninen *et al.* (2011) used PPA for updating a list of alien invertebrate pest species in Finland. More recently, Singh *et al.* (2013) extended the application of PPA to plant parasitic nematodes, and Singh *et al.* (2015) incorporated the PPA approach into a quantitative PRA method called pest screening and targeting (PeST). Broader extensions were also made by Morin *et al.* (2013) who applied PPA for the first time to studying and prioritizing of weeds and by Eschen *et al.* (2014) who applied it to assemblages of bacteria and nematode pests.

1.2.2 Methodological approaches to PPA

Clustering *is the process through which the data is divided into meaningful groups* and it is usually one of the first steps in analysing data (Davies & Bouldin, 1979). The aim of clustering is partitioning the data into groups such that the observations in a cluster (group) are more similar to each other than observations in different clusters (Mangiameli *et al.*, 1996). In cluster analyses, there are no predefined classifications and the clustering algorithm has the task to divide the data into natural groups. Algorithms that perform clustering are called unsupervised learning algorithms.

1.2.2.1 Introduction to Self-Organizing maps

A self-organizing map (SOM) is an algorithm that belongs to the family of artificial neural networks. It is an information-processing paradigm inspired by the functioning of vertebrate brains (Kohonen, 2013). A SOM neural network is composed of two layers of neurons: the input layer and the output layer. Let the data be organized in a matrix where the rows are the input patterns (observations) and the columns are the input neurons (variables). The output layer is represented by a rectangular grid with $m * n$ neurons (also called cells or units) laid in a hexagonal lattice which have meaningful neighbourhood relationship. A SOM organizes information spatially, mapping similar input patterns onto adjacent output neurons. The SOM algorithm can convert complex, non-linear statistical relationships between high-dimensional data items into simple geometric relationships to be visualised in a low-dimensional display (Kohonen, 1982, 2001), thus performing both vector quantization and vector projection. A SOM projects the vectors in the input space onto the output space while attempting preserve the topological relationships observed in the input space.

The SOM is trained iteratively through a large number of epochs. An epoch is the processing of all the input patterns once, thus each input pattern will be processed as many times as the number of epochs. Each neuron of the input layer, has as many weights as the input patterns, and can thus be regarded as a vector in the same space as the patterns.

Output neurons are initialized randomly (are given a set of coordinates (weights) in the multidimensional output space) and then they are trained. When the SOM is trained to an input pattern, the distance between that specific pattern and every neuron in the output space is calculated. Then the closest neuron in terms of distance (Euclidean distance) is defined as the winning neuron, and the pattern is mapped onto that winning neuron. As a consequence, the neuron moves toward the input pattern position in order to improve its representation and this movement is translated into a change in its coordinates (weights). The extent of this movement is controlled by a parameter usually referred to as learning rate.

In order to preserve the topology of the input patterns in the output space, it is essential to correct both the position of the wining neuron and also the position of its neighbouring neurons. Thus, the network is progressively organized (unfolded) with certain parts of the input space being represented by certain subsets of neighbouring neurons. The first part of the training phase is the coarse training or unfolding, where the neurons of the output space are spread out and pulled towards a general area of the multidimensional space, thus defining its general shape. The second is the fine tuning phase, where the SOM matches the neurons as far as possible to the input patterns, thus decreasing the quantitization error.

If a SOM has been successfully trained, then patterns that are close in the input space will be mapped to neurons that are close (or the same) in the output space. This quality is called topology preserving.

At the end of the learning process, there will be a difference between an input pattern and the neuron it is mapped to. This difference is called the quantization error and it is used as a measure of how well the neurons represent the input patterns. Another metric to evaluate the performance of the SOM algorithm is the topological error, which measures the average number of times the nearest and the second nearest neurons of an observations in the input space do not correspond to adjacent neurons in the output space (Kiviluoto, 1996). The higher the topological error, the poorer representation of the input layer onto the output layer.

Formally:

Let x k (with $k = 1$ to the number of training patterns N) be the n-dimensional training patterns; w_{ij} be the neuron in position (i, j); $0 \leq \alpha \leq 1$ be the learning rate and $h(w_{ij}, w_{mn})$ be the neighbourhood function. This neighbourhood function assumes values in $[0, 1]$ and is high for neurons that are close in output space, and small (or 0) for neurons far apart. Let w_{winner} be the winning neuron for a given input pattern.

The algorithm for training the network is then: For each input pattern:

1. Calculate the distance between the pattern and all neurons ($d_{ij} = ||x_k - w_{ij}||$)
2. Select the nearest neuron as winner ($w_{ij} : d_{ij} = min(d_{mn})$)
3. Update each neuron according to the rule $w_{ij} = w_{ij} + \alpha h(w_{winner}, w_{ij})||x_k - w_{ij}||$
4. Repeat the process until stopping criterion is met. Usually, the stopping criterion is a fixed number of epochs. (algorithm transcription adapted from Bação & Lobo (2010))

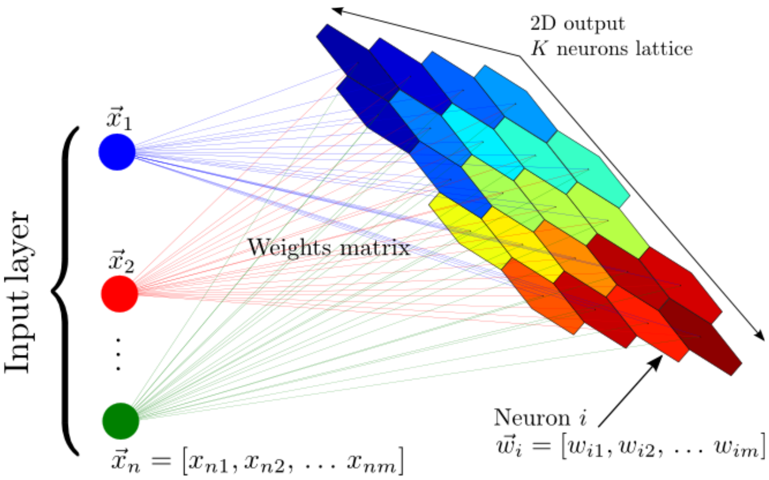

2D output
K neurons lattice
$\vec{x}_1$
$\vec{x}_2$
Input layer
Weights matrix
$\vec{x}_n = [x_{n1}, x_{n2}, \dots x_{nm}]$
Neuron i
$\vec{w}_i = [w_{i1}, w_{i2}, \dots w_{im}]$

Figure 1.1: Visualization of SOM input layer, output layer and its relationships. Adapted from http://matias-ck.com/mlz/somz.html.

1.2.2.2 Introduction to k-means

Other clustering methods have been used for PPA besides SOM. For example, Watts & Worner (2009) used the k-means algorithm to cluster the regional insect pest profiles and compared it to the results obtained by Worner & Gevrey (2006) using SOM PPA. K-means is an unsupervised learning algorithm first described in MacQueen (1967) that produces crisp clusters through a partitional clustering procedure. Partitional clustering attempts to directly decompose the data into a set of unrelated clusters by attempting to determine an integer number of partitions that optimise a certain criterion function (Halkidi *et al.*, 2001). The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed *a priori*. The main idea is to define k centroids, one for each cluster. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The cluster centroids are initialized by placing them as far as possible from each other. Then, each input pattern is associated to its nearest centroid. When all the input patterns have been associated, the first step is completed and an early grouping is done. At this point k new centroids are recalculated. After these k new centroids are obtained, a new binding has to be done between the same data set points and the nearest new centroid, as a result the k centroids change their location step by step until no more changes occur.

The algorithm proceeds (Adapted from Watts & Worner (2009)):

1. Select k seeds as initial centroids. These can be vectors that are generated randomly, or vectors that are selected from the data set being clustered.
2. Calculate the distance from each cluster centroid to each seed.
3. Assign each observation to the nearest cluster.
4. Calculate new cluster centroids, where each new centroid is the mean of all vectors in that cluster.
5. Repeat steps 2 to 4 until a stopping condition is reached.

A key assumption of k-means is that the algorithm expects the data clusters to be spherical and of same size, so that the assignment to the nearest cluster center is the correct assignment.

1.2.2.3 Introduction to Hierarchical clustering methods

In a more recent study (2014), Eschen *et al.* used hierarchical clustering to find similarities in their regional pest profiles. Unlike k-means, hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The result of the algorithm is a tree of clusters, called dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained (Halkidi *et al.*, 2001).

There are many variations of hierarchical clustering algorithms, and they can be roughly divided into:

Agglomerative. These step-wise algorithms merge results from the previous step by merging the two closest clusters into one.

Divisive. These step-wise algorithms split results from the previous step by splitting a cluster into two.

1.3 Global invasive assemblage diversity

Community ecology studies the diversity, abundance and composition of species in communities. Traditionally it focused on the processes that determine the species composition of local communities but eventually recognised that the composition and diversity of species, even at a local scale, depended fundamentally on the composition and diversity of the regional pool of species (Vellend, 2010). The species pool is a concept rooted in island biogeography and refers to all the species that are able to disperse to a focal site, regardless of their ability to tolerate the site's environmental conditions (Cornell & Harrison, 2014). The environmental filter is the relationship between a species and the environment, which acts as a selective force by ruling out species that are unable to tolerate particular environmental conditions. Species that are able to survive at a site may have certain phenotypic traits that reflect their environmental tolerance, and these traits may be shared among other species in the community. Consequently, species at a site exhibit phenotypic convergence in key ecological dimensions compared to a null expectation based on randomly sampling species from the larger species pool (Kraft *et al.*, 2014).

A very general theory of community dynamics is that species either disperse from a regional pool to a particular community, or are created through the evolutionary process of speciation. The relative abundances of these species in the community is determined by deterministic fitness differences between individuals of different species (selection), random changes in species abundances (drift) and ongoing dispersal (Vellend, 2010). Thus, community assembly is influenced by processes operating at a wide range of spatiotemporal scales, and local communities are assumed to reflect the cumulative effects of these processes (HilleRisLambers *et al.*, 2012).

1.3.1 Diversity of herbivore insect assemblages

The global compositional variation of non-native regional insect assemblages has been shown to differ from the global compositional variation of native insect assemblages by Liebhold *et al.* (2016), who also noted that the invasive species compositions differ from what would be expected from island biogiography predictions (Liebhold *et al.*, 2016; Burns, 2015). This suggests strong effects of human mediated dispersal and entry pathways.

For insect crop pests, the history of invasions is intimately linked to the history of agriculture. Crop pest regional assemblages are strongly correlated with the distribution of their host plants (Bebber *et al.*, 2014a), which in turn may also be related to the climate suitability of the region (Bacon *et al.*, 2014; Baker *et al.*, 2005).

Agricultural landscapes are highly disturbed. As Pimentel *et al.* (2005) noted, 90% of the food consumed by humans is being provided by only fifteen plant species. Consequently, agricultural landscapes across the world are very homogenized and environmental and ecological differences between cultivated regions in different parts of the world have been greatly reduced. When a crop pest is introduced to a new and remote area, the expected mismatch between its phenotypic characteristics and the local ecological conditions expected to act as an environmental filter, have been greatly attenuated (Guillemaud *et al.*, 2011).

But the phenomenon of homogenization has not only occurred among crops. Recently, some wide-scale studies on community composition for insect (Chown, 2015; Kuussaari *et al.*, 2010; Vellend *et al.*, 2007) and other crop pests (Bebber *et al.*, 2014a) have shown how the composition of crop pest assemblages around the world are homogenizing too (Bebber *et al.*, 2014b).

1.3.2 Biotic Homogenization of crop pest assemblages

Biotic homogenization (Vitousek *et al.*, 1996; McKinney & Lockwood, 1999; Lockwood & McKinney, 2001) is the phenomena through which the ongoing arrival of non-native species and the habitat disturbance caused by agricultural human practices (Vellend, 2010) increases the spatial and temporal similarity in the taxonomic composition of global biota. The immediate consequence of biotic homogenization is global loss of biodiversity.

Olden & Poff (2003) noted empirical studies of biotic homogeneity were increasing, but there were knowledge gaps in its theoretical aspects, thus they drafted a first conceptual model that comprised 14 theoretical ecological scenarios and mechanisms through which species invasions and extinctions led to different trajectories of biotic homogenization. Olden (2006) emphasized the lack of studies that examined and quantified the homogenization process for communities at multiple spatial and temporal scales. Many studies have found evidence of biotic homogenization and its links to agriculture for local communitites, sometimes to the larger extents of countries or regions. For example, Vellend *et al.* (2007) found homogenization patterns in forest plant communities in North America and Europe due to agricultural land use and (Kuussaari *et al.*, 2010) reported a decrease in diversity of butterflies in intensively cultivated landscapes with simplified land structure. However, Olden & Poff (2003) also reported some studies that questioned the existence of biotic homogenization. An example was Marchetti *et al.* (2006) on community composition of freshwater fish communities in California which reported biotic differentiation, or an increase in community diversity after the introduction of exotic species.

To our knowledge, two studies have measured biotic homogenization at a global extent. Bebber et al.(2014a) studied the distributional changes of 424 crop pests and pathogens over the years 2000 to 2014, and McKinney (2004) studied the plant inventories of 20 localities in the United States to measure whether exotic plants increased the similarity of those localities. Both Bebber *et al.* (2014a) and McKinney (2004) concluded that when considering exotic species only, their data showed signs of biotic homogenization.

Usually, biotic homogeneity is quantified either by comparing community composition measured by a particular similarity index over time or by comparing distances between communities in species space using clustering or ordination (Olden & Poff, 2003; Rahel, 1990). Comparing species assemblages using a similarity index is equivalent to calculating their compositional β diversity.

1.3.2.1 Measures of diversity

Beta diversity encompasses a variety of indices and concepts that reflect different components of between-sites species variability, or species turnover along environmental, spatial or temporal gradient (Whittaker, 1972; Vellend, 2001). There is a huge variety of indices that measure species richness at a spatial level, most of which represent slightly different natural phenomena (Tuomisto, 2010), therefore it is important to clearly define for any particular study the expression used to calculate alpha, beta and gamma diversity (Anderson *et al.*, 2011).

In this thesis we refer to β diversity as species compositional variation between sites, α diversity as the number of different species occurring at one site, and γ diversity as the number of different species in the regional species pool (Whittaker, 1972), (Figure 1.2).

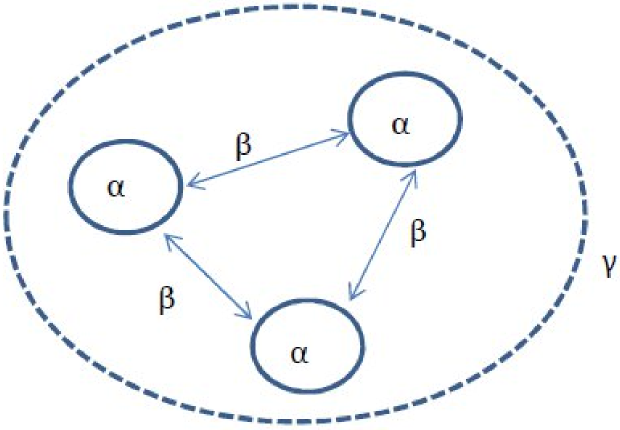


Figure 1.2: Conceptual plot of the α, β and γ diversities for three sites

Two problems associated with β diversity metrics are that community pairwise similarity indices assume site independence, and they can only compare two sites at a time.

Thus, when such metrics are used to compare larger groups, they are calculated by averaging the pairwise metrics across sites. To circumvent these issues, Hui & McGeoch (2014) proposed the zeta diversity (ζ) metric. Zeta diversity (ζ) is a measure of compositional β diversity defined as the number of species shared between any i number of sites or assemblages. Zeta is a simple similarity measurement that reconciles existing descriptors of species incidence and spatial turnover (Hui & McGeoch, 2014) and overcomes the main criticisms posed against the pairwise similarity measures such as Jaccard, Sorensen (Sørensen, 1948) and Shannon (Shannon, 1948). The ζ metric is unaffected by site-dependence, and enables more than two sites to be compared at once. This multi-site comparison feature allows different orders (numbers of features compared at at time) of ζ to describe different levels of similarity that can be very useful to understand the relationship between groups. In this thesis I use the metric ζ as the standard measure of α and β diversity.

1.4 Uncertainty in Ecological Modelling

1.4.1 Model based decision frameworks

Ecological modelling has two primary objectives. The first is to better understand nature, and the second is to use this improved understanding to help decision making. One difficulty of applying modelling to decision making is dealing with uncertainty. Uncertainty in modelling comes from several sources as described in the most important uncertainty taxonomies and reviews. Walker *et al.* (2003) defined a full framework and conceptual basis to integrate the concept of uncertainty arising from scientific modelling into decision-making processes. A policy is defined as 'a set of actions taken by an administration to control the system, to help solve problems within it or caused by it, or to obtain benefits from it (...) Policies are intended to help achieve goals'. In the context of biosecurity, the same rationale applies. Goals like controlling and monitoring invasive species, are strongly controlled by agencies and governmental institutions, and, are the responsability of policy-makers, who sometimes ask for help and advice from applied scientists. Walker *et al.* (2003) synthesised the policy-making process in the following multi-stage iterative process. Uncertainty is incorporated during all the stages of the process to come up with the so called uncertainty matrix. This matrix is a tool for the analysts to identify and char-

Reference from literature	Types of uncertainty considered
(United States Environmental Protection Agency, 1997)	Scenario uncertainty, parameter uncertainty, model uncertainty
(Morgan & Henrion, 1990), (Hofstetter, 1998)	Statistical variation, subjective judgement, linguistic imprecision, intherent randomness disagreement, approximation
(Funtowicz & Ravetz, 1990)	Data uncertainty, model uncertainty, completeness uncertainty
(Bedfort & Cooke, 2001)	Aleatory uncertainty, epistemic uncertainty, parameter uncertainty, data uncertainty, model uncertainty, ambiguity, volitional uncertainty
(Huijbregts *et al.*, 2001)	Parameter uncertainty, model uncertainty, uncertainty due to choices, spatial variability, temporal variability, variability between sources and objects
(Bevington & Robinson, 2002)	Systematic errors, random errors
(Regan *et al.*, 2002)	Epistemic uncertainty, linguistic uncertainty
(Walker *et al.*, 2003)	Location: context uncertainty, model uncertainty, level uncertainty
(Maier *et al.*, 2008)	Data uncertainty, model uncertainty, human uncertainty

Table 1.2: Uncertainty typologies from the literature - 1990 to 2008(Modified from Ascough et al. (2008))

acterize the potential uncertainty in model-based decision processes, and it can be used to assess uncertainty in the same fashion as sensitivity (Matott *et al.*, 2009). Drawing from Walker *et al.* (2003) 's framework, Wattenbach *et al.* (2006) developed a specific decision making process framework applied to ecosystem models.

Clearly, one of the biggest challenges in assessing uncertainty in any kind of ecological or environmental modelling is homogenizing the terms and concepts (Uusitalo *et al.*, 2015a; Aggarwal, 2009). Different taxonomies of uncertainty have been framed by different authors, and the confusion is great. The work published by Ascough *et al.* (2008) brought together most of the research regarding uncertainty in environmental decision-making processes and summarized all these different definitions in a comprehensive table (Table 1.2).

The principal idea that emerged from the Ascough *et al.* (2008) review is that Reagan's (2002) taxonomy of uncertainty could be summarized into the two main groups of reducible and irreducible uncertainty types. Uncertainty that is reducible or epistemic includes: knowledge uncertainty, which comprises process understanding, and model uncertainty; linguistic uncertainty, which arises from the vagueness or ambiguity of definitions and wordings; and decision uncertainty which is very important and is frequently underestimated by the policy assessors. There are also categories of uncertainty that are not reducible and arise from sources that we have no control of. It can also be called random or stochastic uncertainty and is related to the chaotic and unpredictable quality of natural and social processes.

1.4.2 Reducing versus showing uncertainty

Two strategies can be adopted when dealing with uncertainty from a modelling point of view: reducing uncertainty whenever possible and showing uncertainty whenever it is irreducible. For implementing both strategies and for reducing uncertainty, they are highly case specific. A detailed review of methods for assessing uncertainty in predictive methods can be found in Lustig (2016). For showing irreducible uncertainty, a standard framework has been prepared in the form of an ISO document (Joint Committee for Guides in Metrology (JCGM), 2008) in which clear guidelines are given on how to express magnitudes and measurements that contain aleatory uncertainty.

1.5 Research questions and objectives of this thesis

There is a need for quantitative approaches to help the decision making process in the context of pest risk analyses. In this introductory chapter I presented the current state of existing models for pest profile analysis, and the conceptual ecological principles and theories behind the study of regional pest assemblages. I also reviewed current approaches to incorporating uncertainty in modelling to assist decision-making. However, some of these areas need further study, development and validation.

1.5.1 Validation and methodological improvements for PPA

Since the first applications of SOM in ecology, several studies have been instrumental for illustrating their robustness and validity (Chon *et al.*, 1996; Lek & Guégan, 1999; Giraudel & Lek, 2001). Also, Paini *et al.* (2010b) demonstrated that SOM PPA is a tool insensitive to data errors up to 20%. However, all clustering techniques have methodological problems related to their data assumptions. The main issues across clustering methodologies are choosing the number of clusters (Mirkin, 2013) and validating the clustering results (Halkidi *et al.*, 2001).

To validate clustering results, they would ideally be compared to reality, but often the required information is either unavailable or insufficient to both train and validate the models. A common solution is to validate using artificially created, error-free data (Zurell *et al.*, 2010). For SOM PPA this sort of validation was carried by Paini *et al.* (2011) by simulating a virtual world with several invasibility parameters for its regions, and the comparing the SOM PPA rankings to those of the final invasions of the species. However, it is necessary to test the performance of the method against real observed values. Thus, we propose to carry out a model validation for SOM PPA using real occurrence data and compare the predictions of the model to the observed real values.

The SOM algorithm has advantages compared with the other main classification techniques of k-means and hierarchical clustering (HC) that make it particularly useful for analysing ecological data. Its biggest asset is the ability to capture non-linear relationships. Non-linear relationships between observations and variables will often occur in highly dimensional datasets which means some variables will be irrelevant or redundant, and present outlier observations (Mangiameli *et al.*, 1996; Park *et al.*, 2003) However, it is uncertain if SOMs are always the best clustering method for ecological data. Watts & Worner (2009) compared the SOM PPA to k-means algorithm and it seemed k-means performed better, both computationally and quantitatively. Other research studies have used other clustering algorithms such as hierarchical clustering (Eschen *et al.*, 2014) to perform PPA. It is my aim to understand which of these three clustering methods performs better with data such as global pest profiles.

Besides problems relevant to all clustering methods, some specific issues of SOMs have been identified by recent work. They can be summarized as: 1) when global prevalence of a species is small, the SOM PPA apparently is not very accurate at generating good rankings for the species (Singh *et al.*, 2013) 2) regions with pest profiles comprising fewer than eight species are significantly more unstable or difficult to predict (Paini *et al.*, 2011) 3) the analysis can be affected by sampling artefacts; and 4) direct comparison of risk levels between different databases may not be possible (Worner et al. unpublished).

Finally, SOM PPA is an applied modelling methodology that currently lacks any measure of ‘goodness’ or uncertainty, and it would benefit from a measure of uncertainty to better communicate its results to its intended end user, the risk assessor.

1.5.2 Study of the compositional diversity of regional pest assemblages

The PPA methodology is based on fundamental ecological principles such as the assemblage composition and the drivers behind community assembly. The main ecological hypothesis behind PPA is that two regions with similar pest profiles share environmental and historical conditions that enable them able to support similar assemblage of pest species. Therefore, identifying similar regions may help to identify regions that are most likely to succesfully exchange pest species in the future, which would assist pest risk assessment.

This hypothesis is based on robust community ecology principles, but remains empirically untested. In this research I explored the composition of global pests assemblages in terms of composition and diversity and describe their changes over time. By quantifying their levels of biotic homogenization in this way, I aim to better describe the ecological assembly mechanisms on which the PPA hypothesis is based.

1.5.3 Specific objectives

Objective 1: To perform a sensitivity analysis of SOM PPA in order to clarify the issues that other authors have found when using the method.

Objective 2: To validate the SOM PPA outputs by explaining results of the clustering process through the lens of community assembly.

Objective 3: To compare the performance of different clustering methods for PPA and find which method performs best according to the nature of the data.

Objective 4: To test the validity of inferring risks of invasion from clustering of pest profiles.

Objective 5: To measure levels of biotic homogenization among global crop pest profiles.

1.6 Thesis structure

1.6.1 Chapter 2: Sensitivity analyses of SOM PPA

Chapter 2 examines how variation between datasets influences the ouptuts of SOM PPA. The validity of using SOM weights and SOM ranks (model outputs) is tested by applying SOM PPA to 341 datasets and assessing the variability of its outputs.

1.6.2 Chapter 3: Cluster validity and uncertainty assessment of SOM PPA

Chapter 3 explores the value of incorporating an extra step in the SOM PPA method which consists of calculating a cluster validation metric (ζ diversity) that assesses the goodness of cluster of the SOM PPA results. It also proposes the zeta diversity metric as a measure of uncertainty communication for SOM PPA.

1.6.3 Chapter 4: Calculating the degree of Biotic Homogenization for PPA

Chapter 4 analyses the degree of homogenization among global insect pest profiles. The analysis compares the degree of similarity between regional pest profiles in two different years and uses ζ diversity as a new measure of quantification of biotic homogenization.

1.6.4 Chapter 5: The informative power of clustering pest profiles: Comparison of methods

Chapter 5 evaluates the PPA predictive power and a comparison of clustering methods. Three different clustering approaches (SOM, k-means and hierarchical clustering) are applied to a global dataset of pest occurrences in the year 2006 and then validated with an equivalent dataset that documents the same pests' distributions observed in year 2014.

1.6.5 Chapter 6: General discussion

Chapter 6 discusses all results reported in the thesis, their theoretical and practical implications and their contribution to addressing the research topics outlined in the objectives. It also provides concluding remarks and recommendations for future research.

Chapter 2

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

Notes

This chapter is published as:

Roigé, M., Parry, M., Phillips, C., Worner, S.P (2016). Self-organizing maps for analysing pest profiles: Sensitivity analysis of weights and ranks, Ecological Modelling, 342, 113-122.

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 preprocessing is needed before analyses, which has the dual benefits of reducing analysis time and modeller-induced bias.

Keywords

Self-organizing maps, pest profile analysis, clustering, prioritisation, invasive pest assemblages

2.1 Introduction

Over recent decades there has been considerable research on biological invasions and their impacts (Barlow & 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 & 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 programs.

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; International Plant Protection Convention (IPPC), 2006b; 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 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 & 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.

2.2 Methods

2.2.1 Terminology

SOM PPA terminology is sometimes confusing. In table 2.1 we aggregated model nomenclature used across the different studies cited in this paper, and chose one name for each feature.

2.2.2 The self-organizing map algorithm

A SOM is an artificial neural network first described by Kohonen in (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 2.1: SOM PPA terminology

Unified SOM PPA nomenclature		
Name	Short description	Other names
Input matrix	Matrix of regions and species to classify	Input layer, ocurrence 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

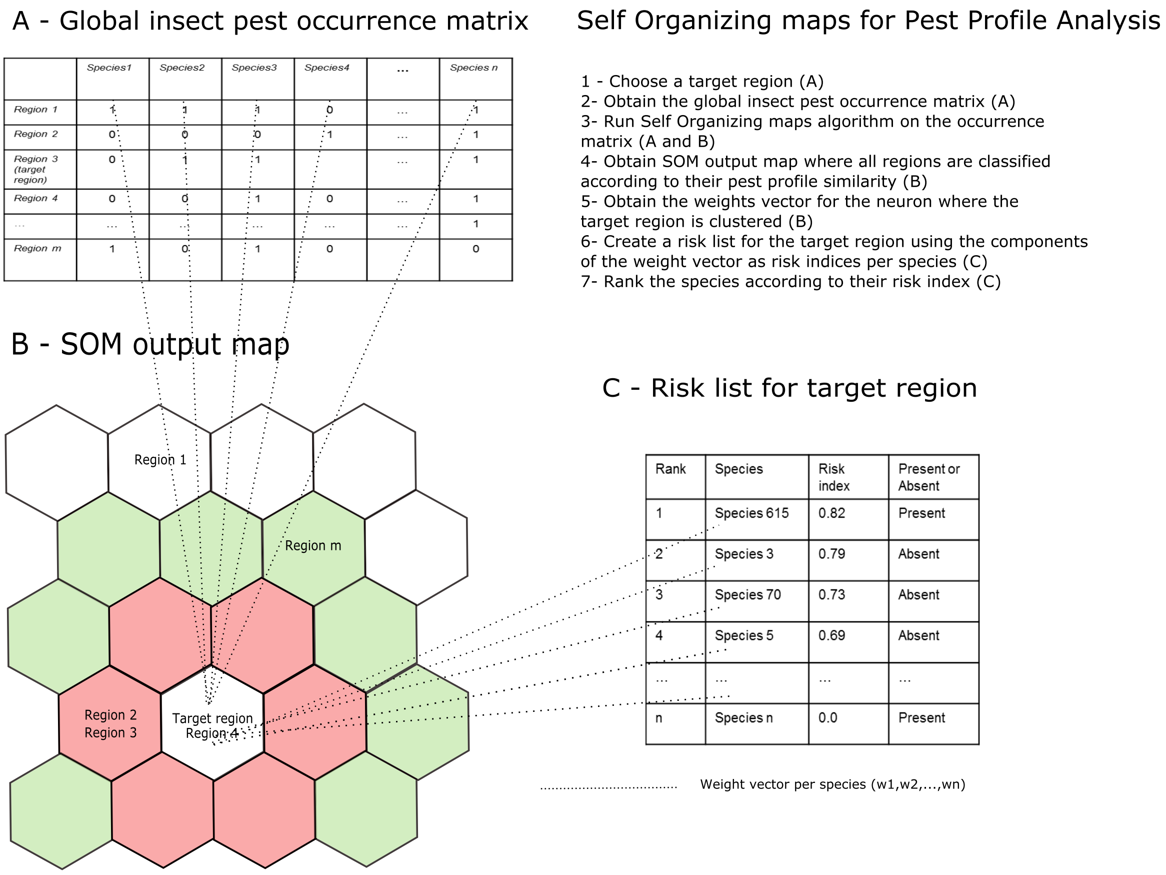
2.2.3 The SOM PPA

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.

Figure 2.1 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



A - Global insect pest occurrence matrix

	Species1	Species2	Species3	Species4	...	Species n
Region 1	1	1	1	0	...	1
Region 2	0	0	0	1	...	1
Region 3 (target region)	0	1	1		...	1
Region 4	0	0	1	0	...	1
...	...	...	...	...	...	1
Region m	1	0	1	0		0

Self Organizing maps for Pest Profile Analysis

1 - Choose a target region (A)
2- Obtain the global insect pest occurrence matrix (A)
3- Run Self Organizing maps algorithm on the occurrence matrix (A and B)
4- Obtain SOM output map where all regions are classified according to their pest profile similarity (B)
5- Obtain the weights vector for the neuron where the target region is clustered (B)
6- Create a risk list for the target region using the components of the weight vector as risk indices per species (C)
7- Rank the species according to their risk index (C)

B - SOM output map

C - Risk list for target region

Rank	Species	Risk index	Present or Absent
1	Species 615	0.82	Present
2	Species 3	0.79	Absent
3	Species 70	0.73	Absent
4	Species 5	0.69	Absent
...	...	...	...
n	Species n	0.0	Present

Figure 2.1: Diagram of SOM PPA process

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.

2.2.4 Data

We used the occurrence data extracted by Worner & 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 (*datasetA*) into ten occurrence matrices, one for each of the ten most common crops worldwide; apples, bananas, cotton, grapes, maize, mangoes, potato, rice, tomatoes and wheat (International Plant Protection Convention (IPPC), 2006b).

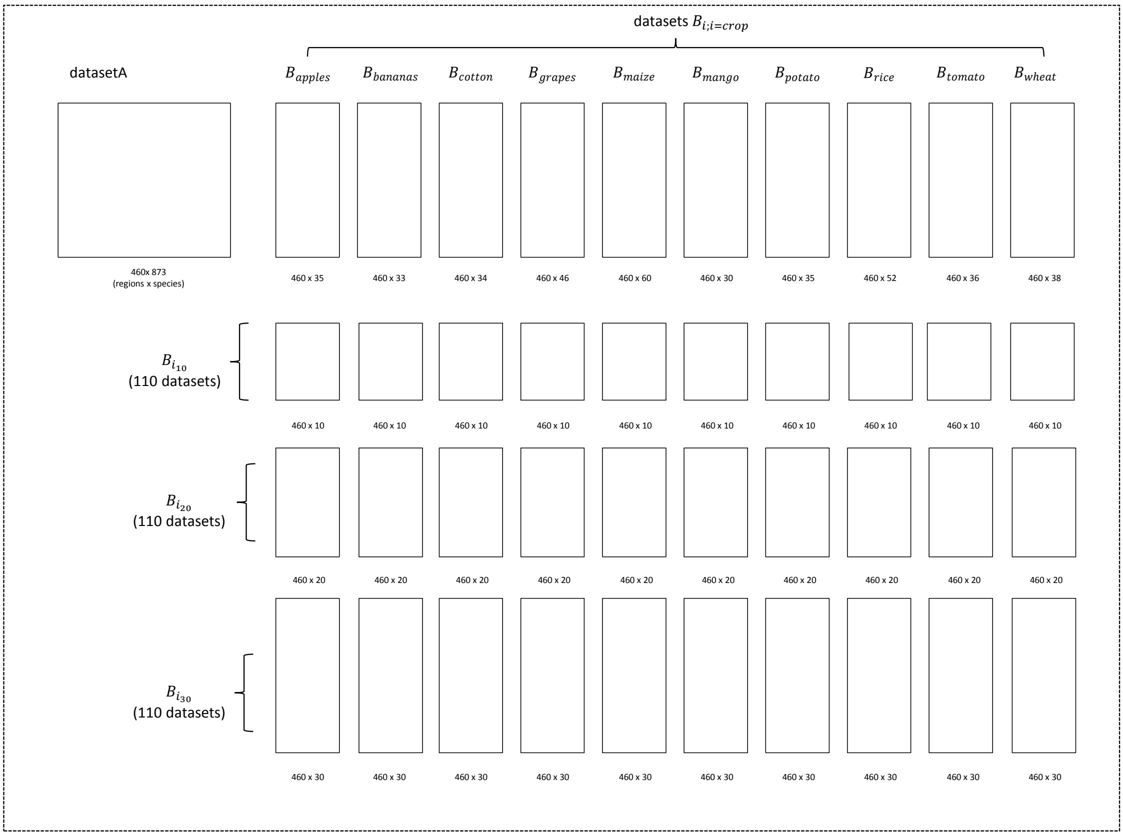


Figure 2.2: Details of datasets preparation and nomenclature

We named these crop restricted matrices $datasetsB_i$, where $i = crop$ (Figure 2.2). 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 the number of species in datasets $datasetsB_i$ was from 33 to 60 species. Finally, for each crop-restricted matrix, we created 33 species restricted matrices through random sampling. Thus, for every $datasetB_i$ we created 11 $datasetBi_{10}$, 11 $datasetBi_{20}$ and 11 $datasetBi_{30}$, where 10, 20 and 30 are the number of species in each, j. The result was a total of 341 datasets ($dataset\ A + 10\ datasetsB_i + 10 * 33\ datasetsB_{i_j}$). All data subsets contained 460 regions.

2.2.5 Weights sensitivity

To investigate the sensitivity of species weights and ranks to data dimensionality, we ran a complete SOM PPA for each of the 341 datasets for one target region (New Zealand) and created a ranked list of species from each run. The SOM initialization parameters were: 108 output neurons as given by the formula $c = 5\sqrt{(n)}$ (Vesanto & Alhoniemi, 2000) where c is number of output neurons and n is number of training samples; Gaussian neighbourhood distribution; linear initialization to ensure proper unfolding; and batch training mode as opposed to sequential training. The software used was SOM Toolbox (Vesanto *et al.*, 2000) for Matlab R2013b (Mathworks, 2013).

We recorded the weight obtained for each species across the 341 SOM PPA. Since not all species occurred in all datasets, we selected 199 species out of the 873 that were each present in 10 or more of the 341 datasets. Because the datasets were created using random sampling, these 199 species comprised an unbiased sample of the total 873. To examine variability of species weights, we calculated descriptive statistics for the 199 species and conducted a graphical exploratory analyses for each. We used R (R Foundation for Statistical Computing, 2014) for statistical tests and data handling.

2.2.5.1 Weights variability

Univariant descriptive statistics were calculated to summarize the central tendency and the variability a of the weight values for each of the 199 species. Boxplots of weights were generated for each species.

2.2.5.2 Weights sensitivity to dataset size

We used the package *lme4* (Bates *et al.*, 2015) to perform a linear mixed effects analysis of the relationship between weight and size of the dataset (number of species). We considered dataset size (number of species) as a fixed effect and species as a random effect. We obtained the p-values by likelihood ratio tests of the full model (with effect) against the model without effect (Bolker *et al.*, 2008). We chose linear relationships to model weight variability and weight sensitivity to dataset size because it was suggested by visual inspection of the weights distributions.

2.2.6 Ranks sensitivity

To evaluate the sensitivity of species ranks to changes in dataset size we analysed their average rank R^* and assessed its variability by computing some dispersion metrics such as their standard deviations and coefficients of variation.

2.2.7 Relationship between weight and species prevalence

We defined prevalence as the number of regions where a species was present. We graphically explored the relationship between prevalence and weight value for a single target region, and performed a linear regression to model the relationship. We expected to find that species with higher worldwide prevalence would generally have higher weights (Singh *et al.*, 2013).

2.3 Results

2.3.1 Weights sensitivity

2.3.1.1 Species weight variability between crop-restricted datasets containing different number of species

Across all species and crops, the mean weight was 0.137 and the median was of 0.024, which indicated a right skewed distribution (see histogram of weight values in A.1). Weight standard deviation was inadequate for measuring variability because it had a parabolic relationship with the mean (Figure A.3). This was unsurprising since weights are constrained between 0 and 1, thus their SD is less near the limits of their range values than in the middle.

To investigate the weight variation within species, we natural log transformed the weights and calculated the coefficient of variation (CV) for each species (figure A.4)(Nakagawa *et al.*, 2015). CV is the ratio of the standard deviation to the mean, thus is unitless. Mean CV across all species was 0.68, which indicated that, on average, weights varied 68% relative to their mean value. Within species, CV varied notably (see figures A.5 and A.6). Due to this within species variability, we used species as a random effect for modelling the relationship between weight and the number of species in the crop-restricted datasets.

2.3.1.2 Species weights sensitivity to datasets of different sizes

The linear mixed effects analysis of the relationship between weight and number of species showed weight was influenced by dataset size ($\chi^2 = 4.08$, p-value = 0.0432), with each additional species increasing it by 0.00002. The intra-species variability of the effect was of 0.2 SD. Species with low average weights showed less variation between datasets of different sizes than species with high average weights.

2.3.2 Ranks sensitivity

2.3.2.1 Species rank variability between crop-restricted datasets containing differing number of species

Many species occurred in few datasets, which made measurements of variation between ranks for these species uninformative. Thus, a subset of 674 species that had less than 10 appearances each across the 341 datasets (in other words less than 10 ranks) were excluded from the analysis. We studied average rank R^* variability for the remaining 199 species that occurred in ten or more datasets (as for weights).

Ranks for a given species usually varied between datasets. Average ranks were normally distributed (Shapiro-Wilk normality test: W= 0.9745, p-value < 0.0001, figure A.7), ranged from 24.88 to 869.40 and had a mean of 460.

As for weights, CV was used to measure variation in average rank because SD had a parabolic relationship with the mean (figure A.8 cf. figure A.3). Of 199 species, 91 had CV < 10%, 168 had CV < 20%, and only two species had CV > 50% (A.9).

Ranks were clearly less variable than weights, with lower CV. Figure 2.3 shows how their respective CV density functions differ.

2.3.3 Species prevalence

We defined species prevalence as the number of regions a species occurred in. Some species plots suggested that species with widespread distributions (e.g. the most widespread pest, *Aphis gossypii* which occurred in 250 regions) tended to have higher weights. This was consistent with some previous studies (Singh *et al.*, 2013). However, a plot of prevalence

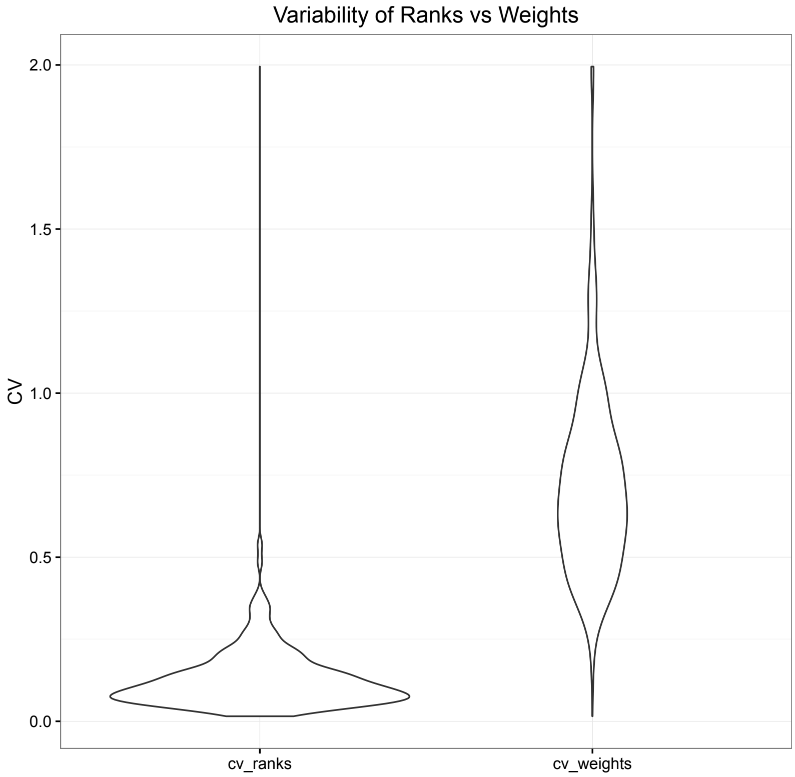


Figure 2.3: Comparison of CV density distribution for ranks and for weights

against weight exhibited no relationship, even after applying logarithmic transformations to the data (Figure A.10). Linear regression also showed no significant relationship (F-statistic 0.1893 on 1 and 871 DF, p-value=0.6636). This indicated that species ranked highly for the target region (New Zealand) were not necessarily those with the largest geographical distributions.

2.4 Discussion

The main aim of this study was to investigate the sensitivity of SOM PPA species weights and ranks to changes in input data. We challenged SOM PPA by creating many different datasets, conducting 341 SOM PPA, and quantitatively investigating how species weights and ranks changed with them. Creating numerous datasets and randomly assigning species to datasets rigorously tested the stability of species weights and ranks when input data varied. As expected, species' weights and ranks both varied, but weights were much more variable than ranks.

2.4.1 Weights vs ranks sensitivity

For the same subset of 199 species, weights had a mean CV of 68% while ranks had a mean CV of 12% (figure 2.3). Due to high variability of weights, we recommend the use of ranks rather than weights in SOM PPA. Another problem using weight as a measure of invasion risk arises from its right skewed distribution, which means few species have high weights and many have low weights. If an analyst naively chose a value of weight to use as a threshold for separating high and low risk species, it would be easy to choose a threshold weight that was too high, meaning many potentially high risk species would be overlooked. This problem is less likely to arise if ranks are used because they are more normally distributed.

2.4.2 Regions with full zero profiles

When working with presence/absence data it is often difficult to identify outliers. A common practice in SOM PPA literature (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Paini *et al.*, 2010b, 2011) is to exclude from analysis those regional profiles that have few species present (usually less than 5) to avoid distorting SOM multidimensional space. This assumes that species-poor regions are under sampled regions and their species records are unreliable. However, we retained regions with no species present in our analysis for the following reasons. Species-poor regions will not always be undersampled and many zeros will be true absences. Thus, they are as meaningful as ones, and excluding true absences can bias SOM results at least as much as including false absences. Even though they are coded as ones and zeros they are a categorical variable, therefore, excluding one category from the analyses has the potential to seriously bias the results (Millar *et al.*, 2011). Moreover, our study aimed to test the sensitivity of SOM PPA to variation in input data, and including regions with no species assisted this testing.

2.4.3 Dataset dimensionality

Our results showed that the sensitivity of weights to the dataset size is species dependant. Although we expected that smaller datasets would yield different potentially lower weights, values of the weight increased only by 0.00002 for each additional species. Thus there will be negligible difference between results from data sets that vary in size by tens of species, or possibly several hundreds.

Our observation that dimensionality has a weak influence on species weights indicates that SOM PPA could be used to study theme restricted matrices such as the crop-restricted datasets used in our analysis. Some other possible examples include datasets restricted by region and taxon. The use of restricted matrices was suggested in Vänninen *et al.* (2011) both for assessing potential to cultivate crops in new regions and for screening 'thematic' lists of potential invaders. We suggest SOM PPA could also be used to assess risks from species associated with particular commodities. It is common practice in the international regulatory framework to initiate a pest risk analyses based on a proposal either to begin importing a new commodity into a recipient region, or to begin accepting a previously imported commodity from a new donor region. This requires analysts to gather data on species associated with the commodity in the donor region and study them further.

SOM PPA is a quantitative tool for analysing complex data, but by no means consitutes a complete pest risk analysis. By creating crop-restricted datasets we have simulated a realistic scenario and tested how SOM PPA would perform over a range of different sized datasets. We chose the 10 most common worldwide crops as recommended by Knight *et al.* (2011) but the results are transferable to other potential dataset restrictions. We have shown how, overall, species ranks are in general stable metrics, and species weights are not related to dataset size. Additionally, real data was used instead of simulated data because the latter approach had already been used by Paini *et al.* in (2011). Our results showed that ranked list of species are more suitable for commodity based risk analyses, due to the stability of the ranks.

2.4.4 Species Prevalence

We found no evidence that species weight is related to species prevalence. Many studies using SOM PPA have stated or implied that widespread species are often highly ranked in the final list (Knight *et al.*, 2011; Singh *et al.*, 2013; Worner *et al.*, 2013) and there was uncertainty about whether prevalence could influence SOM PPA results. However, Watts & Worner (2009) showed in their studies how prevalence and rank are unrelated, and our results confirmed this. The conclusion is robust because our results, and those of

Watts and Worner (2009), were obtained from large sample sizes and numerous SOM PPA replications.

The implications of this finding for SOM PPA are important. Gevrey *et al.* (2006) recommended removing species with low prevalence from input matrices to avoid biasing SOM classifications. However, we argue this is unnecessary because the rank a species will obtain for the target region is unrelated to its prevalence. This means less data preprocessing and less modeller-induced bias arising from subjective decisions about prevalence levels to use as thresholds for exclusion. Our results also indicate that SOM PPA can be reliably used to rank pests with restricted distributions.

Finally, the fact that prevalence is not related to the weight value or the rank for our specific target region does not mean that widespread species will not obtain higher ranks across all target regions. Widespread species will often be generalist feeders with potential to establish in any region where one or more hosts are present (Lewinsohn *et al.*, 2005).

2.5 Conclusions and Future work

The sensitivity analysis of SOM PPA over 341 crop restricted datasets has clearly shown how SOM ranks are stable metrics that respond consistently to a series of extreme changes to the input datasets. Also, we highlight the possibility of using SOM PPA for themed restricted datasets, such as crop-restricted, that may be very useful for certain types of risk assessment or host-based risk assessment.

As a consequence of this research, we can provide new guidelines for SOM PPA; it is not necessary to delete species with low prevalence nor to control for data dimensionality. Thus, future SOM PPA users can employ small themed and restricted datasets of only few species to run the pest profile analyses, thereby reducing the amount of modeller-introduced bias in the approach.

However some questions remain. While this study has investigated the stability of the weights and ranks, it has not determined the validity of weights and ranks as measures of establishment likelihood. While other studies have discussed this issue, more research in this direction is required. A further improvement of the method would be to incorporate an uncertainty or confidence measure in the output, for example a cluster validity metric.

Additionally, this study used a single target region for comparability purposes among all datasets, but could be reproduced for any other target regions as a form of validation. That would help to increase our understanding of the issue encountered by Paini *et al.* (2011) where pest profiles with few species were hard to classify.

Finally, given the variability of weights is higher than ranks, we recommend using a ranked list of species as an output. Also, a ranked list of species gives a better idea of prioritization while a list of species with a weight value between 0 and 1 can be easily mistaken for a probability. Finally, we have shown the skewed weight value renders the numerical weight uninformative.

Chapter 3

Cluster validity and uncertainty assessment for self organizing map pest profile analysis

Acknowledgement

This paper was conceived during a collaborative stay at Monash University in 2014, funded by the Bio Protection Research Centre travel grant 2014. This manuscript has benefited from comments by Audrey Lustig, Ursula Torres, Maria del Mar Rovira Capdevila, Michal Kuchár, Gabriel Moinet, Marta Gallart, Matthew Parry, Craig Phillips and Simon Leboulh. M.R received financial support from the European Commission (Erasmus Mundus partnership NESSIE, ref. 372353-1-2012-1-FR-ERA MUNDUS-EMA22) and Better Border Biosecurity (B3)(www.b3nz.org) research collaboration. MM acknowledges support from Australian Research Council grant DP150103017. The authors state no conflict of interests.

Notes

This chapter is published as:

Roigé, M., McGeoch, M. A., Hui, C. and Worner, S. P. (2017), Cluster validity and uncertainty assessment for self-organizing map pest profile analysis. Methods in Ecology and Evolution, vol.8 pages 349âĂŞ357.

Abstract

1- Pest Risk Assessment (PRA) comprises a set of quantitative and qualitative tools to protect productive ecosystems from the impacts of unwanted biological invasions. 2- Self-organizing maps for Pest Profile Analysis (SOM PPA) is a methodological approach aimed to support PRA. It is based on cluster analysis and extracts information out of current distributions of insect crop pests worldwide, allowing the analyst to generate a list of potential risk species for a target region. 3- SOM PPA currently lacks of a measure of performance able to provide a level of confidence for its outputs. 4- In this study we investigate ζ diversity as an ecologically meaningful and generalizable metric of similarity. The application of ζ allowed us to quantify and thus reveal different levels of similarity across pest profiles. 5- The use of ζ diversity applied to the SOM PPA provides an informative measure of uncertainty for the output of SOM PPA, thus adding major improvements to the methodology while only marginally increasing its complexity.

Keywords

Self organizing maps, Pest profile analysis, Zeta diversity, Uncertainty, Cluster validity, Pest Risk Assessment

3.1 Introduction

Economic globalization and increasing trade present numerous challenges to both natural and productive environments. Along with climate change and loss of biodiversity, biological invasions have gained increasing attention because of their impact on social (Díaz *et al.*, 2006), natural (Walther *et al.*, 2009; Hulme, 2009; Perrings *et al.*, 2005; Vilà, 2013; Vitousek, 1990) and productive ecosystems (Mack *et al.*, 2000; Simberloff, 2011; Horan & Lupi, 2010). There are nine regional organizations under the umbrella of the International Plant Protection Organization (IPPO) which have the responsibility to implement trade standards to help protect global agriculture. In 2006, United Nations Food and Agriculture Organization (FAO) issued a set of recommendations and regulations that serve as a global standard for assessing potential invasive threats, called International Guidelines for pest risk analysis (PRA) (International Plant Protection Convention (IPPC), 2006b).

The initiation stage of a pest risk assessment consists of identifying potential pests. For that, risk assessors prioritize time and resources creating lists of species according to the risk of pest entry, establishment, spread and impact (Worner & Gevrey, 2006; Worner *et al.*, 2013; Venette *et al.*, 2010; Leung *et al.*, 2012). Worner and Gevrey (2006) first described self-organizing maps (SOM) for pest profile analyses (referred to here as a SOM PPA) as a quantitative tool aimed to assist the initiation stage of a PRA by producing a list of species ranked according to risk values that indicate the potential threat they present to the region under consideration. Briefly, a regional pest profile is the information for a region about the presence or absence of global insect crop pests and a SOM is an artificial neural network used to analyse and cluster high dimensional data (Kohonen, 1982) that has been successfully applied to ecological sciences across many scales (Chon, 2011). In SOM PPA, regional pest assemblages are clustered to identify potential pest donor and recipient regions. Clearly the processes of introduction and establishment, as well as the abundance and spread of invasive insects are intricate parts of population dynamics (invasion performance) driven by propagule pressure, the allee effect, stochasticiy, and intraspecific competition, but we are far from having a mechanistic understanding of how all these effects work, let alone how they interact (Leung *et al.*, 2012; HilleRisLambers *et al.*, 2012). Especially for invasive species, abiotic factors, historical processes (Lewinsohn *et al.*, 2005), trade and agricultural production have a great effect on the composition of the assemblage (Worner, 2002; Worner & Gevrey, 2006). This application of SOM to pest profile analysis is based on the assumption that all these complex interactions of biotic and abiotic factors are integrated into a region's resulting pest assemblage (Worner & Gevrey, 2006; Gevrey *et al.*, 2006). A regional pest assemblage is formed by the co-occurrence of insect crop pests and indicates suitable environmental conditions, the presence of suitable hosts (Bebber *et al.*, 2014a) and a particular invasion history of the region (Worner *et al.*, 2013). It follows that two regions with similar assemblages are likely to be donors or recipients of species from each other. A species present in a region with a very similar pest profile to the target region, is highly likely to be able to establish in the target region given the chance to do so (Worner & Gevrey, 2006), therefore, assemblage similarity has been used as a proxy for establishment risk (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Watts & Worner, 2009; Paini *et al.*, 2010a, 2011; Morin *et al.*, 2013; Singh *et al.*, 2013).

Since the risk values generated by SOM PPA are influenced by the individual cluster where the target region is allocated, good clustering is fundamental to obtain meaningful and interpretable risk lists. When performing cluster analysis, validity refers to the process of evaluation of the resulting clusters (Halkidi *et al.*, 2001). For a SOM, there are two measures of goodness. They are the quantization error (Q_e) and the topological error (T_e) (Kohonen, 1990; Vesanto & Alhoniemi, 2000; Vesanto *et al.*, 2000; Uriarte & Martín, 2006). These two metrics are useful to describe the overall performance of the algorithm on the dataset, but they do not provide any specific validity measure for each cluster. Furthermore, out of the many applications of SOM PPA, none has assessed cluster validity of the results (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Paini *et al.*, 2010a, 2011; Watts & Worner, 2011, 2009; Singh *et al.*, 2013, 2015; Morin *et al.*, 2013). In Paini *et al.* (2010b), they calculated the 'percentage similarity in insect assemblage' only between the target region (Australia) and the other regions clustered with it. Even though this approach is informative, some sort of validity measure needs to be computed for every cluster, to assess their relative performance. Therefore, SOM PPA would benefit from containing an integrated cluster validity assessment to support the reliability of its outputs.

In addition to the need for a cluster validity measure SOM PPA requires some form of uncertainty assessment. Quantitative tools are subject to a certain degree of uncertainty (Maier *et al.*, 2008), which can be reducible or irreducible (Regan *et al.*, 2002; Ascough *et al.*, 2008). Irreducible uncertainty in environmental modelling should be acknowledged, quantified and communicated (Walker *et al.*, 2003; Wattenbach *et al.*, 2006). However, SOM PPA outputs do not incorporate any form of uncertainty communication.

Here, because of its ecological relevance and simplicity, we propose zeta ζ diversity (Hui & McGeoch, 2014) as both a potential cluster validity measure for SOM PPA and the use of this cluster validity assessment as a means of communicating the uncertainty associated with the analysis. In a biological community, diversity can be measured by total richness or gamma (γ) (Whittaker, 1972), which is the total number of species. Gamma diversity can, in turn, be partitioned into alpha (α), which is the number of species in each site or local species richness, and beta (β) which refers to either to a compositional species variability or to species turnover along an environmental, spatial or temporal gradient (Whittaker, 1972; Vellend, 2001; Anderson *et al.*, 2011; Jost *et al.*, 2011; Tuomisto, 2010). We use ζ, defined as the number of species shared between any i sites, as a measure of compositional diversity because it is a simple and direct measurement that reconciles existing descriptors of species incidence and spatial turnover (Hui & McGeoch, 2014).

Our primary objective in this paper is to improve the current SOM PPA methodology by incorporating a cluster validity measure. A secondary objective is to show how the cluster validity measure can also serve as an estimate of output uncertainty.

We exemplify this by applying a SOM PPA analysis to a global occurrence dataset and then evaluating the results using ζ as a measure of compositional similarity. With the aim of harmonizing terms and concepts in uncertainty communication (Uusitalo *et al.*, 2015b), we propose the use of the cluster validity assessment provided by ζ as a strategy for communicating the uncertainty associated with the output of SOM PPA.

3.2 Methods

3.2.1 Data

Data used in this study were originally extracted by Worner and Gevrey (Worner & Gevrey, 2006; Gevrey *et al.*, 2006) from the Crop Protection Compendium (CPC) (CABI, 2007) and organized in a matrix of 452 sites (rows or regions) and 873 species (columns) referred to, in this study, as the global pest occurrence matrix. This matrix contained all geographical distributions of phytophagous insect pests considered of relevance for global crop protection by the plant protection organizations which comprise the CPC consortium. Presences were coded as 1 and absences were coded as 0. The geographic areas represented in the Compendium consisted of countries, regions or states of countries, all of different size. In a few cases, large countries appeared more than once in the matrix, both as a whole and also divided into their states. As expected, smaller regions contained, in general, fewer species than larger regions, following a classical species area relationship (SAR) (He & Legendre, 1996).

3.2.2 Self-organizing maps for Pest Profile Analysis: SOM PPA

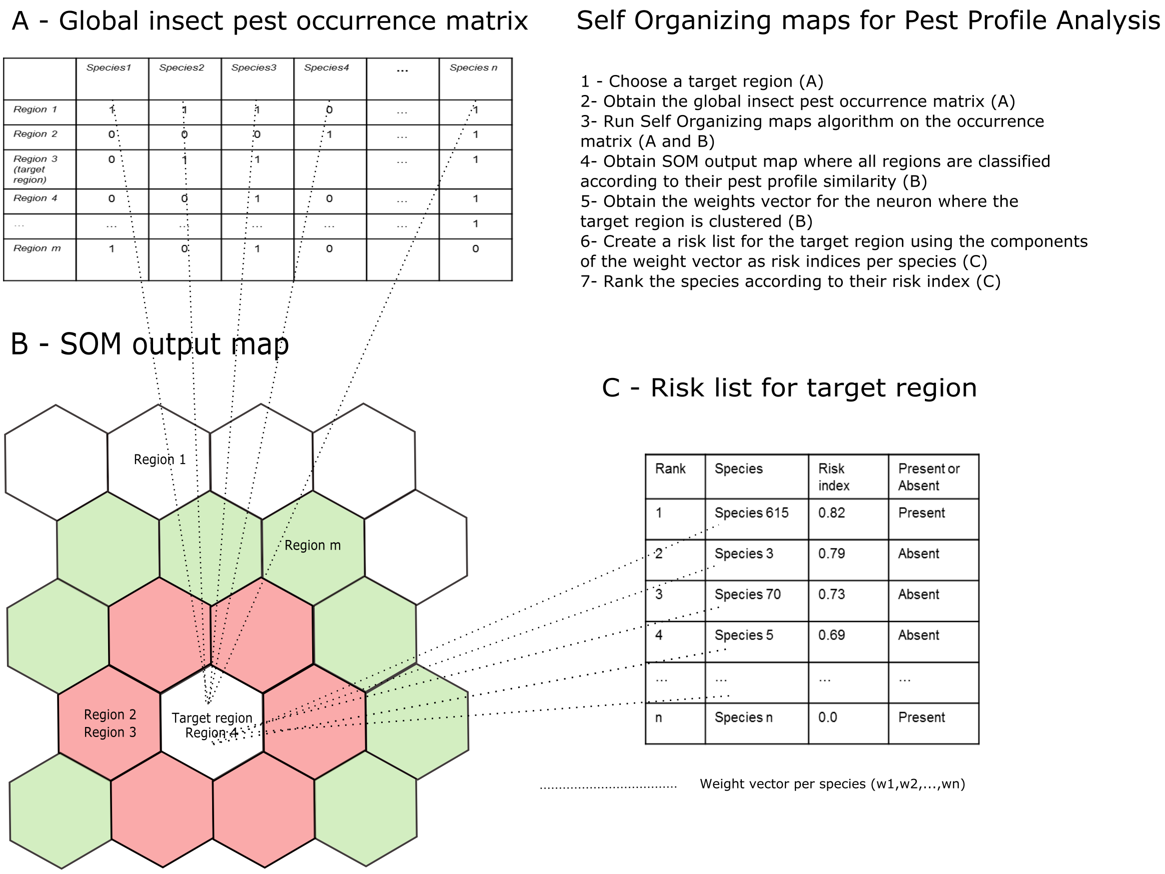
The use of self-organizing maps for pest profile analyisis (SOM PPA) (Worner & Gevrey, 2006; Gevrey *et al.*, 2006) can be broken down into two steps (Figure 3.1).

First, the SOM algorithm performs a process of ordination, vector quantization and vector projection of the regional pest profiles (rows) of the global occurrence matrix. Through a machine learning process which involves two layers of an artificial neural network (Kohonen, 2001, 2013), the SOM algorithm, 1) captures the similarity relationships within the regional profiles of the global occurrence matrix (input layer), 2) distributes them in a multidimensional space according to their similarity (vector quantization), and 3) projects them onto the SOM output map (output layer) where the regional profiles are classified in such a way that their similarity to one another is maximized (vector projection). In other words, the SOM algorithm converts the highly dimensional pest occurrence matrix (each species equals one dimension) into a two dimensional map of clusters (ordination). Regional pest profiles close together in the output map are more similar than those far apart. The output map is arranged in neurons, sometimes referred to as cells, and since the number of neurons is less than the number of regional pest profiles in the input matrix, multiple regional profiles are mapped onto a smaller number of output neurons, creating clusters of regions.

Let n be the number of species in the occurrence matrix. Each neuron of the map has a coordinates 'weights vector' of length n which defines its position in the n-dimensional space created by the algorithm. Since each species is a dimension, each component of the weight vector reveals the strength of association between a neuron (and the regional profiles clustered in it) and each species, thus representing the importance of the species in the classification process. Ecologically, the weight vector component is interpreted as the strength of association between the regional assemblage and the particular species (Gevrey *et al.*, 2006; Paini *et al.*, 2010b), and because occurrence data are binary, weight values are constrained between 0 and 1.

The second step in SOM PPA is to create a species risk list for the region of interest. Once the SOM algorithm has clustered the regional profiles, the neuron to which the target region is allocated is identified and the weights vector of that neuron is extracted. Species are then ranked according to their respective weight vector component, thus creating the risk list.

Since the weight value ranges between 0 and 1, species with weights closer to 1 are assumed to be more strongly associated with the pest assemblage of the target region. Therefore a species with high weight value that is absent from the target region is assumed to be likely to establish if they ever have the chance to do so, that is, if a pathway exists and there is a viable propagule size. (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Watts & Worner, 2009, 2011; Paini *et al.*, 2010b,a, 2011; Singh *et al.*, 2013; Morin *et al.*, 2013).



A - Global insect pest occurrence matrix

	Species1	Species2	Species3	Species4	...	Species n
Region 1	1	1	1	0	...	1
Region 2	0	0	0	1	...	1
Region 3 (target region)	0	1	1		...	1
Region 4	0	0	1	0	...	1
...	...	...	...	...	...	1
Region m	1	0	1	0		0

Self Organizing maps for Pest Profile Analysis

1 - Choose a target region (A)
2- Obtain the global insect pest occurrence matrix (A)
3- Run Self Organizing maps algorithm on the occurrence matrix (A and B)
4- Obtain SOM output map where all regions are classified according to their pest profile similarity (B)
5- Obtain the weights vector for the neuron where the target region is clustered (B)
6- Create a risk list for the target region using the components of the weight vector as risk indices per species (C)
7- Rank the species according to their risk index (C)

B - SOM output map

C - Risk list for target region

Rank	Species	Risk index	Present or Absent
1	Species 615	0.82	Present
2	Species 3	0.79	Absent
3	Species 70	0.73	Absent
4	Species 5	0.69	Absent
...	...	...	...
n	Species n	0.0	Present

Figure 3.1: Diagram of SOM PPA. Regions clustered in red neurons are more similar to the target region than regions clustered in green neurons

3.2.3 Zeta diversity (ζ)

There are many indices that measure the spatial dimensions of species richness and compositional similarity, most of which represent slightly different natural phenomena (Vellend, 2001; Tuomisto, 2010). For this particular study, the following definitions apply. In a study with i regional profiles, gamma diversity (γ) is the total species counts across the i profiles, alpha diversity (α_i) is the local species richness in profile i (Whittaker, 1972), and zeta ζ is the average number of species shared by i assemblages or regional profiles (Hui & McGeoch, 2014), which can be pairs ($i = 2$), triplets ($i = 3$) or larger groups.

The first order zeta diversity (ζ_1) is the average number of species of all the assemblages or profiles at each site (average species richness, or $\bar{\alpha}$) ζ_2 is the average number of shared species between any two regional profiles, ζ_3 is the average number of shared species among three profiles and so on, until ζ_i, the average number of shared species by all i regional profiles. Therefore the value of ζ monotonically declines with i, and the form of its decline (zeta decline) is considered to be ecologically meaningful (Hui & McGeoch, 2014). For our case study we computed the values of γ, α and ζ_i and interpreted their implications in the context of SOM PPA. To perform these analyses we used the packages *zetadiv* (Latombe *et al.*, 2015) and *vegan* (Oksanen *et al.*, 2016) in R (R Foundation for Statistical Computing, 2014).

3.2.4 Case study: Assessing SOM PPA cluster validity using ζ diversity

We ran a SOM PPA for the pest occurrence matrix. We built a map of 12 x 9 neurons in a hexagonal lattice. We used 'batch' training for the algorithm, a gaussian neighborhood relationship, 5000 iterations as training length and obtained a quantization error of 5.85 and a topological error of 0.01 (Vesanto *et al.*, 2000; Vesanto & Alhoniemi, 2000). The SOM analysis was performed using Matlab (Mathworks, 2013) and the package SOM Toolbox (Vesanto *et al.*, 2000).

Once the SOM output map was created, we considered each neuron of the map as one study and systematically computed all its diversity components (see minimal working example in B.2). Gamma γ was the total number of species present in all the regions clustered in a neuron. Alpha was calculated as the individual species richness measures for each one of the regional assemblages in a neuron, and finally we computed ζ_1 to ζ_i, as the average number of species present in i number of assemblages or regional profiles in a neuron.

3.2.4.1 Using normalized ζ as a clustering validity measure

Since ζ describes the number of shared species between i profiles, its value is highly dependent on the region's total species richness. Given that the objective was to use ζ as a value to compare cluster validity amongst neurons and richness can vary substantially across neurons, we computed its normalized version, ζ_i/ζ_1, which can be simply interpreted as the proportion of species in common between i regional profiles.

3.2.4.2 Other similarity metrics

To justify the choice of ζ over other diversity measures, we compared it to a classic measure of β diversity, Sorensen's index (Sørensen, 1948). We computed Sorensen's index using the function vegdist() of the R package *vegan* (Oksanen *et al.*, 2016), which provided Sorensen's dissimilarity, which we then inverted to obtain Sorensen's similarity, (Sorensen similarity = 1- Sorensen's dissimilarity).

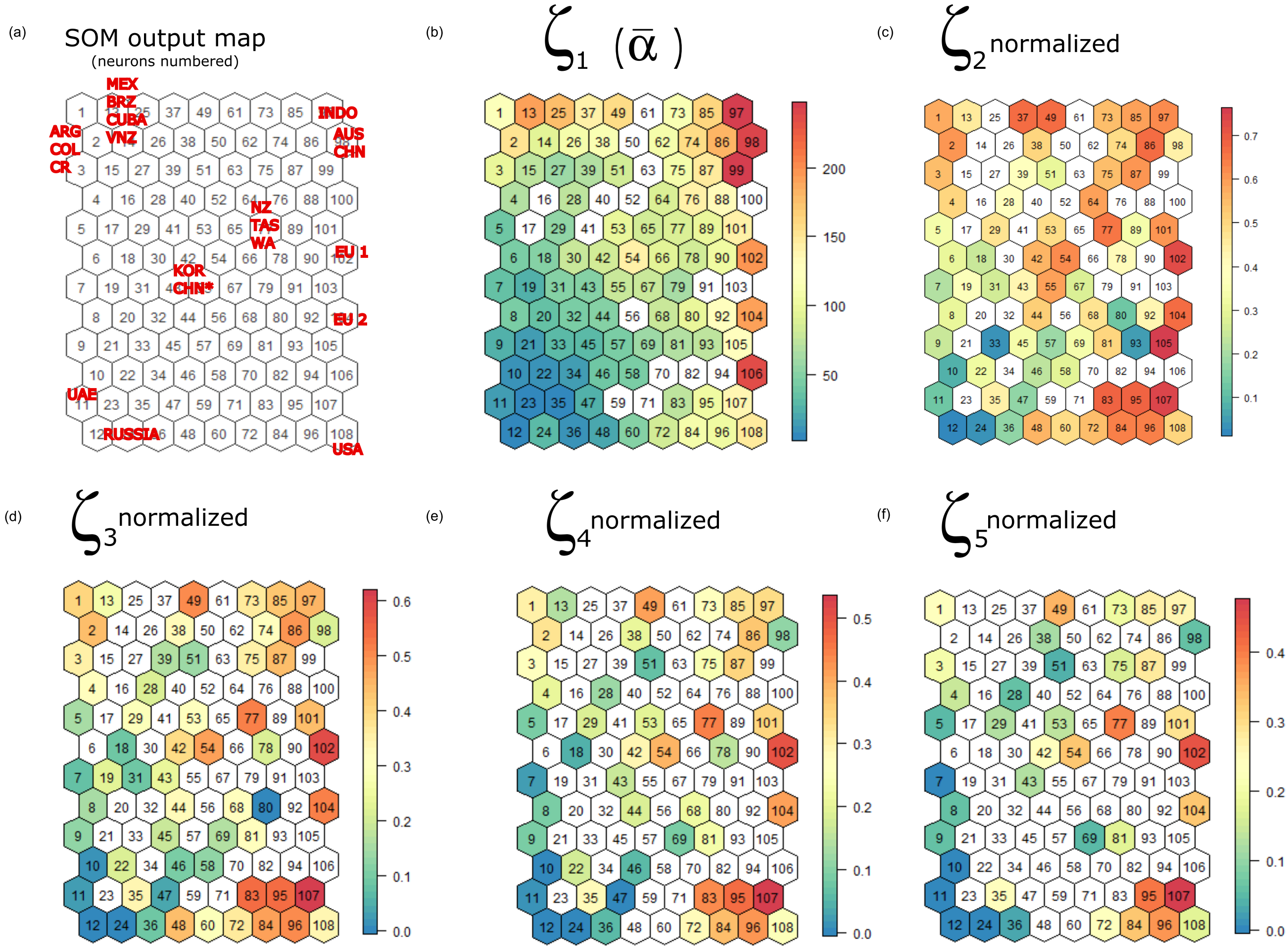
3.2.4.3 Using normalized ζ as an uncertainty measure

To use ζ as an estimate of output uncertainty, we classified the neurons in the SOM output map according to their ζ values. We looked at three different ζ scenarios per neuron. We expected to differentiate neurons with low pairwise similarity ζ_2, neurons with good pairwise similarity ζ_2 but not good groupwise similarity ζ_{3-5}, and neurons with high pairwise similarity and high groupwise similarity.

3.3 Results

3.3.1 General overview of world regions on the SOM output map

The SOM algorithm produced a widely spread classification of the 452 regions across the output map. Figure 3.2a shows the location of some representative world regions. The USA and its states were located at the bottom right corner, in neurons 48, 60, 72, 84, 96, 108, 107, 95, 103, along with Canada in neuron 106. On the right edge, a long list of European and other Mediterranean regions (Algeria, Morocco, Turkey) were located in neurons 104 and 102. A cluster of Middle Eastern countries was located nearby, in neuron in neuron 101. Noticeably, seven of the world's most populated assemblages: Australia, China, India, Japan, Indonesia and Thailand (along with many other Indian provinces, and south-east Asian regions) were allocated together on the top right of the map (neurons 98 and 97). African sub-Saharian regions, (including South Africa) were mapped onto the top edge (neurons 37, 49 and 38). The upper left corner (1, 2, 3, 13) contained Central and South American regions, along with some Caribbean islands. The bottom left corner comprised a large number of regions for which there was not a clear grouping. The most prominent characteristic amongst this last group was probably the number of small islands in it. However, regions such as Greenland, Kuwait or United Arab Emirates were also located in this bottom left corner. At the centre of the map (neurons 53,42,54,55) several provinces of China, North and South Korea were grouped together. And in neuron 77, New Zealand was located at the right half of the map alongside South Australia, Tasmania, Victoria and Western Australia.



ARG - Argentina, COL- Colombia, CR- Costa Rica, MEX- Mexico, BRZ- Brazil, VNZ- Venezuela, UAE- United Arab Emirates, KOR- Korea, CHN*- Provinces of China, NZ- New Zealand, TAS- Tasmania (Aus), WA- Western Australia, INDO- Indonesia, AUS- Australia, CHN- China, EU1- Spain Greece, Portugal, EU2- Austria, Belgium, Bulgaria, Switzerland, Germany, former Czechoslovakia, Denmark, France, UK, Hungary, Italy, Netherlands, Poland, Romania, Sweden.

Figure 3.2: ζ decline in SOM map neurons. (a) shows SOM map with some regions located according to their regional insect pest profiles. (b) shows ζ_1 (average species richness of all the regions clustered in a neuron). (c) corresponds to normalized ζ_2 per neuron (average proportion of species any two regions in that neuron share). And (d,e,f) correspond respectively to normalized ζ_3, normalized ζ_4 and normalized ζ_5

.

3.3.2 Case study: Assessing SOM PPA cluster validity using ζ diversity

Values of ζ_1 revealed the areas of the SOM output map with, on average, richer regional profiles (Figure 3.2b). These were the top (1, 15, 25, 37, 49, 61, 73, 85, 97) and right edge (97 to 108) neurons. The distribution of ζ_1 across the SOM output map showed a clear gradient of species richness. The bottom left area of the SOM map grouped together regions with low species counts while the top right grouped regions with high species counts. However, low species counts did not imply low values of ζ_{2-5}. The plot of ζ_1 showed a clear diagonal gradient that was not found in plots ζ_{2-5} (Figure 3.2).

We normalized ζ to account for the effect of richness (ζ_i/ζ_1). Normalized ζ_2 (Figure 3.2 c) was high across the map except in the bottom left corner where a high number of small islands were clustered. Richness was not high either for neurons 80 and 92, which comprised the regions of Georgia, Russia, Serbia and Montenegro, Ukraine and former Yugoslavia. The remaining neurons showed a ζ_2 value higher than 0.4, which meant pairs of regions clustered in these neurons shared, on average, 40% of their species. Neurons 105 (Finland and Norway), 107 (States of the USA) and 102 (Mediterranian European regions) showed very high ζ_2 levels, with approximately 70% of pairwise shared species.

The more interesting finding was that for some neurons, similarity remained high with increasing ζ orders (Figure 3.2 d,e,f) while for some neurons, relative similarity decreased for orders ζ_{3-5}. For example, neurons 49 (Sub-Saharian African regions), 77 (New Zealand and Southern Australian regions), 101 and 102 (Middle-East and Mediterranean regions) 104 (European regions) and 83, 95, 107, 84, 96 (United States) are neurons that obtained high ζ_2 values and maintained high values for ζ_{3-5}. In contrast, neurons 85, 97, 86 and 98 showed very high similarity according to ζ_2 but this relative similarity faded with higher orders of ζ. This last group of neurons are precisely the ones that contain the highest species richness regions (Australia, China, India, Japan, Indonesia Thailand and several Indian provinces).

It is also important to note that neurons with low normalized ζ_i values were mostly grouped in the bottom left corner: neurons 10, 11, 12, 24, 36, 33, 80 and 92, grouped in the same area of the map where we found extremely low α values. For example, neuron 12 had a very low total richness $\gamma_{12} = 162$, lower than the average $\bar{\gamma} = 172.1$, especially considering it had the highest number of regions allocated, $i_{12} = 38$. Such a large number of regions containing such a a small number of species suggested that neuron 12 was populated by regional profiles with very low species counts. This was confirmed by looking at the ζ_1 plot, where neuron 12 had one of the lowest values (in blue). Neurons 10, 11, 24 and 36, which all neighboured neuron 12, had a small number of regions allocated, but still low values of ζ_1. These observations suggest that SOM algorithm allocated the 'odd' profiles, or the regional profiles that had no common patterns with the rest, into the bottom left corner of the map. Similarly, neurons 80 and 92 (Georgia, Russia, Serbia and Montenegro, Ukraine and Former Yugoslavia) even though they were not allocated to the bottom left corner, also shared the characteristic of having very low species counts. The same applies for neuron 33 (Crete and Arunchal Pradesh), where average species richness (ζ_1) was quite low.

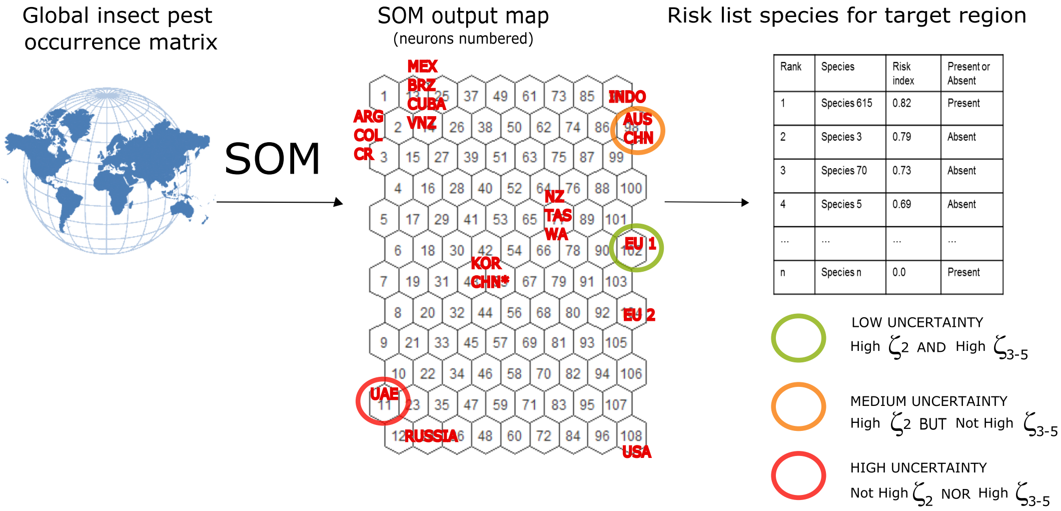
3.3.2.1 Other similarity metrics: Sorensen similarity index

The Sorensen similarity index (1-Sorensen) and $Zeta_2$ (ζ_2) were confirmed to be equivalent measures (Figure in B.3). Values for both indices were identical up to the fourth significant figure except for cells 94, 87, 88 an 93, for which, because of lack of sufficient number of regional assemblages, one of the two indices could not be computed.

3.3.3 Using normalized ζ as an output uncertainty measure for SOM PPA

We showed how using ζ allowed us to differentiate neurons from one another. Figure 3.3 summarizes these findings. Some neurons showed questionable clustering (low ζ_2), whilst others showed reliable clustering (neurons which maintained high values along higher ζ orders). Some other neurons showed high initial similarity (high ζ_2) that faded with higher orders. Undoubtedly, ζ provided crucial information about the clustering goodness in the SOM output map.

Consequently, we propose to incorporate ζ into the SOM PPA approach by using three levels of assemblage similarity that emerged from the study of ζ_i (Figure 3.3). Case 1) If the neuron where the target region is allocated has a very low relative ζ_2 value



Rank	Species	Risk index	Present or Absent
1	Species 615	0.82	Present
2	Species 3	0.79	Absent
3	Species 70	0.73	Absent
4	Species 5	0.69	Absent
...	...	...	...
n	Species n	0.0	Present

Figure 3.3: Scheme of the three levels of uncertainty for the output risk list that emerged from the ζ values on the SOM output map

compared to the other neurons in the SOM output map, the risk list computed by SOM PPA is highly uncertain and should be interpreted with extreme care. Case 2) If the target region is clustered in a neuron with high ζ_2 but not high groupwise similarity ζ_{3-5}, that list can be interpreted, albeit with care, since it is based on assemblage similarities that are, to a degree, superficial. Case 3) When the target region is allocated into a neuron with high ζ_{2-5}, the list of potential risk species can be interpreted as the assemblage similarity is high both superficially (pairwise ζ_2) and for higher ζ orders, assuming Worner and Gevrey's (2006) hypotheses that assemblage similarity can be used as a proxy for establishment risk. As an example, the risk list produced for New Zealand by Worner & Gevrey (2006) would obtain a low uncertainty level according to ζ, since neuron 77 has consistently high levels of normalized ζ_{2-5} (Figure 3.2). On the other hand, a risk list computed for the United Arab Emirates, which is located in neuron 11, would not be a reliable list since that neuron has low ζ_2 values and low ζ_{3-5} values, hence, high uncertainty.

Sometimes only one region is allocated to a neuron, which can impede the calculation of ζ diversity for that neuron. In that case, an option is to perform a two-step clustering of the SOM map neurons (Vesanto & Alhoniemi, 2000) and then, calculate ζ for each cluster of neurons, considering each cluster of neurons, as if they were a single neuron or study.

3.4 Discussion

3.4.1 Assessing SOM PPA cluster validity using ζ diversity

Numerous scientific approaches have been developed and adopted to conduct Pest Risk Assessment. Yet, in a comprehensive review of over 300 quantitative methodologies applied to risk assessment, Leung *et al.* (2012) found that quantitative risk assessments are underused in policy, because of their data requirements and their lack of generalizability (McGeoch *et al.*, 2016). In fact, data requirements for SOM PPA are readily accessible. SOM PPA is able to extract valuable information using only a global occurrence matrix as an input. Data can be incomplete, for many reasons that range from under sampling (**?**McGeoch *et al.*, 2012; McNeely *et al.*, 2001) to trade interests and lack of transparency of trading nations. In any case, we need to acknowledge data as a limitation of any kind of modelling (Venette *et al.*, 2010). Regarding generalizability, clustering regional pest profiles, or by extension, invasive species profiles, was described by Worner & Gevrey (2006), and since then, a number of studies have used the approach, and broadly applied it to other taxa like fungi (Paini *et al.*, 2011), weeds (Morin *et al.*, 2013) or bacteria and nematodes (Eschen *et al.*, 2014), using SOM as well as other clustering approaches such as k-means and hierarchical clustering. For an extended discussion about the validity of using pest assemblage similarity to infer potential of establishment see the review by Worner *et al.* (2013).

In this study we emphasize the generalisability of the SOM PPA approach and of the significant improvements to it that we have proposed here. By taking account of the number of shared species between i regional assemblages, zeta (ζ_2) described the similarity between groups of pest assemblages as equivalent (as shown by Hui & McGeoch 2014) matched that of another commonly used measure, Sorensen index (Sørensen, 1948) (Figure in B.3). More importantly, higher orders of ζ provided additional insights about similarity and composition that pairwise measures cannot provide. Using ζ as a measure that provides three different levels or degrees of similarity substantially improved the interpretation of SOM PPA, and at the same time is applicable to many other clustering approaches or diversity studies.

Moreover, ζ, unlike other pairwise similarity measures, does not require assumptions about site and species independence (Hui & McGeoch, 2014). A potential limitation of the use of ζ is its dependence on species richness (Hui & McGeoch, 2014), but this can be simply overcome by using normalized ζ,(ζ_i/ζ_1), which makes the similarity of two sets of sites fully comparable.

3.4.2 Using normalized ζ as an uncertainty measure

The goal of communicating uncertainty in an environmental modelling context is to minimize the possibility of making an incorrect decision about a potentially adverse outcome (Matott *et al.*, 2009). In terms of SOM PPA, making the wrong decision consists in failing to choose the appropriate species to initiate a PRA, potentially wasting time and resources. In SOM PPA, as in any real-world application of a method, the end user needs to have some confidence measure to base their decisions on, therefore, uncertainty has to be acknowledged quantified and communicated (Wattenbach *et al.*, 2006; Ascough *et al.*, 2008).

Reducible uncertainty in environmental or ecological modelling can be decreased through better practices such as increasing data accuracy (see McGeoch *et al.* (2012) for doing this with invasive species lists), using computational approaches like bootstraping, sensitivity analysis of the model parameters (Matott *et al.*, 2009) or Monte Carlo simulations (Ascough *et al.*, 2008). However, actively reducing uncertainty is not easy since its sources are not simple to identify. In the context of SOM PPA, there are many factors that will affect degree of confidence we can have in the outputs. SOM PPA uncertainty is the degree to which similarity across regions in a neuron is high, and remains high, across zeta orders. In other words a SOM PPA output list with low uncertainty is one where the neuron which allocates the target region has high average number of shared species across pairs of regions and declines slowly with increasing numbers of regions in the comparison. We have shown that groups of regions with very low species counts, or with very rare species obtained very low ζ values. We could assume that for these regions data inadequacy in the form of incomplete species is a major uncertainty source. However, there are regions which are very rich in species that also have low values of ζ_{3-5}, suggesting these assemblages are only superficially similar. Thus, the source of uncertainty for these neurons could be the self-organizing map algorithm itself. Cluster analysis is significantly affected by uncertainty because different attributes may have different uncertainty levels (Aggarwal, 2009). Future research performing sensitivity analysis of the SOM algorithm is recommended in order to quantify model uncertainty. Other than that, given the complexity and sometimes impossibility of reducing the uncertainty, the use of ζ to categorize it into three levels and communicate it is a major improvement.

To conclude, incorporating a measure of uncertainty based on ζ diversity helps advance the reliability of SOM PPA as a technique, even though it marginally increases its complexity (Figure 3.3). It is important to note that the measure is case-specific and therefore it is not possible to predefine thresholds of ζ to characterize what to consider goodness of clustering. More important is the relative ζ values per neuron, in other words, how similar are the assemblages clustered in one neuron compared with similarity of those clustered in another neuron. Despite the extra effort required, we believe adding an uncertainty measure improves the methodology. This is especially so because using three categories alongside the risk list communicates uncertainty in a simple, informative and meaningful way, which is needed in environmental modelling (Venette *et al.*, 2010), pest risk assessment (Leung *et al.*, 2012) or by extension, any applied science aimed at supporting the decision making process (Walker *et al.*, 2003; Wattenbach *et al.*, 2006).

Chapter 4

Biotic homogenization of global assemblages of insect crop pests

Notes

This chapter was submitted, peer reviewed and finally rejected in the journal Global Ecology and Biogeography

A reviewed version is intended to be submitted as:

Roigé, M., Lustig,A., Parry, M., Phillips,C. and Worner, S. P. (2017), Biotic homogenization of global assemblages of insect crop pests. Journal of Biogeography

Acknowledgement

This manuscript has benefited from comments and suggestions by Ursula Torres and Marta Gallart. M.R received financial support from the European Commission (Erasmus Mundus partnership NESSIE, ref. 372353-1-2012-1-FR-ERA MUNDUS-EMA22) and Better Border Biosecurity (B3)(www.b3nz.org) research collaboration. The authors state no conflict of interests.

Abstract

Aim: Biotic homogenization is a major consequence arising from habitat fragmentation and biological invasions, yet its magnitude remains unknown due to the lack of studies that measure this process at a global scale and over different extents. In this study, we quantify the level of biotic homogenization for insect crop pests worldwide and its advance over a period of 12 years (2003-2014). We also propose a general method for quantifying biotic homogenization at different study extents

Location: Global

Time Period: 2003 to 2014

Major taxa studied: Herbivorous insect crop pests

Methods: We characterized β diversity for 423 of the world's regional pest assemblages comprising the presence and absence of 711 species at two different times (2003 and 2014) using the ζ diversity metric. We then calculated the homogenization rate by comparing the ζ values from 2003 to ζ 2014 over ten different study extents

Results: We showed a general spread of global insect crop pests where the records of occurrence increased from 36,107 in 2003 to 41,110 in 2014. Most of the 423 regions (82%) studied also showed an increase in species richness. Beta diversity decreased in general, although at different rates for different study extents.

Main conclusions: There has been a global increase in crop pest species richness paired with a decrease in β diversity, therefore, there is, for the studied taxa, a clear pattern of biotic homogenization. Additionally, our results show that homogenization rate is correlated to the number of assemblages compared at a time. In other words, measures of homogenization are dependent on the study extent, such that pair-wise measures of β diversity may underestimate the true homogenization rate.

Keywords

biotic homogenization, beta diversity, zeta diversity, pest assemblages, insects

4.1 Introduction

Biological invasions and habitat destruction are widely accepted as the main drivers of biodiversity loss (Mack *et al.*, 2000; McGeoch *et al.*, 2010). Ongoing translocations of non-native species through human assisted dispersal increases their potential to establish in new communities throughout the world. An expected consequence of these translocations is increasing similarity in the taxonomic composition of the global biota. This phenomenon is referred to as global biotic homogenization (Vitousek *et al.*, 1996; McKinney & Lockwood, 1999; Lockwood & McKinney, 2001; Olden & Poff, 2003).

The first conceptual framework for studying biotic homogenization was provided by Olden & Poff (2003). It described fourteen different theoretical ecological scenarios where the number of wining (invaders) and losing (extinct) species defined different homogenization outcomes. Olden (2006) reviewed and summarized existing studies of biotic homogenization. The author identified knowledge gaps and emphasized the lack of clear empirical evidence for homogenization, arguing that claims about its magnitude and impacts on biodiversity remained undefended. Olden *et al.* (2016) warned about bias in such studies towards certain taxa and pointed out that insects were the least investigated taxon. Olden *et al.* (2016) also reported some contrasting studies of biotic differentiation. For example, Marchetti *et al.* (2006) investigated the community composition of freshwater fish in California and recorded an increase in community diversity following non-native species introductions. Similarly, Vellend *et al.* (2013) concluded after a meta-analyses of species diversity studies at small scales, that there were no changes in local communities species richness despite a clear global species diversity decrease.

The consequences of increasing biotic similarity remain unknown. Olden *et al.* (2004) argued that it is unclear whether biotic homogenization will promote higher levels of new species creation or will limit species diversity by diminishing the geographic isolates needed for speciation. Therefore, to understand the magnitude, causes and effects of biotic homogenization, Olden *et al.* (2004) and others (Gonzalez *et al.*, 2016) called for studies that quantify homogenization at a global extent across different taxa.

To test for biotic homogenization, it is necessary to obtain species distribution data over space and time at a global extent. Researchers of agriculture and forestry have compiled such data on herbivore community composition on economically important hosts (Lewinsohn *et al.*, 2005). Perhaps the most comprehensive database of occurrences of agricultural pest species is the CABI Crop Protection Compendium (CPC) (Bebber *et al.*, 2014a; CABI, 2007). The Compedium was recently used to extract the global distributions of plant pests and diseases to model and predict future invasions using ordination approaches (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Watts & Worner, 2009; Paini *et al.*, 2010a). Bebber *et al.* (2014b) also used it to investigate the economic and physical determinants of the global distributions of crop pest and pathogens. Following that study, Bebber (2015) studied the potential range-expanding effects of global warming on their distributions. Bebber *et al.* (2014a) concluded that biotic homogenization was occurring rapidly among the plant pests and diseases recorded in the CPC. However, their conclusion was based on distributional changes of 424 crop pests and pathogens over the period 2000 to 2011, which is a rather small subset of the available recorded species in the CPC. Also, they reported results for both pests and pathogens jointly and did not specifically quantify the degree of biotic homogenization for insects pests alone.

Usually, biotic homogeneity is quantified either by comparing assemblage composition measured by a particular similarity index over time, or by comparing distances between assemblages or communities in species space using clustering or ordination techniques (Olden & Poff, 2003; Rahel, 1990). Comparing species assemblages using a similarity index is equivalent to calculating their compositional β diversity. Beta diversity encompasses a variety of indices and concepts that reflect different components of between-sites species variability, or species turnover along environmental, spatial or temporal gradients (Whittaker, 1972; Vellend, 2001). Two well known problems associated with β diversity metrics are that community pairwise similarity indices assume site independence and that they are limited to the comparison of two sites at a time. When pairwise similarity indices are used to compare more than two assemblages, communities or sites, they are calculated by averaging the pairwaise metrics across sites. To circumvent these issues, Hui & McGeoch (2014) proposed the zeta diversity (ζ) metric, which is a measure of compositional β diversity defined as the number of species shared between any i sites or assemblages. Zeta is a simple similarity measurement that reconciles existing descriptors of species incidence and spatial turnover (Hui & McGeoch, 2014) and overcomes the main criticisms of the pairwise similarity measures such as Jaccard, Sorensen (Sørensen, 1948) and Shannon diversity metric (Shannon, 1948).

Zeta is unaffected by site-dependence, and allows more than two sites to be simultaneously compared. Multi-site comparisons allow different orders of ζ (the number of features compared at one time) to describe different levels of similarity which is useful for understanding relationships within groups. For example, ζ was recently used to validate results from a cluster analysis, and revealed different grades of group similarity than other pairwise measures could not (Roige *et al.*, 2016). In this study, we use ζ orders to provide different homogenization values for different study extents, which is particularly relevant since spatial and temporal extent have significant effects on reported measures of biotic homogenization (Olden *et al.*, 2016).

Our primary objectives in this paper are: First, to empirically test the hypothesis of biotic homogenization for agricultural insect pest assemblages, describe their diversity components at regional and global scales, and quantify their changes over the time 2003 to 2014. Second, to evaluate ζ diversity as a measure of assemblage composition and the ratio ζ_{2014} to ζ_{2003} as a generalizable measure of homogenization advance.

4.2 Methods

4.2.1 Data

The 2003 global insect pest occurrence matrix used in this study was originally compiled by Worner & Gevrey (2006) and Gevrey *et al.* (2006) from the Crop Protection Compendium (CPC) (CABI, 2007). It was arranged in 459 rows (regions) and 844 columns (species). An equivalent matrix of the same size was extracted from the CPC in 2014 with updated records for the same species, and is referred to as the 2014 global pest occurrence matrix. To reflect changes in CPC nomenclature during 2003-2014 and ensure proper comparison, we filtered both matrices so that they contained identical species and regions. This created two matrices of 423 regions and 711 species that summarized the geographical distributions of phytophagous insect pests considered relevant to global crop protection by CABI. Presences were coded as 1 and absences were coded as 0. Geographic areas summarized in CPC consisted of countries, regions or states of countries, all of different sizes.

4.2.2 Using ζ diversity metric as a measure of α, γ and β diversities

In a study with i regional assemblages, gamma diversity (γ) is the total species counts across the i regions, alpha diversity (α) is the local species richness (Whittaker, 1972) or number of species in each of the i regions, and zeta ζ is the average number of species shared by i regions (Hui & McGeoch, 2014), which can be pairs or larger groups.

Zeta can be calculated for any i. The number of regions i defines the order of ζ. For instance, ζ_1 is the average number of shared species across all the regions considered in a study (average species richness, conceptually equivalent to $\bar{\alpha}$), ζ_2 is the average number of shared species between two regions (equivalent to pairwise β diversity measures), ζ_3 is the average number of shared species between three regions and so on, until ζ_i, which defines the average number of shared species between all i regions.

To quantify local α diversity we computed 423 different values representing the number of species present in each region. In each matrix,γ diversity totalled 711. We measured β diversity by calculating ζ for the 2003 and 2014 occurrence matrices, from ζ_1 to ζ_{10} (higher orders of ζ were not reported because 0 was reached within first 10 orders). To assess their significance, we compared all metrics (ζ_{1-10}) to those generated from randomized occurrence matrices comprising the same number of presences and absences as in the 2003 and 2014 occurrence matrices respectively. Details on how the random matrices and their ζ values were generated are given in C.3.

4.2.3 Zeta diversity as a quantification of biotic homogenization

To test for biotic homogenization Olden & Poff (2003) suggested quantifying the advance of species distributions over time. In this study we use ζ to describe the change in the distributions of regional insect crop pest assemblages during an eleven year interval. We computed ζ_{1-10} for the 2003 and 2014 global pest occurrence matrices and compared their raw and normalized values (normalized $\zeta_i = \zeta_i/\zeta_1$).

We computed the ratio $(\zeta_{2014}/\zeta_{random_{2014}})/(\zeta_{2006_{1-10}}/\zeta_{random_{2006}})$ and interpreted it as the homogenization rate of advance. The homogenization ratio indicates, for each order of ζ, how much the proportion of shared species between i regions changed from 2003 to 2014. Normalized ζ_1 explains the proportion of shared species between i regions, thus normalized ζ_1 is always 1. Higher orders of the normalized ζ_i ratio show how the proportion of species shared between i regions increased between 2003 and 2014. We used R (R Foundation for Statistical Computing, 2014) to compute α, γ, ζ and the homogenization ratio, specifically the package 'zetadiv' (Latombe *et al.*, 2015). Details and code for computating ζ are in C.1.

4.3 Results

4.3.1 Diversity components of regional insect pest assemblages

4.3.1.1 Global species richness: γ diversity

The analysis showed an increase in γ diversity over time. In 2003 there were 36,107 records of presence for the 711 species across 423 regions. In 2014 the number of presences for the same 711 species had increased to 41,110.

4.3.1.2 Local species richness: α diversity calculated by ζ_1

Most (82%) of the regions had more species in 2014 than in 2003. Eighteen regions experienced increases of more than 100% and average species richness per region ($\bar{\alpha}_{2006}$) increased from 85.36 species in 2003 to 97.18 in 2014 ($\bar{\alpha}_{2014}$). Sixteen regions (4%) had fewer species in 2014 than in 2003 and 23 (5.4%) regions showed no change between years. In 2003, the most species-rich region was USA with 381 species. In 2014, it was India with 322 species, while the species counts for USA decreased by 80 to 301. The 16 regions where species richness declined included the ten most species-rich regions, which lost 10-20% of their species. The remaining six regions had low species richness in 2003, thus a decrease of one or two species represented a high percentage loss. The most dramatic declines in species richness were observed in Caroline Islands (98.4%), Rodriguez Islands (72,7%) and Nusa Tenggara (43.5%) but again, those were not the norm. Figure 4.1 shows the percentage change from α_{2003} to α_{2014} for each region. There were 489 records of 236 species that changed from present in 2003 to absent in 2014.

4.3.1.3 Beta diversity, explained by ζ

Beta diversity decreased from 2003 to 2014 because the values of ζ_{1-10} for the 2014 occurrence matrix were higher than those for the 2003 matrix. This means the number of species shared by i regions in 2014 was higher than in 2003, thus diversity among the regional assemblages was reduced (Table 4.1). The changes in the values of ζ differed from those observed in the random datasets. Randomly generated datasets showed log-linear decreases in ζ with increasing orders, as opposed to the real data, that showed a structured decrease in ζ with increasing orders (Figure C.1).

The value of ζ_1 (Table 4.1) corresponds to the average number of species per region, or average α diversity. The average number of species shared between any two regions (ζ_2) was 17.49 in 2003 and increased to 21.81 in 2014. In proportional terms, it increased from 20,49% of shared species to 22.45%. The proportion of species shared between any three assemblages (normalized ζ_3) increased from 6.04% to 7.07%. Higher orders of ζ followed the same trend.

Table 4.1: Columns 1 and 2 show ζ values for the 2003 and 2014 pest occurrence matrices, units are species counts. Columns 3 and 4 show the average percent shared species between i assemblages, where i is the order of ζ.

	ζ 2003	ζ 2014	Normalized ζ 2003	Normalized ζ 2014
ζ_1	85.36	97.19	-	-
ζ_2	17.49	21.81	20.49	22.45
ζ_3	5.15	6.87	6.04	7.07
ζ_4	1.93	2.79	2.26	2.87
ζ_5	0.80	1.27	0.94	1.31
ζ_6	0.36	0.65	0.42	0.67
ζ_7	0.17	0.35	0.20	0.36
ζ_8	0.09	0.19	0.10	0.20
ζ_9	0.04	0.11	0.05	0.12
ζ_{10}	0.02	0.07	0.03	0.07

4.3.2 Homogenization ratio

In 2003, two randomly chosen regional assemblages worldwide shared an average 20.49% of their species, and in 2014 this increased to 22.45%. These figures gave an homogenization ratio of 1.10 (Figure 4.2), which provides a measure of the advance of biotic homogenization. Higher orders of zeta, ζ_{3-10}, showed this increase also occurred at larger study extents comprising more regions. For example, for groups of three regions the homogenization ratio increased 1.18 times from 2003 to 2014, and for groups of 10 regions it increased by 2.67 times. In summary, the proportions of species shared by i regions were always higher in 2014 than in 2003 (i.e,. homogenization ratios were always > 1) and increased as more regions were included (Figure 4.2).

4.4 Discussion

Despite being one of the largest taxonomic groups of the animal world, arthropods remain understudied with respect to both invasion biology (Bacon *et al.*, 2014) and biotic homogenization (Olden *et al.*, 2016). Moreover, their impacts on human activity, although already enormous, are likely often underestimated (Bradshaw *et al.*, 2016). Herbivorous insect crop pests are critically important due to their devastating effects on agricultural ecosystem services, and the threats they pose to endemic ecosystems (Díaz *et al.*, 2006; Hulme, 2009; Perrings *et al.*, 2005; Vilà, 2013; Horan & Lupi, 2010). We investigated regional assemblages of these pests and provided the first numerical estimates of their homogenization rate.

We showed changes in insect crop pest species distributions between 2003 and 2014 conformed to the hypothesis of biotic homogenization. Beta diversity decreased globally while local richness generally increased. In particular, local species richness, α diversity, increased for 384 regions, the average α diversity also increased, and globally, the number of new records for the 711 species also increased. Exceptions involved sixteen regions that had fewer species in 2014 than 2003, which probably arose from CPC data corrections. These comprised only 489 records (0.16%) of the 300,753 analysed. We did not manually check those records because of the large amount of time that task would require. Also, given the tiny proportion these records represent, adjusting them would either have only marginally increased, or left unchanged, the homogenization estimates.

4.4.1 Biotic homogenization scenarios

Previous studies stressed the need to quantify biotic homogenization in terms of its magnitude, spatial extent and also to describe how it covaries with other diversity components (Olden, 2006). Theory about biotic homogenization predicts that the arrival of a new species in a new community will first increase the number of species, augmenting both α and β diversities in the community (Olden, 2006). Then there will be a transient time during which community dynamics (extinctions of resident species through competitive interaction) shapes the form of the assemblages generally resulting in an increased local α diversity (communities will tend to have higher number of species) at the expense of a decreased global beta diversity, the same set of invaders will be globally successful, thus community composition will be more similar across the world (McKinney, 2004; Rosenzweig, 2001; Lewinsohn *et al.*, 2005).

Olden *et al.* (2004), described a set of fourteen theoretical ecological scenarios where the number of winning (invaders) and losing (extinct) species defined different homogenization outcomes. The insect crop pest assemblages considered in this study are anthropogenically created groupings in which extinctions are rare (Pimentel *et al.*, 2001). Thus, we expected changes in species distributions between 2003 and 2014 would conform to one of the two scenarios simulated in Olden & Poff (2003) that involved species invasions without extinctions of local species. In scenario one, a single species invaded different communities, which in their simulations always led to biotic homogenization. In scenario two, different species invaded different communities, and two different outcomes were possible depending on the number of species and the number of communities they invaded at one time. If numerous species simultaneously invaded many communities, then those communities would become more homogeneous. In contrast, if numerous species invaded relatively few communities, then those communities would become less homogeneous. Our results were clearly indicative of homogenization whereby many non-native species each invaded numerous regions, which created greater community homogeneity throughout the world, thus causing β diversity to decrease. Therefore, the insect crop pest invasions case belong to Olden & Poff (2003) scenario two in which the number of simultaneous establishments (which leads to homogenization) compensates the number of new species invading a single assemblage or region (which leads to differentiation). The high number of simultaneous establishments is consistent with the bridgehead effect, commonly observed among crop pest invasions (Guillemaud *et al.*, 2011). That is, in their invasion routes, crop pest species establish intermediate successful invasive populations (bridgehead populations) in secondary regions from where they can disperse into additional regions, thus multiplying their potential routes of dispersal and increasing the probability they will arrive in several new regions at similar times.

4.4.2 Homogenization ratio and study extent

Beta diversity measured by ζ_{2-10} showed a significant decrease and the different homogenization ratios for different ζ orders showed that the more regions we compared at a time, the higher the homogenization advance.

Two other studies have measured biotic homogenization for invasive species at a global extent (Bebber *et al.*, 2014a; McKinney, 2004). Bebber *et al.*'s (2014a) temporal (2000-2014) study of biotic homogenization comprised 424 species of crop pests and pathogens and used Shannon's diversity index to measure assemblage similarity. McKinney (2004) used Jaccard similarity index to measure the pairwise similarity of 20 localities which lead to 190 pairwise site comparisons of plant species. Both Bebber *et al.* (2014a) and McKinney (2004) concluded that when considering exotic species only, their data showed signs of biotic homogenization. In this study we have augmented the extent of their studies, comparing 711 species across 423 sites. We also applied a multi-site similarity metric that allowed us to show how homogenization ratio advanced at different study extents, when considering different numbers of assemblages (or regions).

Olden (2006) questioned whether calculating a mean pairwise change in community similarity was a valid measure of biotic homogenization. We argue that pairwise similarity metrics such as Jaccard, Sorensen and Shannon indices describe only one dimension of homogenization. We propose that ζ can encapsulate more dimensions and can differentiate almost continuously what happens at different study extents. Furthermore, we exemplified how ζ can quantify the advance of biotic homogenization, which is a major advantage over traditional pairwise β diversity measures. Our analyses suggested that traditional beta diversity computations may underestimate biotic homogenization by only accounting for two sites at a time.

That is a key result that could allow biodiversity studies to quantify biodiversity across spatial extents. There is much controversy about how biodiversity estimates at local scales may be underestimates of real biodiversity changes at global scales (Loreau, 2002; Gonzalez *et al.*, 2016). In this study, ζ explored the changes in diversity between different number of assemblages, although those were not over a spatial gradient. We suggest that ζ orders can be useful for future research at characterizing diversity changes over the local-global continuum.

Our results were only meaningful up to ζ_{10}, because the value of zeta reached zero at higher orders. But if the 2003 regional assemblages had had more shared species, then ζ would be informative at higher orders, theoretically up to where i equals the total number of regions. The present study is, to our knowledge, the first quantitative demonstration of how biotic homogenization is influenced by the study extent when measured with assemblage similarity metrics.

4.5 Conclusions

1- Global biotic homogenization is occurring amongst assemblages of invasive insect crop pests, and its rate can be estimated using ratios of normalized ζ diversity measured at different times.
2- Our results are consistent with the species invasion-only scenarios of Olden & Poff (2003) because local α increases while there is a decline in global β diversity.
3- Biotic homogenization rates increase with study extent (number of regions studied), which is a fact overlooked by classic pair-wise similarity measures.

4.6 Figures

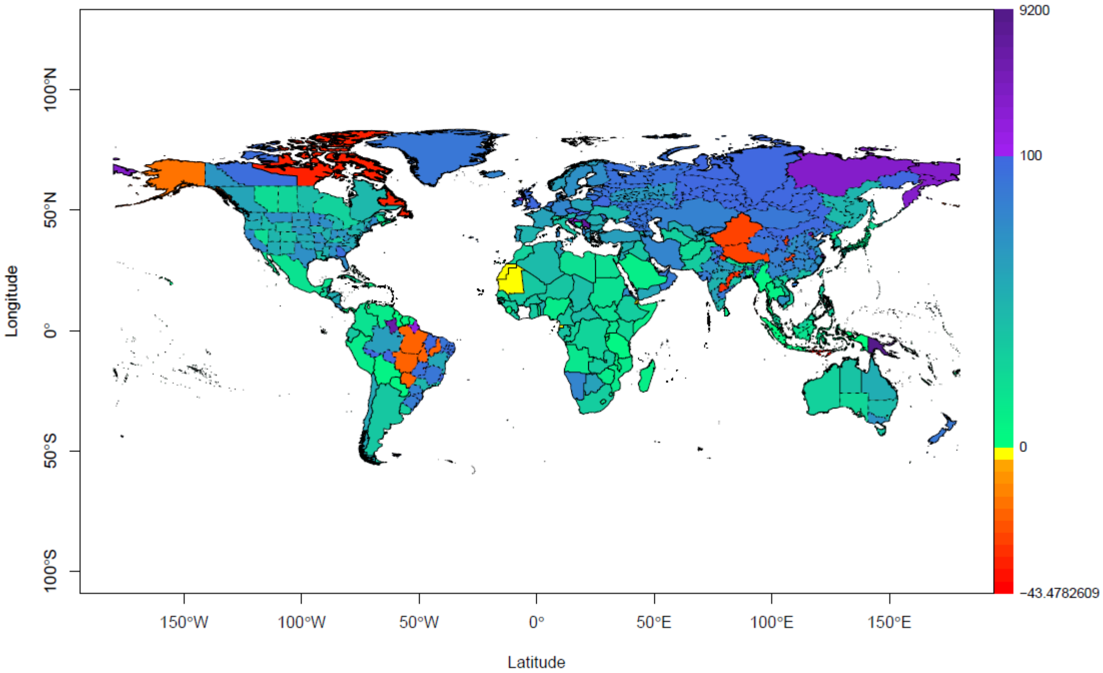


Figure 4.1: Percentage increase of α for all the regions considered in the CABI CPC database from 2003 to 2014. Warmer colors indicate percentage decreases in α, yellow indicates no change in α and colder colors indicate percentage increases in α.

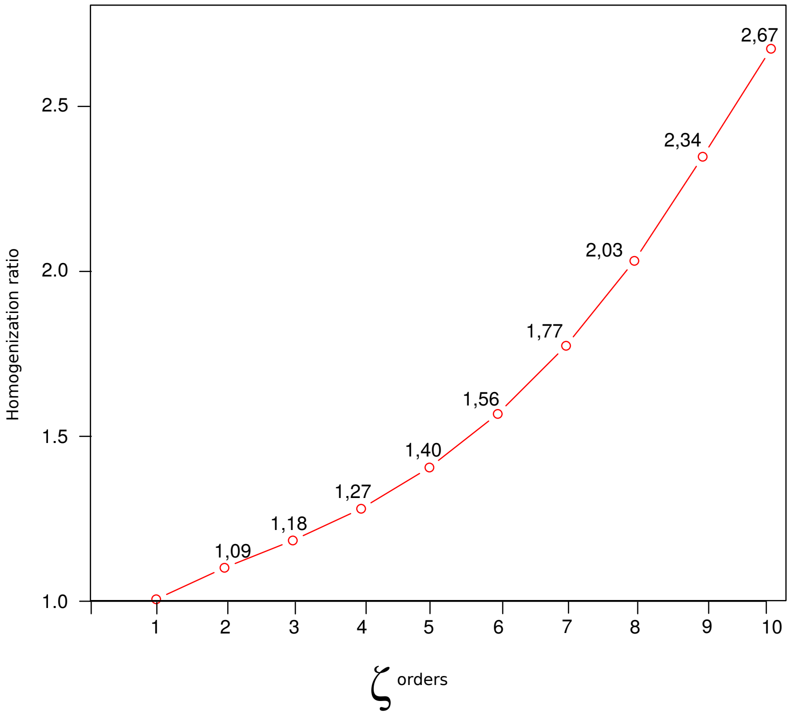


Figure 4.2: Plot of homogenization ratios calculated as a ratio of normalized ζ values for 2014 to normalized ζ values for 2003

Chapter 5

Comparison of clustering methods for pest profile analysis

Keywords

clustering method, SOM, Hierarchical clustering, k-means, pest profile analysis, comparison

5.1 Introduction

Pest Profile Analysis (PPA) is a quantitative approach which aims to assist pest risk analysts' decisions about prioritising management of invasive pest species. PPA was first described by Worner & Gevrey (2006) and is based in the study of regional insect pest assemblages. In PPA, world regions are compared according to similarities in the composition of their pest species (pest profile). Regions with comparable pest profiles are assumed to share similar biotic and abiotic characteristics (thus enabling their comparable suites of pests to become established). Therefore, localities with similar pest profiles are assumed to be high risk of exchanging pests with one another in the future.

Cluster analysis is the formal study of methods for discovering natural groupings or patterns in data (Jain, 2010). The first use of clustering methods applied to PPA was in Worner & Gevrey (2006); Gevrey *et al.* (2006) who used self organizing maps (SOM) to cluster 459 world regions according to similarities between their crop pest profiles. After clustering the regions, SOM weights were used to infer strength of association between each species and each region's pest profile, and interpreted as a proxy for risk of establishment. A detailed explanation of how the SOM weights were used can be found in Roigé *et al.*

(2016). Other clustering methods have been explored for use in PPA. For example, Worner and Gevrey (2006) conducted a hierarchical cluster (HC) analysis comprising single and complete linkage clustering using the Jaccard coefficient and Euclidean distance as similarity metrics on their original (2006) data, but they reported that the HC analyses failed to organize the data and the results could not be interpreted.

Later, Watts & Worner (2009) compared SOM to the k-means clustering method for estimating establishment risks. Because there is no equivalent to SOM weights in k-means algorithm, Watts & Worner (2009) estimated the risk of invasion by computing the relative frequency of occurrence of each pest species in each cluster. After the comparison, the authors concluded that while both algorithms were able to produce meaningful clusters, k-means seemed to perform better in terms of goodness of cluster (measured by Shannon's entropy index) and was more efficient in terms of running time.

A third application of clustering for PPA was carried in Eschen *et al.* (2014), who analysed the distributions of around 1000 invertebrate pests and pathogens of woody hosts in 344 global regions using the European Union as a target region. They applied Ward's method of hierarchical clustering. To generate risk indices, Eschen *et al.* (2014) used Watts & Worner (2009) approach of calculating the relative frequency of a species in a certain cluster. Eschen *et al.* (2014) did not compare HC to SOM, but called for future research to do so.

SOM, k-means and HC are unsupervised learning techniques for clustering, and as such, they all present methodological problems related to their assumptions about cluster shape (Jain *et al.*, 1999; Kovács *et al.*, 2005). The main issues are choosing the number of clusters (Mirkin, 2013) and evaluating the results (Halkidi *et al.*, 2001). It is difficult to evaluate clustering results because the characteristics that separate a 'good' from a 'bad' cluster differ depending on the research question. Thus, assessing a clustering procedure's output, is highly subjective (Jain *et al.*, 1999). In general, there are three ways to validate cluster results; internal, external and relative (Jain *et al.*, 1999). To evaluate the ability of any method to discover patterns in data, ideally the results are compared to reality (external validation). Often, however, the required information is unavailable or there are insufficient data to both train and validate the models. A common solution to circumventing lack of validation data is to validate using artificial, error-free data (Zurell *et al.*, 2010).

Paini *et al.* (2011) were the first to use artificial data to validate SOM results for PPA and showed an average 97% predictive power of SOM. K-means and HC have not yet been validated for application PPA, and none of the three approaches have been validated using external data. More general studies have compared the performance of k-means, SOM and HC (Waller *et al.*, 1998; Astel *et al.*, 2007; Mangiameli *et al.*, 1996; Mingoti & Lima, 2006; Bação *et al.*, 2005) but the results were often contradictory. The overall conclusion from these studies was that is always preferable to test the algorithms with the data they are going to be applied to.

Regarding choice of number of clusters, Gonzalez *et al.* (2010) conducted a sensitivity analysis of SOM which showed the results were highly sensitive to the size or number of output neurons (equivalent to clusters) used in the analysis. K-means is similarly sensitive to the initial k seeds (which then become the clusters) (Jain, 2010), and similarly, in HC, the level at which the tree (dendrogram) is cut determines the number of clusters and whether they are meaningful (Zaki & Meira, 2013).

Our main objective in this study was to assess which of the three clustering methods performs best for pest profile analysis (PPA). We used SOM, k-means and HC in an *a posteriori* multiple technique analysis of global crop pest distribution data from CABI CPC 2003 (CABI, 2007). We assessed the predictive power of the three cluster analyses by comparing their predictions (the risk indices of pest establishment) to the real observed data recorded in 2014. We used a combination of metrics for evaluating predictive power, with a focus on answering questions that pest risk analysts would be interested in.

5.2 Methods

5.2.1 Data

We used the data originally extracted by Worner and Gevrey (Worner & Gevrey, 2006; Gevrey *et al.*, 2006) from Crop Protection Compendium International (CPC) (CABI, 2007), organized in a matrix of 459 regions (rows) and 844 species (columns), and referred to as 2003 dataset. Presences were coded as 1 and absences were coded as 0. The geographic areas represented consisted of countries, or regions, provinces or states within countries, all of different sizes. In a few cases, large countries appeared more than once in the matrix; as a whole and also divided into their states. To externally evaluate the methods' predictive power, we extracted more recent data (2014) for the same species and regions (henceford referred to as 2014 dataset). To ensure that 2003 and 2014 data sets were fully comparable, we deleted any species that only appeared in one of the two data sets (this occurred, mostly due to changes in taxonomic classifications). Our final sets of test and observed data comprised records for 711 species and 423 regions.

5.2.2 Test design

We replicated the procedures carried in Worner & Gevrey (2006), Watts & Worner (2009) and Eschen *et al.* (2014) (explained in section 5.2.3) with the purpose of realistically comparing the three existing approaches that had been used across PPA literature (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Watts & Worner, 2009; Paini *et al.*, 2010a, 2011; Morin *et al.*, 2013; Singh *et al.*, 2013; Eschen *et al.*, 2014).

For each clustering method, we used an optimization rule to choose the number of clusters. For the SOM method, we applied the heuristic used in Worner & Gevrey (2006), $nc = 5\sqrt{(c)}$, as recommended by Vesanto *et al.* (1999). For *k-means*, we chose the number of clusters that maximized the value of the Krzanowski-Lai index (Krzanowski & Lai, 1988), as recommended in Walesiak (2016), and for HC we followed Eschen *et al.* (2014) and chose the number of clusters that had the minimum Davies Bouldin index (Davies & Bouldin, 1979).

Finally a *control* test was also conducted which involved calculating the relative frequencies (rf) of all species in the 2003 data and using these frequencies as risk indices to make predictions for presence/absence of the species in 2014, without applying any clustering method. Details on the specifications of all the algorithms and the control are given in Table 5.1. After obtaining the risk indices for each classifier, they were scaled using the formula ((x-mean)/range)), so that their distributions would all range between 0 and 1.

Table 5.1: Summary table of tested methods and their specifications

	Number of clusters (nc)	Formula to choose nc	Risk index calculation
Control	1		relative frequences
SOM	99	$nc = 5\sqrt(c)$	SOM weights
k-means	97	highest Krzanowski-Lai	relative frequencies
HC	4	first minimum Davies Bouldin	relative frequencies

5.2.3 Clustering methods

5.2.3.1 Self-organizing maps

Kohonen (1982) described the self-organizing maps (SOM) algorithm as a tool to perform ordination vector quantization and classification. The SOM neural network is composed of two layers of elements or neurons: the input layer and the output layer. The input layer comprises the elements to be classified, and the output layer is represented by a map of a rectangular grid with m x n neurons (also called cells) laid in a hexagonal lattice (Worner & Gevrey, 2006). Sample vectors from the input data are presented to the algorithm, which evaluates their distance to the output neurons and chooses the best matching unit (BMU). The BMU is the closest output neuron in terms of Euclidean distance. At each iteration, the coordinates of the output neurons and their neighbour neurons are updated according to their BMU. The algorithm ends after a predetermined number of iterations. At the end of the iterations, each sample vector from the input layer is assigned to a neuron of the output layer. Since the number of output neurons is smaller than the number of sample vectors, more than one sample vector is assigned to the same output neuron, thus creating clusters. Each sample in the input layer is linked to the neurons in the output layer through a coordinates vector called a weight. SOMs can convert complex, non-linear statistical relationships between high-dimensional data into simple geometric distances that can be visualized on a low-dimensional display while preserving the original topological relationships. Therefore, distances between the original data and distances between the data items after they are mapped onto the output grid, are consistent. Preserving metric and topological relationships is a unique advantage of SOM over some other clustering methods (Kohonen, 1990). Based in the original algorithm, Vesanto *et al.* (1999) built a software implementation of SOM algorithm for Matlab, called the SOM Toolbox. There are two packages that implement self-organzing maps in R, they are the package 'kohonen' and the package 'som', but both have fewer optionalities than SOM Toolbox.

5.2.3.2 Self organizing maps for analysis of pest profiles

Worner & Gevrey (2006) used data comprising 844 phytophagous insect pests for 459 geographical areas for pest profile analysis. Data pre-processing included deleting those species that occurred in less than 2% of the regions. The number of clusters (or neurons in the self-organizing map) was 108 and it was chosen using the formula $c = 5\sqrt{n}$ where c is the number of training samples (sample vectors). The SOM was trained using a batch algorithm, and linear initialization. After the SOM ordination, each neuron of the output layer (or each cluster) corresponded to a virtual vector of weights, with weights comprising values in the interval [0,1]. Each weight can be interpreted as a risk index or degree of association of each species with the sites in each cluster. To perform the SOM analysis, (Worner & Gevrey, 2006) used the SOM Toolbox (Vesanto *et al.*, 1999) for Matlab (Mathworks, 2013).

In our study we replicated the original Worner & Gevrey (2006) parameters for the SOM. We also used the formula $c = 5\sqrt{n}$ to calculate the number of neurons but obtained 99 clusters instead of 108 because our data was smaller (711 species). SOM was trained using the batch algorithm and linear initialization and we also used the SOM weights obtained after ordination as risk indices for each species. We used SOM Toolbox (Vesanto *et al.*, 1999) for Matlab (Mathworks, 2013).

5.2.3.3 K-means

The k-means algorithm is a partitional algorithm which starts with the user defining k and finds a partition such that the squared error between the mean of a cluster and its points is minimized (Jain, 2010). K are the initial seeds (or centroids, or cluster centres) that will each become a cluster and it is the most critical choice of all k-means parameters. There are a number of heuristics to choose k, and they are usually based on running the algorithm with different values of k and choosing the one that minimizes a given criterion (Tibshirani *et al.*, 2001).

K-means works by comparing the sample vectors (units to be clustered) with each seed by computing the Euclidean distance between the vector and the seed, and then assigning the sample vector to the closest cluster seed. At each iteration the seeds are recalculated by averaging the vectors that have been assigned to it. The procedure is repeated until a stopping condition is met. The stopping condition is usually when changes in the seed between iterations are either close to zero or meet a threshold value (Johnson & Wichern, 2007).

5.2.3.4 K-means for analysis of pest profiles

Watts & Worner (2009) directly compared SOM and k-means classifications using the same data as Worner & Gevrey (2006). Once the clusters were generated with the k-means algorithm, risk values were computed by calculating the relative frequency of the species in the cluster and assigning that relative frequency to every region of the cluster. Watts & Worner (2009) chose the same number of clusters as in Worner & Gevrey (2006) which did not follow any heuristic, and was applied solely for comparing the two methods.

Therefore, to choose the optimal number of clusters, we ran a simulation for all possible number of clusters from $k = 2$ to $k = 423$ (number of regions) and chose the k that obtained the highest value of the Krzanowski-Lai index, using the function clusterSim of the package 'clusterSim' (Walesiak, 2016). When k-means is applied to binary data it is recommended to normalize the data beforehand (Legendre & Gallagher, 2001), thus, data were normalized according to the formula of unitization ((x-mean)/range).

5.2.3.5 Hierarchical clustering

Hierarchical clustering methods group the data by successive merging (agglomerative) or dividing (divisive) the sample units. The result is a dendogram or tree diagram where the branches represent clusters (Johnson & Wichern, 2007). The clusters in the hierarchy range from fine-grained to coarse-grained. At the lowest level of the tree, each point is its own cluster and in the highest level of the tree, all points are one cluster (Zaki & Meira, 2013). The height of the dendrogram chosen to cut the tree determines the number of resulting clusters.

Computing an agglomerative hierarchical clustering follows this process: Initially, there are as many clusters as units, and a matrix of pairwise distances between clusters is calculated. Clusters with low pairwise distances are merged, then the distance matrix is calculated again, and the closest clusters are merged once more. The process is repeated until all the sample units re grouped in one cluster. Divisive hierarchical clustering works in the same way but in an opposite direction, from one cluster to as many clusters as units.

There are many algorithms that perform hierarchical clustering, and most of them are variants of the single-link, complete-link and minimum variance methods (Jain *et al.*, 1999).

5.2.3.6 Hierarchical clustering for analysis of pest profiles

In Eschen *et al.* (2014) the authors used hierarchical clustering with Ward linkage (which is a type of minimum variance hierarchical clustering algorithms (Ward, 1963)) and reported that the other linkage methods did not produce clusters that could be interpreted. To choose the number of clusters, they used the Davies-Bouldin index (Davies & Bouldin, 1979). More specifically, they calculated the index and plotted it against the number of clusters, subsequently choosing the first minimum in the resulting curve as the optimal degree of cluster separation. Following their same procedure, we used the Ward linkage method of hierarchical clustering and plotted the results against each Davies-Bouldin index. In Eschen *et al.* (2014), to generate the risk values for each species and region, the authors calculated for each cluster the relative abundance of every species, and assigned it to all the regions in the cluster as a proxy for the establishment risk of that particular species to all those particular regions. Thus, the more times a species was present in the regions in the cluster, the higher risk this species represented for the remaining of regions of the cluster. We replicated the Eschen *et al.* (2014) same procedure to calculate establishment risk indices.

5.2.4 Evaluation methods

5.2.4.1 Confusion matrix and derived metrics

In binary classification problems, given the model predictions and the true observations, there are four possible outcomes that can be expressed in a confusion matrix (Table 5.2).

Translated into PPA terminology, a species that was predicted as likely to establish (obtained a high risk index) that did become established, it is a **true presence** (top left in Table 5.2). A species that was predicted as unlikely to establish (obtained a low risk index) that did become established, it is a **false absence** (top right in Table 5.2). A species that was predicted as likely to establish that did not become established is a **false presence**, and a species that was predicted as unlikely to establish and did not become established, is a **true absence**.

Table 5.2: Confusion matrix

		prediction	
		present	*absent*
reality	*present*	True presences TP	False absences FA
	absent	False presences FP	True absences TA

Two commonly used performance measures are the false absence rate and the true presence rate. The true presence rate is also known as sensitivity and it is a widespread performance measure that quantifies both the ability of the model to detect true presences and avoid false absences (Fielding & Bell, 1997). Sensitivity, in the context of PPA, is the model's ability to correctly predict species that did become established.

Equivalent measures can be derived from the top row of the confusion matrix, which are the true absence rate and the false absence rate (Table 5.2, top row). They are the proportions of correctly and incorrectly predicted absences of all real absences. The true absence rate is also called specificity (Fielding & Bell, 1997) and quantifies the model's ability to correctly detect the species that did not become established.

There are a plethora of performance metrics that can be derived from the confusion matrix, but they have one main problem and that is that they are threshold dependent. That is, any method that generates scores ranging between 0 and 1 (as the PPA risk indices do) needs a threshold value over which the method's predictions values are considered presences and below which are considered absences. This value of the threshold is dependent on the distribution of the risk indices of each method, therefore it is not suitable to compare different methods by comparing their performance at a given value of the threshold. As a consequence, many strategies have been developed to assess the overall performance of a model to be able to compare the prediction accuracy of different models.

5.2.4.2 ROC evaluation

The receiver operating characteristics (ROC) is a technique to select classifiers based on a visualization of their performance (Fawcett, 2006). The ROC consists of a two-dimensional plot of the results of the classification model for a set of thresholds. The true positive rate, or sensitivity, is plotted in the y-axis and the false positive rate, or specificity (actually 1-specificity is plotted instead to obtain ordered results), is plotted on the x-axis. Thus, the plot shows all the possible trade/offs between benefits and costs of the model. A classifier model is optimal if it lies on the convex hull of the set of points in ROC space.

An interesting property of the ROC space is that ROC curves are insensitive to changes in class distribution. If the proportion of present to absent species changes in a data set, the ROC curves will not change. The explanation of this phenomena is in the confusion matrix (Table 5.2). The class distribution or the proportion of absences to presences, is the relationship of the top to the bottom row. Any performance metric that uses values from both columns will be sensitive to class skews. Whereas ROC graphs, since they are based upon true presence rate and false presence rate, each dimension is a strict row ratio, thus do not depend on class distributions (Fawcett, 2006).

5.2.4.3 Overall performance metrics

The most well known performance metric to compare prediction models across a wide range of science disciplines is the area under the curve (AUC) which is literally the area comprised under the ROC curve of the model in the unit squared ROC space. The AUC reports the probability that the model will rank a randomly chosen present species higher than a randomly chosen absent species (Krzanowski & Hand, 2009) and is equivalent to the Wilcoxon test of ranks (Fawcett, 2006). The AUC can also be defined as the mean specificity value assuming a uniform distribution for the sensitivity (Anagnostopoulos *et al.*, 2012).

However, it has been reported by many that the AUC as a metric has one important flaw. That is, the AUC compares all the values of Sensitivity to Specificity in a way that assigns the same relative severity of misclassification cost to wrongly classifying a presence (false presence or Type I error) than to wrongly classifying an absence (false absence or Type II error) (Lobo *et al.*, 2008; Hand, 2009; Anagnostopoulos *et al.*, 2012; Hand & Anagnostopoulos, 2014).

In terms of biosecurity, it is clearly more important to avoid predicting false absences than it is to avoid predicting false presences. Species correctly predicted to have high potential to establish may not have done so (false presences), either through chance, or because effective border control measures had excluded them. Such false presences could naively be considered as incorrect predictions, but may not be because with more time, the predictions of high risk could prove true. Similarly, in areas such as medical diagnosis of life threatening diseases, a false alarm (false presence) generally cost less than a missed case (false absence) (Anagnostopoulos *et al.*, 2012), however it is very difficult to ask the end user or the researcher to specify the real cost of one misclassification over the other (Hand & Anagnostopoulos, 2014) .

As a consequence, Hand (2009) developed a metric called H measure that is analogous to AUC while explicitly accounts for different misclassification costs for different errors (Type I error and Type II error, which are also called commission and omission error in Lobo *et al.* (2008)). Specifically, the H measure treats missclassifications of the smaller class as more serious than those of the larger class (Hand & Anagnostopoulos, 2013), which in biosecurity terms translates into penalizing the PPA methods much more for the false absences they produce rather than for the false presences or true absences.

To perform the computation of the confusion matrix, ROC plots and H measure and AUC we used the package 'hmeasure' (Anagnostopoulos *et al.*, 2012) for the statistical software R (R Foundation for Statistical Computing, 2014).

5.3 Results

5.3.1 Distributions of risk indices per method

The risk indices generated after applying each clustering method are summarized in Table 5.3. All methods have continuous distributions of the risk indices, all of them having a range from 0 to 1, and similar values of their mean, except the *control* method which has a higher mean value (0.20). Figure 5.1 shows the distribution densities of the risk indices for each method and also for the 2014 observed data. The plot shows that all the distributions of the risk indices generated present right side skewness, that is, a much bigger number of zeros and close to zero values than close to one or one values. The 2014 observed data are binary, thus it only comprises values one and zero.

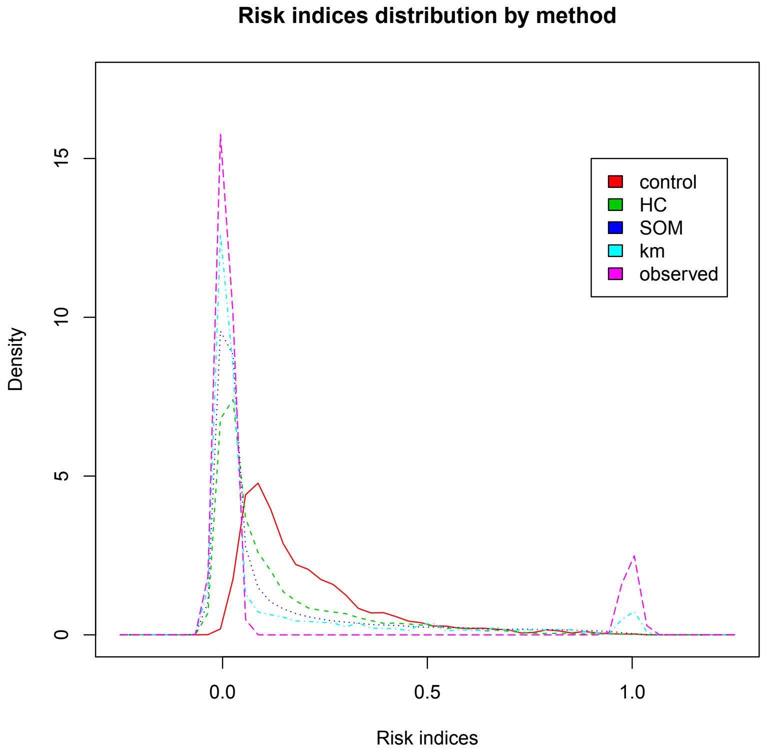


Figure 5.1: Kernel density plots of the risk indices for each method and for observed 2014 data (binary)

Table 5.3: Summary statistics of risk indices per method and for 2014 observed data. Observed 2014 data are binary

	control	*HC*	*SOM*	*k-means*	*observed 2014*
min	0.004	0.00	0.00	0.000	0.000
median	0.142	0.049	0.015	0.000	0.000
mean	0.200	0.12	0.119	0.128	0.136
max	1	1	0.999	1	1

5.3.2 Aggregate performance metrics

Figure 5.2 shows the ROC curve for each clustering method. HC presents a higher ROC curve than all the others, *control* has the second highest curve, and SOM and k-means have ROC curves that cross, which deems impossible to determine which one is better (Hand, 2010).

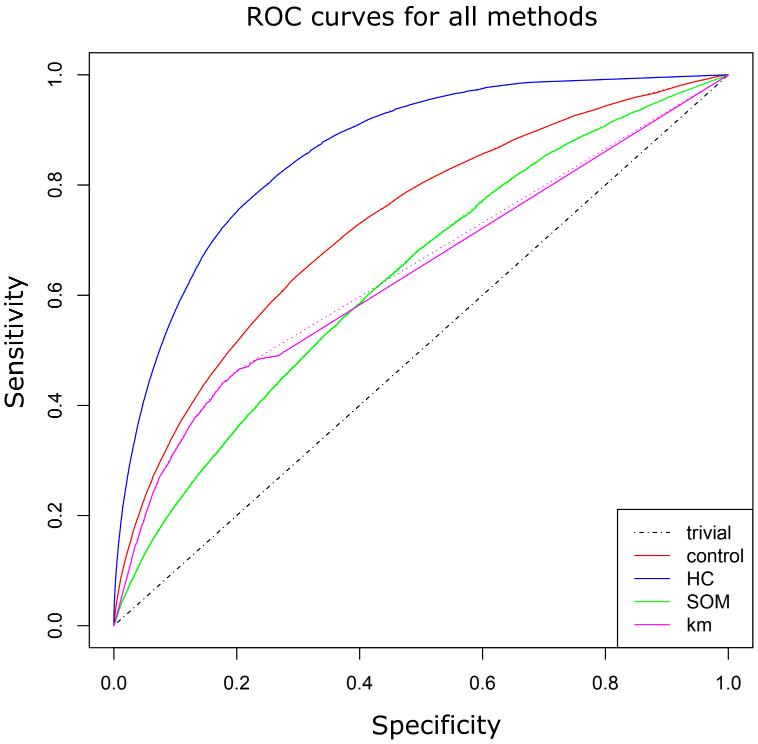


Figure 5.2: ROC space plots for all the methods. Specificity values in Y axis and (1-Sensitivity) values in the X axis

Both performance metrics in Table 5.4 rank the clustering methods in the same order. HC is the clustering method with better performance, followed by *control* and then k-means and SOM. However, the 3rd and 4th position is disputable because their ROC curves cross. Column four shows the Minimum cost-Weighted Error Rate ('MWL') which depicts the threshold values at which, compared to itself, each individual clustering method performs better, accounting for the misclassification cost implicit in the H measure (Anagnostopoulos *et al.*, 2012).

Table 5.4: Aggregate (threshold-independent) performance metrics calculated for each method

	H measure	AUC	MWL
control	0.17	0.73	0.16
HC	0.40	0.86	0.11
SOM	0.06	0.63	0.19
k-means	0.12	0.64	0.17

5.3.3 Results by clustering method

5.3.3.1 Control

The *control* method calculated the relative frequencies of all species in all regions in 2003, and used those values as estimates of risk that the species pose to all the other regions in 2014. This turned out to be a relatively good predictor of species presences in 2014 in terms of AUC (0.73, Table 5.4), and the second best predictor in terms of H measure (0.17, Table 5.4). Also, the gap between control and HC is much larger than in terms of H measure than in terms of AUC, suggesting that *control* might be good at handling absences, that is, true absences and false presences, or in other words, specificity. However, when that factor is controlled for (using the H measure instead) its performance decreases, suggesting that its sensitivity is not as good.

5.3.4 Hierarchical Clustering

Hierarchical clustering had the best overall performance. Its high AUC value, greater than any other method, (0.86, Table 5.4) indicated that the model was good at discriminating the true presences, true absences, false presences and false absences. The same superiority was reported by the H metric. Moreover, the difference between the first and second classified methods in terms of H metric is notable. HC is the first with value 0.40 and the next one is *control* with value 0.17.

5.3.4.1 SOM

The value of AUC of 0.63 for SOM is virtually equal at the value of AUC for k-means, which is 0.64 (both in Table 5.4), but as shown in Figure 5.2 their ROC curves cross each other, therefore, one cannot be ranked superior to the other. Interestingly, SOM produced a very low H measure, of 0.06. Since the H measure is an equivalent measure to AUC that allocates more value to correctly classifying the smaller class (in our case the presences), the big difference between SOM's H measure and SOM's AUC value seems to indicate that SOM is not particularly good at detecting false absences (Type II or omission error) and true presences, the two components of sensitivity. In other words, when specificity and sensitivity are considered equally important (AUC), SOM method obtains a 0.63, while its performance drops to very low (0.06) when sensitivity is more important than specificity (H measure).

5.3.5 K-means

K-means obtained an AUC of 0.64 and a H measure of 0.12 (Table 5.4). Compared to its closest rated method, SOM, k-means performed better in terms of H measure, but virtually equal in terms of AUC. The drop between the value of its AUC and its H measure also shows, as it was the case for the SOM, that when model's sensitivity is considered more important than model specificity, the model loses predictive power.

5.4 Map of the HC clusters

Figure 5.3 shows the spatial distribution of the clusters created by HC which yielded the better results in terms of prediction power between 2003 and 2014. Pink cluster (contains for example Alaska and Greenland) contains the countries or regions with very low number of species (average number of species per region is 16.35). The other three clusters, green (roughly Eurasian regions), blue (southern regions) and purple (United States and Canada) have high and similar number average species per region in 2014 (128.60, 110.84, 109.33, respectively).

All the clusters increased their number of species from 2003 to 2014. Regions in cluster 1 (pink) increased an average of 7 species, regions in cluster 2 (green)increased by an average of 22 species, regions in cluster 3 (blue) an average of 9.75 species and regions in cluster 4 (purple) increased an average of 13.14 species.

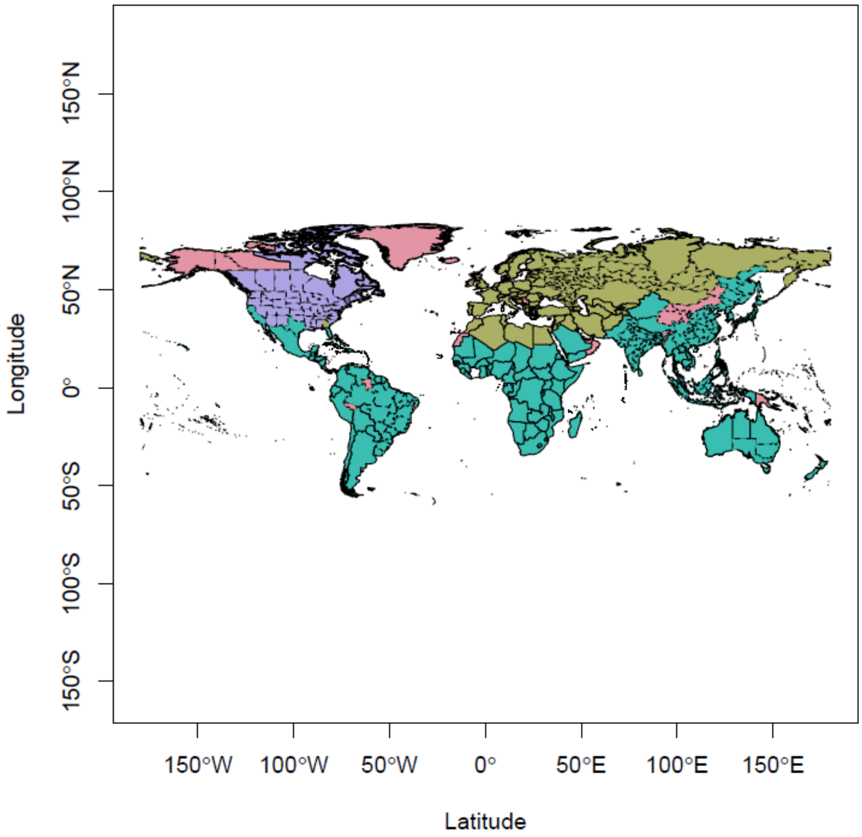


Figure 5.3: Map of the 4 clusters obtained by HC

5.5 Discussion

We compared the performance of three methods for PPA with the aim of elucidating their relative strengths and weaknesses and to identify which was most suitable for PPA. The results indicate there are no simple answers. Although HC clearly performed better in terms of both the performance measures used, there are certain factors that need to be considered.

5.5.1 Issues encountered

First, it could be argued that the better performance of HC might not be attributable to the model *per se*, but to the number of clusters. The results seem to show that high number of clusters (SOM had 99 and k-means had 97) performed worse than a low number of clusters (*control* had 1 cluster and HC had 4). Initially, this study was designed as a comparison among three existing clustering models for PPA. Therefore, we intended to follow the original parametrizations (choice of number of clusters) that were deemed optimal in the original papers (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Watts & Worner, 2009; Eschen *et al.*, 2014). However, after obtaining preliminary results for the performance measures, we observed that differences in model performance could also be caused by different parametrizations and did not necessarily result from the choice of the model. Therefore, we tentatively tested different parametrizations of the three models, when possible. We tried increasing the number of clusters for HC, and decreasing it for SOM and k-means for several different number of clusters. However, it is unconsistent to compare different clustering algorithms using non optimal number of clusters, (for example, we did not try a SOM analysis with only 4 clusters, since this would be wrong from the algorithmic point of view, SOM is known to perform well when the number of clusters is chosen by the formula $c = 5\sqrt{n}$, and suboptimally otherwise) and the results seemed to indicate a dependency of the performance on the cluster number. However, the exploration was not rigorous and not by any means complete, thus, the results were inconclusive. Therefore, we acknowledge the need to further extend this analysis towards a full sensitivity analysis of how the parameter number of clusters changes the performance of the clustering methods in terms of both AUC and H measure. This is a very interesting and needed investigation that should be undertaken in the near future, since it opens interesting questions behind the theoretical hypothesis of PPA.

5.5.2 Number of clusters and environmental filter hypotheses

Our results seemed to indicate that methods with fewer clusters were better at predicting the observed data. If that were true, it could potentially affect the current interpretation of the theoretical hypothesis behind pest profile analyses (PPA).

The ecological hypothesis of environmental filtering (Kraft *et al.*, 2014) suggests that geographical areas with similar pest profiles share biotic and abiotic conditions that shaped the composition of these profiles. In PPA, the environmental filtering hypothesis is used to infer likelihood of pest species exchange between regions with similar composition of pest species assemblages (Worner & Gevrey, 2006; Gevrey *et al.*, 2006; Paini *et al.*, 2010a, 2011; Morin *et al.*, 2013; Singh *et al.*, 2013; Roigé *et al.*, 2016; Roige *et al.*, 2016). It follows that if we divide the 423 world regions into only 4 or 5 clusters (like HC does), the regions grouped together are going to be less similar to one another than if we divide the 423 into 99 or 100 groupings (like SOM and k-means do). We expected that high number of clusters would group more similar regions, as found in Roige *et al.* (2016), where the cluster similarity was assessed by an external similarity metric (ζ diversity) and was found to be high for the majority of the resulting clusters. Highly similar clusters were expected to lead to better predictions, however, results seemed to indicate the contrary, therefore it is necessary to consider whether there might be a conceptual problem within the hypothesis.

One reason why four clusters might predict better than any larger number of clusters could be environmental filtering working at a global scale by defining four large biogeographic regions (see map in Figure 5.3). While it is unarguable that climatic suitability acts as a driver of species distributions, broad-scale patterns in insect diversity are not evident (Diniz-Filho *et al.*, 2010). The local effects such as local habitats and microclimates and insect-plants relationships have a major effect on insects global distributions (Diniz-Filho *et al.*, 2010). However, up-scaling processes known to act at local scales usually fail to explain broadscale patterns of (insect) species distributions because of the complexity of the processes, their interactions and emergent properties (Hortal *et al.*, 2010). Henceforth, there could be a climatic or anthropogenical driver behind the division of world's insect pest assemblages in four groups, that is worth to further explore.

Another potential explanation could be that the global insect assemblages are still in an inceptive state of homogenization. Chapter 4 investigated the state of general similarity found in insect pest assemblages around the world and found that for the same data investigated in this present study, regional insect pest assemblages are still very different (Roige *et al.*, 2016; Bebber *et al.*, 2014a). Also, as shown in Figure 5.1, the majority of the regional profiles are filled with zeroes. It could be that bigger clusters allow for much more room for species exchange between their regions (because they contain more regions) and now, in this moment of time, are a better predictor than fewer smaller clusters. Nevertheless, in a future moment in time when regional species assemblages will be more similar to one another, we will potentially need finer resolution to identify which countries can act as donors and receivers, which translates into investigating finer clusters, that is, clusters that have fewer regions in them but much more similar to one another, as the ones that SOM and k-means methodologies produce.

Related to that, Eschen *et al.* (2014) noted that using relative frequencies as predictors instead than SOM weights does not provide an estimate of risk, thus one can automatically assume a risk of zero for all the species that are not present in a cluster, whereas SOM weights still give some risk value to some of the species outside of the cluster of the target region. However, the relative frequency approach might be conceptually right. If we assume that species pose a risk by being present in the regions that are similar to the region of interest, species that are not present in any of the similar regions (regions in the same cluster) do not present a risk (hence, their risk is zero). Furthermore, the results presented in this research show that, in terms of predicting power, HC with 4 clusters and using relative frequencies, and *control* with no cluster and using relative frequencies, are better predictors than the SOM weights for 99 clusters or k-means relative frequencies for 97 clusters.

While we suggested some potential explanations regarding why fewer clusters could lead to better predictions, we want to emphasize the need to further investigate this controversial result that opens some very interesting questions.

5.5.3 Performance metrics

5.5.3.1 Comission and omission errors equal weights

Using a sole performance measure to judge three different models would have been inadequate, mainly for two reasons. First, each clustering algorithm works differently and can reach different local optima as well as different performance measures are more sensitive towards certain type of errors. Second, performance measures should be carefully applied taking into account what specific questions we are trying to answer. For example, there has been controversy in the field of presence/absence modelling regarding the use of AUC-ROC. One of the relevant remarks is that AUC weights omission and commission errors equally (Lobo *et al.*, 2008) but it has also been shown that in the case of true absences and true presences (which is our case) the AUC, if correctly interpreted, can be informative (Jiménez-Valverde, 2012). For an optimal use of the ROC-AUC technique, we followed the recommendations in Pontius & Parmentier (2014) , which included not presenting AUC as one standalone measure of performance, but provide it along with other insightful measures, and also not presenting the AUC as a single value, but providing as well the plot of each method in the ROC space.

When evaluating the performance of a method for PPA, we are interested in whether the model is able to correctly predict the true presences with few or no false absences. False presences and true absences are less useful for evaluating performance because they are subject to the many stochasticities of species invasions. We discussed in the section 5.2.4 why false presences are not good indicators of model performance for PPA. True absences are similarly less useful for performance evaluation in PPA because true absences provide only weak support for predictions of low risk. A species predicted as low risk that did not establish, maybe did not establish because it did not have a chance to do so, that is, never had the chance to physically arrive to the target region, but might in the future establish, if a new pathway is formed through, for example, a new commercial relationship between two regions. Therefore, we are more interested in false absences, which demonstrate that model predictions were incorrect, and true presences, which demonstrate that model predictions were correct (top row of the confusion matrix, Table 5.2). Both performance metrics used in this study are based in sensitivity and specificity, and therefore, include in their calculations the values of false absences and true presences and , in case of H metric, they give them a higher weight in the computation of the metric. In other words, the H metric weights the commission and omission errors differently, and considers more important the discovery of the smaller class in the dataset, in our case, discovering (and failing to discover) new presences (Hand, 2009; Anagnostopoulos *et al.*, 2012).

5.5.4 Choosing the best method

A model has been validated when its prediction uncertainties are sufficiently small, and that depends on the model objectives but also on other criteria which fall beyond scientific or technical arguments (Usunoff *et al.*, 1992). PPA can be defined as an extrapolation prediction model (it does not extrapolate spatially like species distribution models do, but temporally). PPA uses distributional species data and similarity information from the data to infer future species distributions (areas where the species are susceptible to establish). Prediction models that extrapolate are conceptually simple and require little data (Sutherland, 2006) and that is a desirable attribute in terms of PPA. On the other hand, those models assume that the conditions do not change, and also that we are certain about the drivers of the distributions in the first place (Sutherland, 2006).

The distinction of how model uncertainty originates from model input, model parameters and model structure is problematic (Refsgaard *et al.*, 2007). In this study we have experienced how entangled those uncertainties are for PPA. Uncertainty arose from the chose of models (which clustering method to chose), from the parameters of the models (what number of clusters), from the data itself (how many errors were in the original 2003 and the observed 2014 validation data), from the choice of the performance measures (which predictive characteristics of the model are better explained by the AUC or the H measure). But most importantly, the conceptual uncertainty that arose from the realization that maybe fewer bigger clusters were able to predict better than a higher number of smaller clusters.

An experiment is deemed satisfactory for model discrimination if simulations with the same data using all models are significantly different (Usunoff *et al.*, 1992). In this study we have provided different results for at least one of the methods. In other words, HC has been ranked as the best method by the performance metrics that we have used. However, our results are less than conclusive. We have encountered new sources of variability such as the number of clusters that need to be further studied before we can conclude anything about the superiority of any method. The causes of the failure to conclude are that our experimental designed did not initially account for the parameter uncertainty of each method and also because this parameter uncertainty discovered possible conceptual errors in the models design.

5.5.5 Implications for PPA

To judge which is the best model to cluster regional pest profiles for PPA we must be mindful of risk analysts' main applications. It is notable that all three clustering methods are better than random. A risk assessor faced with the challenge of defining quarantine measures and trade regulations and a large list of potentially hazardous species would likely benefit from relatively quick methods of filtering species such as the clustering methods presented in this study. Another interesting result is that *control*, which estimated how geographically widespread a species is, is a good predictor of risk. This suggests that risk analysts under strong time pressure could obtain robust estimates of invasion risk using the relative frequency of species presence in different world regions.

5.5.6 Perspectives and future research

This study has clearly shown the need for further research. There are two main questions that arise from this comparison. First question is; is there an effect of the number of clusters on the prediction performance of the methods? And the second question is: is there a flaw in the assumption behind PPA? It is accepted that conceptual uncertainty seems to comprise a big source of prediction uncertainty (Refsgaard *et al.*, 2007), that means, sometimes the hypothesis behind our models are not exhaustive enough. In our case, there is plenty of biological and ecological sense behind the hypothesis of assemblage similarity and environmental filtering, but maybe we are not looking at the right scale (number of clusters) to identify the proper drivers. Future research is needed to design a test for the hypothesis of the number of clusters. Ideally, a sensitivity analysis of the parameter number of clusters should be conducted for all the methods, in order to conclude whether it actually is a source of distortion in the predictions or, on the contrary, the apparent relationship found in the present study is just spurious.

The second question opened by the results of this study is to investigate, for the given clustering results for each method, which sort of 'natural groupings' the clustering methodologies have revealed. It would be very interesting to see, for example for the 4 resulting clusters of HC, what are the drivers behind and compare them to the three potential explanations we gave earlier in this discussion.

Although it must be acknowledged the highly hypothetical character of the following discussion, I proceed to explore some ideas regarding what could be the drivers behind the four clusters shown in 5.3. The first considered argument considered was the climatic or biogeographical driver. The map in Figure 5.3 shows rough broadscale climatic patterns that seem to be consistent with Earth's biogeographic realms. Biogeographic realms are divisions of Earth that account for the biogeographical patterns of biotic organisms, and correspond to floristc kingdoms or zoogeographic regions. The 4 clusters obtained by HC mostly group Palearctic regions together in green cluster, Nearctic regions in purple cluster, and then gathers together Oceania, Neotropic, Afrotropic, Indo-Malay and Australasia biogeographic realms in blue cluster (Olson *et al.*, 2001). However, since the blue cluster groups five biogeographic realms, we considered further potential drivers.

Another option is that the map is showing some anthropogenically influenced pattern. We tentatively explored the possibility that the map was reflecting the origin and length of stay of the species, that is, assuming that most of the pest species recorded in CABI CPC would have Palearctic origin (green cluster) then they would have been exported or migrated to Nearctic region (purple cluster) and are more recently arriving to the third big cluster, the blue one that contains the southern world regions. However, it could also be that the direction of the species movement was another. For example, that is the purple cluster (Nearctic, or USA and Canada mostly) sending species to other two clusters.

Again, any of these hypothesis are highly speculative and would indeed require of a thorough investigation each.

However, the investigations of these hypothesis need to be preceded by a full sensitivity analyses of the parameter number of clusters for each one of the clustering methodologies, that clarifies whether the resulting map of 4 clusters using HC is indeed a valid result.

5.6 Conclusions and Future work

1- Comparing current clustering methods for PPA revealed interesting features of the methods. Hierarchical clustering with few number of clusters showed the best predictive power.

2- Predictive power seems to diminish with increasing number of clusters, which is an intriguing result that warrants further investigation. Potential explanations have been provided, but there is a need for a complete sensitivity analyses of the parameter number of clusters for each method.

3- All three clustering methods investigated are suitable for PPA, and have potential to provide useful first filters for risk analysts. They might also identify some risk species that could be overlooked using qualitative approaches based on expert-knowledge. However, knowledge of each models' limitations must be clearly communicated to risk analysts.

4- The species worldwide prevalence (relative frequency) seems to provide good predictive power of future establishments.

Chapter 6

General Discussion

6.1 Preamble

Quantitative methods for pest risk assessment are a valuable tool to risk assessors and biosecurity agencies worldwide (Jarrad *et al.*, 2015). A plethora of quantitative approaches have been described in recent years (reviewed by Leung *et al.* (2012)) and more are being developed at the moment. Venette *et al.* (2010) suggested many priorities for further development of quantitative methods for pest risk assessment. This thesis focused on one of those methods; pest profile analysis (PPA).

Pest profile analysis is a young approach to pest risk prioritisation, first described in Worner & Gevrey (2006). Several studies have subsequently applied it (Paini *et al.*, 2010b, 2011; Morin *et al.*, 2013; Singh *et al.*, 2013, 2015; Qin *et al.*, 2015; Paini *et al.*, 2016), though few studies have tested its validity. Worner *et al.* (2013) listed the studies that performed model validation and sensitivity analyses of the PPA approach between 2006 and 2013. They comprised: A sensitivity analyses of the SOM algorithm to the changes of status from present to absent (Worner & Souquet, 2010); A comparison of SOM PPA predictions based on 2007 data to observations made in 2011 (Suiter, 2011); and a validation using simulated data, where SOM PPA was used to rank fungal species in a virtual simulated world (Paini *et al.*, 2011).

Worner *et al.* (2013) recommended that research in PPA continued particularly to develop protocols for: Conducting comparable studies; detecting and removing outliers; choosing the initial number of clusters; validating clusters; and reconciling information obtained from different clustering methods. However, these recommendations have not yet been fully implemented. The research presented in this thesis is the first to address a big number of Worner *et al.* (2013) recommendations such as thoroughly testing PPA for all target regions, comparing three clustering methods (Chapter 5), evaluating theoretical assumptions of the model (Chapter 4), and making methodological improvements (Chapter 2, Chapter3).

In summary, this research has addressed each one of the objectives presented in the Introduction by performing a sensitivity analysis of SOM PPA (Chapter 2), validating SOM outputs by ecologically explaining clustering results (Chapter 3 and Chapter 4), measuring the level of biotic homogenization among global crop pest profiles (Chapter 4), comparing the performance of different clustering methods for PPA (Chapter 5) and assessing and discussing the validity of inferring risks of invasion from clustering results (Chapter 5).

6.2 Methodological improvements

This thesis has addressed several of Venette et al.'s (Venette *et al.*, 2010) high priority recommendations for improving quantitative methods for pest risk assessment:

- A recommendation to **provide greater documentation of model development and assessment** was addressed in Chapters 1-4 by unifying the terminology used in previous research, creating a graphical representation to explain the steps that comprise SOM PPA and by simplifying the data pre-preprocessing time by showing that is not necessary to remove any regions from the analysis.

- A recommendation to **improve models' representation of uncertainty** was addressed in Chapter 3 by incorporating the ζ metric a measure of uncertainty for SOM PPA outputs. Moreover, the evaluation of cluster validity and uncertainty assessment using ζ metric can be transferred to the other two clustering techniques used in PPA, k-means and HC.

- A recommendation to **increase communication with decision makers on the interpretation of model outputs** was addressed in Chapter 2 by recommending the use of a ranked list of species instead of SOM weights to communicate the model outputs, and in Chapter 3 by suggesting the incorporation of ζ metric as a simple visual depiction of uncertainty.

6.3 Model validation

Validating a model involves showing that it meets certain performance requirements and is, thus, suitable for its intended use (Rykiel, 1996). Validation attempts to define the degree of confidence users should have in a model (Power, 1993). In general, models should be accompanied by a sensitivity or uncertainty analysis showing how their theoretical premises, the data, and uncertainties associated with the premisses and the data influence models' outputs (Kirchner *et al.*, 1996; Venette *et al.*, 2010).

The concept of validation can be ambiguous and can be subject to differing interpretations, thus, different authors approached validation in different ways. In this thesis, different types of validation tests have been conducted for PPA. This discussion will first focus on 'operational' which test how well models' outputs meet the standards required for their purpose (Rykiel, 1996). I then discuss the 'conceptual validity' of the PPA approach, which is concerned with whether the theories and assumptions underlying the models are correct or at least justifiable (Rykiel, 1996).

6.3.1 Operational validation tests

A comprehensive list of different validation procedures can be found in Rykiel (1996), and a several of them have been applied in this thesis. They are: 1) Comparing one model to another ('comparative test'); 2) Showing that a model gives consistent results from the same data and parameters ('internal model validity test'); 3) Assessing model outputs when a parameter is set at an extreme value ('extreme conditions test'); 4) Making predictions then comparing them with real observed data ('historical data validation test'); 5) Testing for the model's ability to reproduce proper relationships between variables, regardless of their quantitative values ('event validity test'). An internal model validity test, extreme conditions test, sensitivity analysis and assessment of model validity for PPA were all performed in Chapter 2 by changing the initial data input of the model to datasets with very few species and recording the model results in those extreme conditions. The historical data validation test was conducted in Chapter 5 for all the three clustering methods.

Sensitivity analysis is an important component of validation because it helps understand how the model responds to different data or parameters. A sensitivity analysis measures how a model's outputs change when one parameter value is varied. In Chapter

2, the sensitivity of SOM PPA to the input data was found to be significant. However, it also showed that ranked lists of species are a more stable measure of invasion risk than SOM weights for individual species. This means that ranks would be more comparable between studies, whilst also easier to understand for a risk analyst.

Subsequent research (Chapter 5) suggested the sensitivity analysis conducted in Chapter 2 might not have been comprehensive enough. It suggested that additional parameters such as the 'number of clusters' should have been studied for the SOM, and also for the HC and k-means clustering methods. This parameter was excluded from the Chapter 2 sensitivity analysis because each clustering method has its own numerically optimal strategy for defining it. However, Chapter 5 showed that the mathematically optimal choice for each algorithm might not be conceptually optimal for PPA. Rykiel (1996) argued that frequently there is a disparity between the parameters the natural system is sensitive to and the parameters the model is sensitive to. Chapter 5 suggests that PPA is conceptually sensitive to the number of clusters, whereas the algorithms used to perform PPA have a different mathematical sensitivity to the same parameter. In Chapter 3, an event validity test was applied to SOM PPA by showing that the resulting clusters were ecologically meaningful. The ζ diversity metric was used to test the validity of the regions clustered by the SOM, which helped to validate some clusters and discard others. In Chapter 5, observed data was used to validate the predictions of each clustering model. Chapter 5 was the first time the PPA method was validated using observed data for all target regions, and it included not only SOM, but also HC and k-means.

6.3.2 Conceptual validation of ecological principles

The fields of philosophy and ecology have both sustained extended discussions about different approaches to hypothesis testing. In general, scientific method follows a deductive approach, exemplified and formalized in Platt (1964). That is, for a specific problem, potential explanations are explicitly listed, experiments or tests are conducted and incorrect hypotheses are systematically eliminated leaving fewer possibilities within which the truth must lie (Quinn & Dunham, 1983; Platt, 1964). However, some authors have argued that in community ecology this approach is often infeasible, because it deals with problems at a systems level where interactions are complex, composite effects are common, true controls are rare, replicates are difficult to obtain and experiments take too long (Hobbs *et al.*, 2006). Moreover, many causes can contribute to an observed pattern, which means hypotheses may not be mutually exclusive. Thus, the model of 'strong inference' formalized by Platt (1964) is frequently inapplicable to community ecology questions. Instead, Quinn & Dunham (1983) and Hobbs *et al.* (2006) advocate for a more inductive approach. This involves identifying potential causes for a problem, evaluating the contributions of each cause to the main problem, and cautiously generalizing these contributions to other situations. This process of knowledge discovery is an inductive process called 'learning by accumulation of evidence' (Quinn & Dunham, 1983; Hobbs *et al.*, 2006), and is contrary to the deductive process of 'falsifying wrong hypotheses' formalized by Platt (1964). However, Simberloff (2010) argues that any natural event (not only in community ecology) is likely to be caused by many actors, and this multiplicity of causes is a poor excuse for using an inductive process. Therefore, Simberloff (2010) and Marquet *et al.* (2014) advocate the application of Platt's (1964) 'strong inference' approach. They also suggest that searching for confirmatory evidence is always easier, and more seductive, than searching for falsification.

The PPA methodology is based on fundamental ecological principles of community assembly. The main ecological hypothesis behind PPA is that two regions with similar pest profiles share a set of environmental and historical conditions that make them suitable for the assembly of similar sets of pest species. Therefore, identifying similar regions can be used to identify regions that are likely to share pest species in the future, thus, providing predictive power to pest risk assessment.

This hypothesis can be divided into several less-inclusive hypotheses. The first is that species assembly is a non-random process, or at least not entirely random, but influenced by a set of underlying distributional mechanisms. The second is that identifying regions with similar pest assemblages enables us to predict potential future establishments. The third is that this predictive power arises from biotic and abiotic similarities between regions.

The alternative to hyphotesis one, that 'assemblage composition is a random process' has been investigated. Hubell (2001) described the neutral theory of biodiversity and biogeography(NTB), which assumes all individuals of all species are competitively identical and any trait variation between species has no influence on their abundance and speciation rates (Mcgill *et al.*, 2006). An important implication of the NTB is that communities are assembled by random stochastic processes. Thus it contradics 100 years of community ecology theory, which explained species distributions by niche differentiation (Chave, 2004) whereby differing characteristics between species and their complex interactions drive community composition.

Null model tests are interesting for studying community assembly (Gotelli & Graves, 1996). They are randomization models that hold some elements of ecological data constant and allow others to stochastically vary, to create new assemblage patterns. These are the patterns expected in the absence of a particular assemblage mechanism. The NTB hypothesis can be tested by conducting null model tests of community composition datasets (Gotelli & Mcgill, 2006). If a community significantly differs from the null model, then there is likely to be an underlying niche assembly process operating. The NTB hypothesis was tested for regional pest assemblages via a null model analysis conducted by Watts & Worner (2009). The PPA concept rests upon the assumption that insect crop pest assemblages are a non-random collection of species. This thesis did not explicitly test hypotheses relating to the NTB, but did find evidence of biotic homogenization (Chapter 4). This implies that pest assemblages are becoming more similar, thus, their assembly is unlikely to be purely random.

The second hypothesis is that similarities between pest profiles can be used to predict future establishments. The alternative is that 'pest profiles have no predictive power', which, Chapter 5 suggested is incorrect because using species assemblages to predict future establishments provided better predictions than the control. What remains unclear is the number of clusters that maximizes predictive power. Chapter 5 showed that fewer clusters probably yield better predictive power than more clusters but reasons for this remain unknown, and further investigation is needed.

The third hypothesis is that predictive power arises from biotic and abiotic similarities between regions. PPA assumes that the predictive power of clustering invasive species assemblages is related to the biotic and abiotic conditions that shape assemblage composition through the process of environmental filtering. However, which abiotic and biotic conditions are the important drivers of species assemblages' composition or presence in a region, is often difficult to determine for many species. The relative influence of abiotic and biotic conditions on species distribution is not a new question in community ecology. Climatic suitability often plays an important role in defining species distributions (Roura-Pascual *et al.*, 2011; Eyre *et al.*, 2012; Sutherst, 2014; Bebber *et al.*, 2014a; Williamson, 2006). However, some species, particularly invasive species, can establish in new regions with climates that differ from those of their native distribution (Guisan & Edwards, 2002), thus, we need to remain cautious when using only climatic suitability as a predictor. For example, in Chapter 5 it was shown that the world biogeographical realms appear to be an important variable influencing pest species distributions. Other important variables are human modification of habitats (Roura-Pascual *et al.*, 2011) and, in the case of agricultural plant pests, the presence of a host plant of the pest(Bebber *et al.*, 2014b,a). On the other hand, herbivorous pests have been shown to have host-shifts, mostly towards other hosts phylogenetically related to the original ones (Lewinsohn *et al.*, 2005), thus, the presence or absence of the host should not be the only factor to take into account when explaining insect pest species distributions. PPA, specially SOM PPA, creates meaningful groups of regions. Chapter 3 showed that these groupings had biological meaning by assessing them using the ζ metric but also showed how they appeared to account for many of the known drivers of species distributions, which are; climate suitability, host availability, trade relationships between countries and geographical proximity (or spatial dependency). However, whether those drivers are enough to explain pest species distributions, and whether the predictive power of clustering pest assemblages lies in the ability of the PPA method to reflect those drivers remains unclear. Chapter 5 indicated that the best predictive power was reached using a method, HC with 4 clusters, that roughly divided the world in four groups of regions depicting biogeographical realms, compared with the SOM, which is able to divide the world into around 90 groups of regions that do present big similarities. Therefore, the predictive power of clustering pest profiles seems to lie in broad scale similarity patterns rather than in fine-scale high similar groupings.

6.4 Final remarks

There was insufficient time in this study to further explore why few large clusters yielded better predictive power than numerous small clusters (Chapter 5). However, I encourage further research to explain this. Future PPA research would benefit both from improvements to the quality of the CABI CPC data, which are known to contain errors and misclassifications, and from access to additional data sources. However, reliable global data are difficult to obtain and the process of data acquisition, validation and codification is very time consuming. Overall, this research has emphasized the usefulness of quantitative models in pest risk assessment, and has provided evidence that the investigated quantitative methods are valid predictors of species distributions that can, indeed, convert scientifically relevant data into decision relevant information.

References

Aggarwal, C. C. (2009). *Managing and mining uncertain data*. Berlin: Springer.

Anagnostopoulos, C., Hand, D. J., & Adams, N. M. (2012). Measuring classification performance : the hmeasure package .

Anderson, M. J., Crist, T. O., Chase, J., Vellend, M., Inouye, B., Freestone, A. L., Sanders, N. J., Cornell, H., Comita, L., Davies, K., Harrinson, S., Kraft, N., Stegen, J., & Swenson, N. (2011). Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. *Ecology letters*, *14*(1), 19–28.

Arnold B . Erickson (1960). Review Reviewed Work (s): The Ecology of Invasions by Animals and Plants by Charles S . Elton. *The Journal of Wildlife Management*, *24*(2), 231–233.

Ascough, J. C., Maier, H. R., Ravalico, J. K., & Strudley, M. W. (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*, *219*(3-4), 383–399.

Astel, A., Tsakovski, S., Barbieri, P., & Simeonov, V. (2007). Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water research*, *41*(19), 4566–78.

Bação, F., & Lobo, V. (2010). Introduction to Kohonen's Self-Organising Maps. *Instituto Superior de Estatistica E Gestao de Informacao*, (p. 22).

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. *Computational ScienceâĂŞICCS 2005*, *3516*, 476 – 483.

Bacon, S. J., Aebi, A., Calanca, P., & Bacher, S. (2014). Quarantine arthropod invasions in Europe: The role of climate, hosts and propagule pressure. *Diversity and Distributions*, *20*(1), 84–94.

Bacon, S. J., Bacher, S., & Aebi, A. (2012). Gaps in Border Controls Are Related to Quarantine Alien Insect Invasions in Europe. *PLoS ONE*, *7*(10), 1–9.

Baker, R., Cannon, R., Bartlett, P., & Barker, I. (2005). Novel strategies for assessing and managing the risks posed by invasive alien species to global crop production and biodiversity. *Annals of Applied Biology*, *146*(2), 177–191.

Barlow, N. D., & Goldson, S. (2002). Alien invertebrates of New Zealand. In D. Pimentel (Ed.) *Biological Invasions. Economic and environmental costs of alien plant, animal, and microbe species*, (pp. 195–217). New York: CRC Press.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., & Grothendieck, G. (2015). Package âĂŸlme4'.

Bebber, D. P. (2015). Range-Expanding Pests and Pathogens in a Warming World. *Annual Review of Phytopathology*, *53*, 335–356.

Bebber, D. P., Holmes, T., & Gurr, S. J. (2014a). The global spread of crop pests and pathogens. *Global Ecology and Biogeography*, *23*, 1398–1407.

Bebber, D. P., Holmes, T., Smith, D., & Gurr, S. J. (2014b). Economic and physical determinants of the global distributions of crop pests and pathogens. *New Phytologist*, *202*(3), 901–910.

Bedfort, T., & Cooke, R. (2001). *Probabilistic Risk Analysis: Foundations and Methods.*. Cambridge: Cambridge University press.

Bevington, P., & Robinson, D. (2002). *Data reduction and error analysis for the physical sciences*. New York: McGraw-Hill.

Blackburn, T. M., Essl, F., Evans, T., Hulme, P. E., Jeschke, J. M., Kühn, I., Kumschick, S., Marková, Z., Mrugała, A., Nentwig, W., Pergl, J., Pyšek, P., Rabitsch, W., Ricciardi, A., Richardson, D. M., Sendek, A., Vilà, M., Wilson, J. R. U., Winter, M., Genovesi, P., & Bacher, S. (2014). A Unified Classification of Alien Species Based on the Magnitude of their Environmental Impacts. *PLoS biology*, *12*(5).

Blackburn, T. M., Pyšek, P., Bacher, S., Carlton, J. T., Duncan, R. P., Jarosík, V., Wilson, J. R. U., & Richardson, D. M. (2011). A proposed unified framework for biological invasions. *Trends in Ecology and Evolution*, *26*(7), 333–339.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-s. S. (2008). Generalized linear mixed models : a practical guide for ecology and evolution. (Table 1).

Bradshaw, C., Leroy, B., Bellard, C., Albert, C., Roiz, D., Barbet-Massin, M., Fournier, A., Salles, J.-M., Simard, F., & Courchamp, F. (2016). Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, *7*(September 16), 14.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231.

Burgman, M. (2005). *Risks and decisions for conservation and environmental management*. Cambridge: Cambridge University press.

Burns, K. C. (2015). A Theory of Island Biogeography for Exotic Species. *The American Naturalist*, *186*(4), 441–451.

CABI (2007). Crop Protection Compedium, Global Module.

Chave, J. (2004). Neutral theory and community ecology. *Ecology letters*, *7*(3), 241–253.

Chon, T., Park, Y., Moon, K., & Cha, E. (1996). Patternizing communities by using an artificial neural network. *Ecological Modelling*, *90*(1), 69–78.

Chon, T.-S. (2011). Self-Organizing Maps applied to ecological sciences. *Ecological Informatics*, *6*(1), 50–61.

Chown, S. L. (2015). Editorial overview: Global change biology: Insects in a hot, crowded and connected world. *Current Opinion in Insect Science*, *11*, iv–vi.

Clout, M. (2002). Biological Invasions. Economic and Environmental costs of alien plant, animal and microbe species. In *Biological Invasions. Economic and Environmental costs of alien plant, animal and microbe species*, chap. Ecological, (pp. 185–195).

Cornell, H. V., & Harrison, S. P. (2014). What Are Species Pools and When Are They Important ? *Annual Review of Ecology, Evolution, and Systematics*.

Davies, D. L., & Bouldin, D. W. (1979). IEEE Transactions on pattern analysis and machine intelligence. (2), 224–227.

Deboeck, G. J., & Kohonen, T. (1998). *Visual exploration in finance with self organizing maps*. Springer Finances.

Dengler, J. (2009). Which function describes the species-area relationship best? A review and empirical evaluation. *Journal of Biogeography*, *36*(4), 728–744.

Díaz, S., Fargione, J., Chapin, F. S., & Tilman, D. (2006). Biodiversity loss threatens human well-being. *PLoS Biology*, *4*(8), e277.

Diniz-Filho, J. A. F., de Marco, P., & Hawkins, B. A. (2010). Defying the curse of ignorance: Perspectives in insect macroecology and conservation biogeography. *Insect Conservation and Diversity*, *3*(3), 172–179.

Elton, C. S. (1958). *The ecology of invasions by animals and plants*. Chicago: Kluwer Academic Publishers, university ed.

EPPO (2004). EPPO Standards Diagnostic protocols for regulated pests. Protocoles de diagnostic pour les organismes r{é}glement{é}s.

Eschen, R., Holmes, T., Smith, D., Roques, A., Santini, A., & Kenis, M. (2014). Likelihood of establishment of tree pests and diseases based on their worldwide occurrence as determined by hierarchical cluster analysis. *Forest Ecology and Management*, *315*, 103–111.

Eyre, D., Baker, R. H. A., Brunel, S., Dupin, M., Jarošik, V., Kriticos, D. J., Makowski, D., Pergl, J., Reynaud, P., Robinet, C., & Worner, S. (2012). Rating and mapping the suitability of the climate for pest risk analysis. *EPPO Bulletin*, *42*(1), 48–55.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/ absence models. *Environmental Conservation*, *24*(1), 38–49.

Funtowicz, S., & Ravetz, J. (1990). *Uncertainty and Quality in Sicence for Policy*. Dodrecht: Kluwer Academic Publishers.

Gevrey, M., Worner, S. P., Kasabov, N., Pitt, J., & Giraudel, J.-L. (2006). Estimating risk of events using SOM models: A case study on invasive species establishment. *Ecological Modelling*, *197*(3-4), 361–372.

Giera, N., Bell, B., Jones, C., Warburton, B., & Cowan, P. (2009). Economic Costs of Pests to New Zealand. Tech. rep., MAF Biosecurity New Zealand.

Giraudel, J., & Lek, S. (2001). A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling*, *146*(1-3), 329–339.

Gleason, H. A. (1922). On the Relation Between Species and Area. *Ecology*, *3*(2), 158–162.

Gonzalez, A., Cardinale, B. J., Allington, G. R. H., Byrnes, J., Endsley, K. A., Brown, D. G., Hooper, D. U., Isbell, F., O'Connor, M. I., & Loreau, M. (2016). Estimating local biodiversity change: A critique of papers claiming no net loss of local diversity. *Ecology*, *97*(8), 1949–1960.

Gonzalez, P., Neilson, R. P., Lenihan, J. M., & Drapek, R. J. (2010). Global patterns in the vulnerability of ecosystems to vegetation shifts due to climate change. *Global Ecology and Biogeography*, *19*(6), 755–768.

Gotelli, N. J., & Graves, G. (1996). *Null models in ecology*. Washington DC: Smithsonian Institution Press.

Gotelli, N. J., & Mcgill, B. J. (2006). Null versus neutral models: what's the difference? *Ecography*, *29*(5).

Guillemaud, T., Ciosi, M., Lombaert, É., & Estoup, A. (2011). Biological invasions in agricultural settings: Insights from evolutionary biology and population genetics. *Comptes Rendus - Biologies*, *334*(3), 237–246.

Guisan, A., & Edwards, T. C. (2002). Generalized linear and generalized additi v e models in studies of species distributions : setting the scene. *157*.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2/3), 107–145.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123.

Hand, D. J. (2010). Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Statistics in Medicine*, *29*(14), 1502–1510.

Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, *34*(5), 492–495.

Hand, D. J., & Anagnostopoulos, C. (2014). A better Beta for the H measure of classification performance. *Pattern Recognition Letters*, *40*(1), 41–46.

He, F., & Legendre, P. (1996). On Species-Area Relations. *The American Naturalist*, *148*(4), 719–737.

HilleRisLambers, J., Adler, P., Harpole, W., Levine, J., & Mayfield, M. (2012). Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics*, *43*(1), 227–248.

Hobbs, R. J., Arico, S., Aronson, J., Baron, J. S., Bridgewater, P., Cramer, V. A., Epstein, P. R., Ewel, J. J., Klink, C. A., Lugo, A. E., Norton, D., Ojima, D., Richardson, D. M., Sanderson, E. W., Valladares, F., Vil??, M., Zamora, R., & Zobel, M. (2006). Novel ecosystems: Theoretical and management aspects of the new ecological world order. *Global Ecology and Biogeography*, *15*(1), 1–7.

Hofstetter, P. (1998). *Perspectives in Life Cycle Impact Assessment: A structured approach to combine models of the technosphere, ecosphere and valuesphere*. 12. Dordrecht: Kluwer Academic Publishers.

Honkela, T. (1997). *Natural language processing with self organizing maps*. Ph.D. thesis, Helsinki University of Technology.

Horan, R. D., & Lupi, F. (2010). The economics of invasive species control and management: The complex road ahead. *Resource and Energy Economics*, *32*(4), 477–482.

Hortal, J., Roura-Pascual, N., Sanders, N. J., & Rahbek, C. (2010). Understanding (insect) species distributions across spatial scales. *Ecography*, *33*(1), 51–53.

Hubell, S. P. (2001). *A unified neutral theory of biodiversity and biogeography*. Oxfordshire: Princeton University Press.

Hui, C., & McGeoch, M. a. (2014). Zeta Diversity as a Concept and Metric That Unifies Incidence-Based Biodiversity Patterns. *The American Naturalist*, *184*(5), 684–694.

Huijbregts, M., Norris, G., Bretz, R., A, C., von Bahr, B., Maurice, B., Weidema, B., & Beaufort, A. S. H. D. (2001). Framework for Modelling Data Uncertainty in Life Cycle Inventories. *International Journal*, *6*(Lci), 127–132.

Hulme, P. (2003). Biological invasions: winning the science battles but losing the conservation war? *Oryx*, *37*(02), 178–193.

Hulme, P. E. (2009). Trade , transport and trouble : managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, (pp. 10–18).

Hulme, P. E. (2011). Practitioner's perspectives: Introducing a different voice in applied ecology. *Journal of Applied Ecology*, *48*(1), 1–2.

International Plant Protection Convention (IPPC) (2004). Pest Risk Analysis for quarantine pests. Tech. Rep. 11, United Nations Food and Agriculture Organization, Rome.

International Plant Protection Convention (IPPC) (2006a). ISPM 01 Phytosanitary principles for the protection of plants and the application of phytosanitary measures in international trade. Tech. rep., United Nations Food and Agriculture Organization, Rome.

International Plant Protection Convention (IPPC) (2006b). ISPM No. 2 Guidelines for pest risk analysis. Tech. Rep. 2, Food and Agrigulture Organization of the United Nations.

International Plant Protection Convention (IPPC) (2007). ISPM 2 Framework for pest risk analysis. Tech. rep., United Nations Food and Agriculture Organization, Rome.

IUCN/SSC (2000). IUCN Guidelines for the Prevention of Biodiversity Loss Caused by Alien Invasive Species. Tech. Rep. May.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323.

Jarrad, F., Low-Choy, S., & Mengersen, K. (2015). *Biosecurity surveillance: quantitative approaches*.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, *21*(4), 498–507.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate stastical analysis*. New Jersey: Pearson Education, Inc., 6th ed.

Joint Committee for Guides in Metrology (JCGM) (2008). ISO:2008 Evaluation of measurement data: Guide to the expression of uncertainty in measurement. Tech. Rep. September.

Jost, L., Chao, A., & Chazdon, R. L. (2011). Compositional similarity and beta diversity. In *Biological Diversity. Frontieers in Measurement and Assessment*, chap. 6, (pp. 66–84). Oxford University Press.

Kirchner, J. W., Hooper, R. P., Kendall, C., Neal, C., & Leavesley, G. (1996). Testing and validating environmental models. *Science of the Total Environment*, *183*(1-2), 33–47.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. In *IEEE International Conference on Neural Networks*.

Knight, J., Worner, S. P., Griessinger, D., & Souquet, A. (2011). Enhancements of pest risk analysis techniques. Specification for neural network application in PRA. Tech. Rep. 212459, EPPO.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Kohonen, T. (1990). The self-organizing map. *Neurocomputing*, *21*(1-3), 1–6.

Kohonen, T. (2001). *The self-organizing maps*. New York: Springer- Verlag Berlin Heidelberg GmbH, third ed.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks*, *37*, 52–65.

Kolar, C. S., & Lodge, D. M. (2001). Progress in invasion biology: predicting invaders. *Trends in ecology & evolution*, *16*(4), 199–204.

Kovács, F., Legány, C., & Babos, A. (2005). Cluster Validity Measurement Techniques. In *6th International symposium of hungarian researchers on computational intelligence*.

Kraft, N. J. B., Adler, P. B., Godoy, O., James, E., Fuller, S., & Levine, J. M. (2014). Community assembly, coexistence, and the environmental filtering metaphor. *Functional Ecology*, *29*(5), 592–599.

Krzanowski, W., & Hand, D. J. (2009). *ROC curves for continuous data*. Boca Raton: Chapman and Hall.

Krzanowski, W., & Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, *44*(1), 23–34.

Kuussaari, M., Ekroos, J., & Helio, J. (2010). Homogenization of lepidopteran communities in intensively cultivated agricultural landscapes. (pp. 459–467).

Latombe, G., McGeoch, M. A., & Hui, C. (2015). Package 'zetadiv '.

Lawton, J. H., & Gaston, K. J. (1989). Temporal Patterns in the Herbivorous Insects of Bracken : A Test of Community Predictability Author. *Journal of Animal Ecology*, *58*(3), 1021–1034.

Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, *129*(2), 271–280.

Lek, S., & Guégan, J. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological modelling*, *120*, 65–73.

Leung, B., Roura-Pascual, N., Bacher, S., Heikkilä, J., Brotons, L., Burgman, M. a., Dehnen-Schmutz, K., Essl, F., Hulme, P. E., Richardson, D. M., Sol, D., Vilà, M., & Rejmanek, M. (2012). Teasing apart alien species risk assessments: a framework for best practices. *Ecology letters*, *15*(12), 1475–93.

Lewinsohn, T. M., Novotny, V., & Basset, Y. (2005). INSECTS ON PLANTS : Diversity of Herbivore Assemblages Revisited. *Annual Review of Ecology, Evolution, and Systematics*, *36*, 597–620.

Liebhold, A. M., Yamanaka, T., Roques, A., Augustin, S., Chown, S. L., Brockerhoff, E. G., & Pyšek, P. (2016). Global compositional variation among native and non-native regional insect assemblages emphasizes the importance of pathways. *Biological Invasions*, *18*(4), 893–905.

Liu, S., Hurley, M., Lowell, K. E., Siddique, A. B. M., Diggle, A., & Cook, D. C. (2011). An integrated decision-support approach in prioritizing risks of non-indigenous species in the face of high uncertainty. *Ecological Economics*, *70*(11), 1924–1930.

Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151.

Lockwood, J. L., & McKinney, M. L. (2001). *Biotic Homogenization*. Springer.

Loreau, M. (2002). Biodiversity loss and the maintenance of our life-support system. In M. Loreau, S. Naeem, & P. Inchausti (Eds.) *Challenges of a Changing Earth*, chap. 32, (pp. 169–173). Oxford, UK: Oxford University Press.

Low-Choy, S. (2015). Getting the story straight: Laying the foundations for statistical evaluation of the performance surveillance. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.) *Biosecurity Surveillance. Quantitative approaches*, chap. 3, (pp. 43–72). Rome: CAB International, 1st ed.

Lustig, A. (2016). *Complex systems analysis of invasive species in heterogeneous environments*. Ph.D. thesis, Lincoln University.

Mack, R. N., Simberloff, D., Lonsdale, W., Evans, H., Michael Clout, & Bazzaz, F. A. (2000). Biotic Invasions: Causes, epidemiology, global consequences, and control. *Bulletin of the Ecological Society of America*, *10*(November 1999), 689–710.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(233), 281–297.

Maier, H., Ascough, J., Wattenbach, M., Renschler, C., Labiosa, W., & Ravalico, J. (2008). Uncertainty in environmental decision making: issues, challenges and future directions. In A. J. Jakeman, A. A. Voinov, A. E. Rizzoli, & S. H. Chen (Eds.) *Environmental Modelling, Software and Decision Support: State of the Art and New Perspectives*, chap. 5. Elsevier Science.

Mangiameli, P., Chen, S. K., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, *93*(2), 402–417.

Marchetti, M. P., Lockwood, J. L., & Light, T. (2006). Effects of urbanization on California's fish diversity: Differentiation, homogenization and the influence of spatial scale. *Biological Conservation*, *127*(3), 310–318.

Marquet, P. a., Allen, a. P., Brown, J. H., Dunne, J. a., Enquist, B. J., Gillooly, J. F., Gowaty, P. a., Green, J. L., Harte, J., Hubbell, S. P., O'Dwyer, J., Okie, J. G., Ostling, a., Ritchie, M., Storch, D., & West, G. B. (2014). On Theory in Ecology. *BioScience*, *64*(8), 701–710.

Mathworks, T. (2013). MATLAB.

Matott, L. S., Babendreier, J. E., & Purucker, S. T. (2009). Evaluating uncertainty in integrated environmental models : A review of concepts and tools. *Water Resources Research*, *45*(6), 1–14.

McGeoch, M. a., Butchart, S. H. M., Spear, D., Marais, E., Kleynhans, E. J., Symes, A., Chanson, J., & Hoffmann, M. (2010). Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. *Diversity and Distributions*, *16*(1), 95–108.

McGeoch, M. a., Chown, S. L., & Kalwij, J. M. (2006). A global indicator for biological invasion. *Conservation biology*, *20*(6), 1635–46.

McGeoch, M. A., Genovesi, P., Bellingham, P. J., Costello, M. J., McGrannachan, C., & Sheppard, A. (2016). Prioritizing species, pathways, and sites to achieve conservation targets for biological invasion. *Biological Invasions*, *18*(2), 299–314.

McGeoch, M. A., Spear, D., Leynhans, E. J. K., & Marais, E. (2012). Uncertainty in invasive alien species listing. *Ecological Applications*, *22*(3), 959–971.

Mcgill, B. J., Maurer, B. A., Weiser, M. D., Mcgill, J., Maurer, A., & Weiser, D. (2006). Empirical Evaluation of Neutral Theory. *Ecology*, *87*(6), 1411–1423.

McKinney, M. L. (2004). Do exotics homogenize or differentiate communities? Roles of sampling and exotic species richness. *Biological Invasions*, *6*(4), 495–504.

McKinney, M. L., & Lockwood, J. L. (1999). Biotic homogenization: A few winners replacing many losers in the next mass extinction. *Trends in Ecology and Evolution*, *14*(11), 450–453.

McLeod, A. (2015). The relationship between biosecurity surveillance and risk analysis. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.) *Biosecurity Surveillance. Quantitative approaches*, chap. 5, (pp. 109–120). Rome: CAB International, 1st ed.

McNeely, J., Mooney, H., Neville, L., Schei, P., & Waage, J. (2001). Global Strategy on Invasive Alien Species. Tech. rep., CAB International.

Millar, R. B., Anderson, M. J., & Tolimieri, N. (2011). Much ado about nothings: using zero similarity points in distance decay curves. *Ecology*, *92*(9), 1717–1722.

Mingoti, S. a., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, *174*(3), 1742–1759.

Mirkin, B. (2013). *Clustering. A data recovery approach*. Chapman & Hall, 2nd ed.

Morgan, M., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis.*. Cambridge University press.

Morin, L., Paini, D. R., & Randall, R. P. (2013). Can Global Weed Assemblages Be Used to Predict Future Weeds ? *PloS one*, *8*(2), e55547.

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, *6*(2), 143–152.

Oja, M., Kaski, S., & Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. *Neural ...*, (pp. 1–156).

Oksanen, A. J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., Hara, R. B. O., Simpson, G. L., Solymos, P., Stevens, M. H. H., & Wagner, H. (2016). Package vegan.

Olden, J. D. (2006). Biotic homogenization: A new research agenda for conservation biogeography. *Journal of Biogeography*, *33*(12), 2027–2039.

Olden, J. D., Comte, L., & Giam, X. (2016). Biotic Homogenisation. *eLS, John Wiley and Sons*, (pp. 1–8).

Olden, J. D., & Poff, N. L. (2003). Toward a mechanistic understanding and prediction of biotic homogenization. *The American naturalist*, *162*(4), 442–460.

Olden, J. D., Poff, N. L., Douglas, M. R., Douglas, M. E., & Fausch, K. D. (2004). Ecological and evolutionary consequences of biotic homogenization. *Trends in Ecology and Evolution*, *19*(1), 18–24.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. a., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, *51*(11), 933.

Paini, D. R., Bianchi, F. J. J. A., Northfield, T. D., & De Barro, P. J. (2011). Predicting invasive fungal pathogens using invasive pest assemblages: testing model predictions in a virtual world. *PloS one*, *6*(10), e25695.

Paini, D. R., Sheppard, A. W., Cook, D. C., Barro, P. J. D., Worner, S. P., & Thomas, M. B. (2016). Global threat to agriculture from invasive species. *Pnas*, *113*(27), 7575–7579.

Paini, D. R., Worner, S. P., Cook, D. C., De Barro, P. J., & Thomas, M. B. (2010a). Threat of invasive pests from within national borders. *Nature Communications*, *1*, 115.

Paini, D. R., Worner, S. P., Cook, D. C., De Barro, P. J., & Thomas, M. B. (2010b). Using a self-organizing map to predict invasive species: sensitivity to data errors and a comparison with expert opinion. *Journal of Applied Ecology*, *47*(2), 290–298.

Park, Y.-S., Céréghino, R., Compin, A., & Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling*, *160*(3), 265–280.

Perrings, C., Dehnen-Schmutz, K., Touza, J., & Williamson, M. (2005). How to manage biological invasions under globalization. *Trends in Ecology and Evolution evolution*, *20*(5), 212–5.

Pimentel, D., McNair, S., Janecka, J., Wightman, J., Simmonds, C., O'Connell, C., Wong, E., Russel, L., Zern, J., Aquino, T., & Tsomondo, T. (2001). Economic and environmental threats of alien plant, animal, and microbe invasions. *Agriculture, Ecosystems and Environment*, *84*(1), 1–20.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*, *52*(3), 273–288.

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Pontius, R. G., & Parmentier, B. (2014). Recommendations for using the relative operating characteristic (ROC). *Landscape Ecology*, *29*(3), 367–382.

Power, M. (1993). The predictive valN1ation of ecological and environmental models. *Ecological Modelling*, *68*(1-2), 33–50.

Pyšek, P., Richardson, D. M., Rejmánek, M., Webster, G. L., Kirschner, J., Pygekl, P., Rejmanek, M., Williamson, M., & Kirschnerl, J. (2004). Alien plants in checklists and floras: towards better communication between taxonomists and ecologists. *Taxon*, *53*(1), 131–143.

Qin, Y., Paini, D. R., Wang, C., Fang, Y., & Li, Z. (2015). Global establishment risk of economically important fruit fly species (Tephritidae). *PLoS ONE*, *10*(1), 1–9.

Quinlan, M., Stanaway, M., & Mengersen, K. (2015). Biosecurity surveillance in agriculture and environment: a Review. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.) *Biosecurity Surveillance. Quantitative approaches*, chap. 2, (pp. 9–42). Wallingford, UK.: CAB International, 1st ed.

Quinn, J. F., & Dunham, A. (1983). On Hypothesis Testing in Ecology.

R Foundation for Statistical Computing (2014). R: A language and environment for statistical computing.

Rahel, F. J. (1990). The Hierarchical Nature of Community Persistence : A Problem of Scale. *The American Naturalist*, *136*(3), 328–344.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process - A framework and guidance. *Environmental Modelling and Software*, *22*(11), 1543–1556.

Regan, H. M., Colyvan, M., & Burgman, M. A. (2002). A TAXONOMY AND TREATMENT OF UNCERTAINTY FOR ECOLOGY AND CONSERVATION BIOLOGY. *Ecological Applications*, *12*(2), 618–628.

Richardson, D. M., & Pyšek, P. (2012). Naturalization of introduced plants: Ecological drivers of biogeographical patterns. *New Phytologist*, *196*(2), 383–396.

Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Dane Panetta, F., & West, C. J. (2000). Naturalization and invasion of alien plants: Concepts and definitions. *Diversity and Distributions*, *6*(2), 93–107.

Richardson, D. M. D., Pysek, P., Carlton, J. T., Pyšek, P., & Carlton, J. T. (2011). A compendium of essential concepts and terminology in invasion ecology. In *Fifty Years of Invasion Ecology: The Legacy of Charles Elton*, (pp. 409–420).

Richardson, D.M., Pyšek, P., Simberloff, D., Rejmánek, M., Mader, A. (2008). Biological invasions âĂŞ the widening debate: a response to Charles Warren. *Progress in Human Geography*, *32*(2), 295–298.

Ricklefs, R. E. (2004). A comprehensive framework for global patterns in biodiversity. *Ecology letters*, (pp. 1–15).

Roige, M., McGeoch, M. A., Hui, C., & Worner, S. P. (2016). Cluster validity and uncertainty assessment for self organizing map pest profile analysis. *Methods in Ecology and Evolution*, *8*(3), 349–357.

Roigé, M., Parry, M., Phillips, C., & Worner, S. (2016). Self-organizing maps for analysing pest profiles: Sensitivity analysis of weights and ranks. *Ecological Modelling*, *342*, 113–122.

Roques, A. (2012). Biological invasion. *Integrative zoology*, *7*(3), 227.

Rosenzweig, M. L. (2001). The four questions: What does the introduction of exotic species do to diversity? *Evolutionary Ecology Research*, *3*(3), 361–367.

Roura-Pascual, N., Hui, C., Ikeda, T., Leday, G., Richardson, D. M., Carpintero, S., Espadaler, X., Gómez, C., Guénard, B., Hartley, S., Krushelnycky, P., Lester, P. J., McGeoch, M. a., Menke, S. B., Pedersen, J. S., Pitt, J. P. W., Reyes, J., Sanders, N. J., Suarez, A. V., Touyama, Y., Ward, D., Ward, P. S., & Worner, S. P. (2011). Relative roles of climatic suitability and anthropogenic influence in determining the pattern of spread in a global invader. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(1), 220–5.

Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling*, *90*(3), 229–244.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell Systematic Technology Journal*, *27*, 379–423.

Silcock, P., & Guy, N. (2013). Incursions threaten $0b by 2020. *NZ Grower*, (pp. 36–37).

Simberloff, D. (2010). Competition theory, hypothesis testing and other community ecology buzzwords. *The American Naturalist*, *122*(5), 626–635.

Simberloff, D. (2011). How common are invasion-induced ecosystem impacts? *Biological Invasions*, *13*(5), 1255–1268.

Singh, S. K., Ash, G. J., & Hodda, M. (2015). Keeping one step ahead of invasive species: using an integrated framework to screen and target species for detailed biosecurity risk assessment. *Biological Invasions*, *17*, 1069–1086.

Singh, S. K., Paini, D. R., Ash, G. J., & Hodda, M. (2013). Prioritising plant-parasitic nematode species biosecurity risks using self organising maps. *Biological Invasions*, *16*(7), 1515–1530.

Sørensen, T. J. (1948). *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.*. København, I kommission hos E. Munksgaard, biologiske ed.

Suiter, K. A. (2011). Progress on SOM Analysis A comparison of pre-emergent invasive pest lists using distribution data obtained from the GPDD and CABI databases. Tech. rep.

Sutherland, W. J. (2006). Predicting the ecological consequences of environmental change: A review of the methods. *Journal of Applied Ecology*, *43*(4), 599–616.

Sutherst, R. W. (2014). Pest species distribution modelling : origins and lessons from history. *Biological Invasions*, (pp. 239–256).

Sutherst, R. W., & Bourne, a. S. (2008). Modelling non-equilibrium distributions of invasive species: a tale of two modelling paradigms. *Biological Invasions*, *11*(6), 1231–1237.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic.

Törönen, P., Kolehmainen, M., Wong, G., & Castrén, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS letters*, *451*, 142–146.

Tuomisto, H. (2010). A diversity of beta diversities : straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, *33*(November 2009), 2–22.

United States Environmental Protection Agency, U.-E. (1997). Exposure Factors Handbook. Tech. Rep. August 1997.

Uriarte, E. A., & Martín, F. D. (2006). Topology Preservation in SOM. *World Academy of Science, Engineering and Technology*, *21*, 52–55.

Usunoff, E., Carrera, J., & Mousavi, S. F. (1992). An approach to the design of experiments for discriminating among alternative conceptual models.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015a). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*, *63*, 24–31.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015b). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling and Software*, *63*, 24–31.

Vänninen, I., Worner, S., Huusela-veistola, E., Tuovinen, T., & Nissinen, A. (2011). Recorded and potential alien invertebrate pests in Finnish agriculture and horticulture. *Agricultural and food science*, *20*(July 2010), 96–114.

Vellend, M. (2001). Do commonly used indices of beta diversity measure species turnover? *Journal of vegetation science*, *12*, 545–552.

Vellend, M. (2010). Conceptual synthesis in community ecology. *85*(2), 183–206.

Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beauséjour, R., Brown, C. D., De Frenne, P., Verheyen, K., & Wipf, S. (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19456–9.

Vellend, M., Verheyen, K., Flinn, K. M., Jacquemyn, H., Kolb, A., Van Calster, H., Peterken, G., Graae, B. J., Bellemare, J., Honnay, O., Brunet, J., Wulf, M., Gerhardt, F., & Hermy, M. (2007). Homogenization of forest plant communities and weakening of species-environment relationships via agricultural land use. *Journal of Ecology*, *95*(3), 565–573.

Venette, R. C., Kriticos, D. J., Magarey, R. D., Koch, F. H., Baker, R. H. A., Worner, S. P., Raboteaux, N. N. G., McKenney, D. W., Dobesberger, E. J., Yemshanov, D., De Barro, P. J., Hutchinson, W. D., Fowler, G., Kalaris, T. M., & Pedlar, J. (2010). Pest risk maps for invasive alien species. A roadmap for improvement. *Biosicence*, *60*, 349–362.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on neural networks*, *11*(3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (1999). Self-organizing map in Matlab : the SOM Toolbox. *Proceedings of the Matlab DSP Conference*, (pp. 35–40).

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). Self organizing map in MATLAB: The SOM Toolbox. Tech. rep., Helsinki University of Technology, Helsinki.

Vilà, M. (2013). 36 Responses on Invasive Species. In D. Simberloff (Ed.) *Invasive Species. What Everyone Needs to Know*, vol. 342, (p. 424). New York: Oxford University Press.

Vitousek, P. M. (1990). Biological invasions and ecosystem processes: towards an integration of population biology and ecosystem studies. *Oikos*, *57*(1), 183–191.

Vitousek, P. M., D'Antonio, C. M., Loope, L. L., & Westbrooks, R. (1996). Biological invasions as a Global environmental change. *American Scientist*, *84*(5), 468–478.

Walesiak, M. (2016). Package âĂŸ clusterSim '.

Walker, W., Harremoes, P., Rotmans, J., Van der Sluijs, J., Van Asselt, M. B., Janssen, P., & Von Krauss, M. K. (2003). Defining Uncertainty. A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment*, *4*(1), 5–17.

Waller, N. G., Kaiser, H. A., Illian, J. B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika*, *63*(1), 5–22.

Walther, G.-R., Roques, A., Hulme, P. E., Sykes, M. T., Pysek, P., Kühn, I., Zobel, M., Bacher, S., Botta-Dukát, Z., Bugmann, H., Czúcz, B., Dauber, J., Hickler, T., Jarosík, V., Kenis, M., Klotz, S., Minchin, D., Moora, M., Nentwig, W., Ott, J., Panov, V. E., Reineking, B., Robinet, C., Semenchenko, V., Solarz, W., Thuiller, W., Vilà, M., Vohland, K., & Settele, J. (2009). Alien species in a warmer world: risks and opportunities. *Trends in ecology & evolution*, *24*(12), 686–93.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function.

Warren, C. R. (2007). Perspectives on the 'alien' versus 'native' species debate: a critique of concepts, language and practice. *Progress in Human Geography*, *31*(4), 427–446.

Warren, C. R. (2008). Alien concepts: a response to Richardson et al. *Progress in Human Geography*, *32*(2), 299–300.

Warziniack, T. W., Finnoff, D., & Shogren, J. F. (2013). Public economics of hitchhiking species and tourism-based risk to ecosystem services. *Resource and Energy Economics*, *35*(3), 277–294.

Wattenbach, M., Gottschalk, P., Hattermann, F., Rachimow, C., Flechsig, M., & Smith, P. (2006). A framework for assessing uncertainty in ecosystem models. In *Proceedings of the iEMSs Thrid Biennial Meeting: Summit on Environmental Modeling and Software*. Burlington.

Watts, M., & Worner, S. (2011). Improving cluster-based methods for investigating potential for insect pest species establishment: region-specific risk factors. *Computational Ecology and Software*, *1*(3), 138–145.

Watts, M., & Worner, S. (2012). Using artificial neural networks to predict the distribution of bacterial crop diseases from biotic and abiotic factors. *2*(1), 70–79.

Watts, M. J., & Worner, S. P. (2009). Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. *Ecological Modelling*, *220*(6), 821–829.

Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon*, *21*(2), 213–251.

Williamson, M. (1989). mathematical models of invasion.pdf. In J. Drake (Ed.) *Biological Invasions: a Global Perspective*, (pp. 329–350). SCOPE.

Williamson, M. (2006). Explaining and predicting the success of invading species at different stages of invasion. *Biological Invasions*, *8*(7), 1561–1568.

Williamson, M., & Fitter, A. (1996). The Varying Success of Invaders. *Ecology*, *77*(6), 1661–1666.

Worner, S., Gevrey, M., Eschen, R., Kenis, M., Paini, D., Singh, S., Watts, M., & Suiter, K. (2013). Prioritizing the risk of plant pests by clustering methods; self-organising maps, k-means and hierarchical clustering. *NeoBiota*, *18*, 83–102.

Worner, S. P. (1991). Use of models in applied entomology: The need for perspective. *Environmental Entomology*, *20*(3), 768–773.

Worner, S. P. (2002). Predicting the invasive potential of exotic insects. In G. J. . Halman, & C. P. Schwalbe (Eds.) *Invasive Arthropods in Agriculture. Problems and solutions*, chap. 7, (pp. 119–137). Science Publishers Inc.

Worner, S. P., & Gevrey, M. (2006). Modelling global insect pest species assemblages to determine risk of invasion. *Journal of Applied Ecology*, *43*(5), 858–867.

Worner, S. P., & Souquet, A. (2010). A retrospective analysis of the use of ecological theory and pest species assemblages to prioritise pests. In *4th International Pest Risk Modelling and Mapping workshop: Pest risk in a changing world*. Port Douglas.

Zaki, M. J., & Meira, M. J. (2013). *Data Mining and Analysis: Fundamental Concepts and Algorithms*.

Zurell, D., Berger, U., Cabral, J., & Jeltsch, F. (2010). The virtual ecologist approach: simulating data and observers. *Oikos*, *119*(4), 622–635.

Appendix A

Supplement chapter 2

A.1 Weights variability

Histogram of the weight of 199 species over 341 datasets

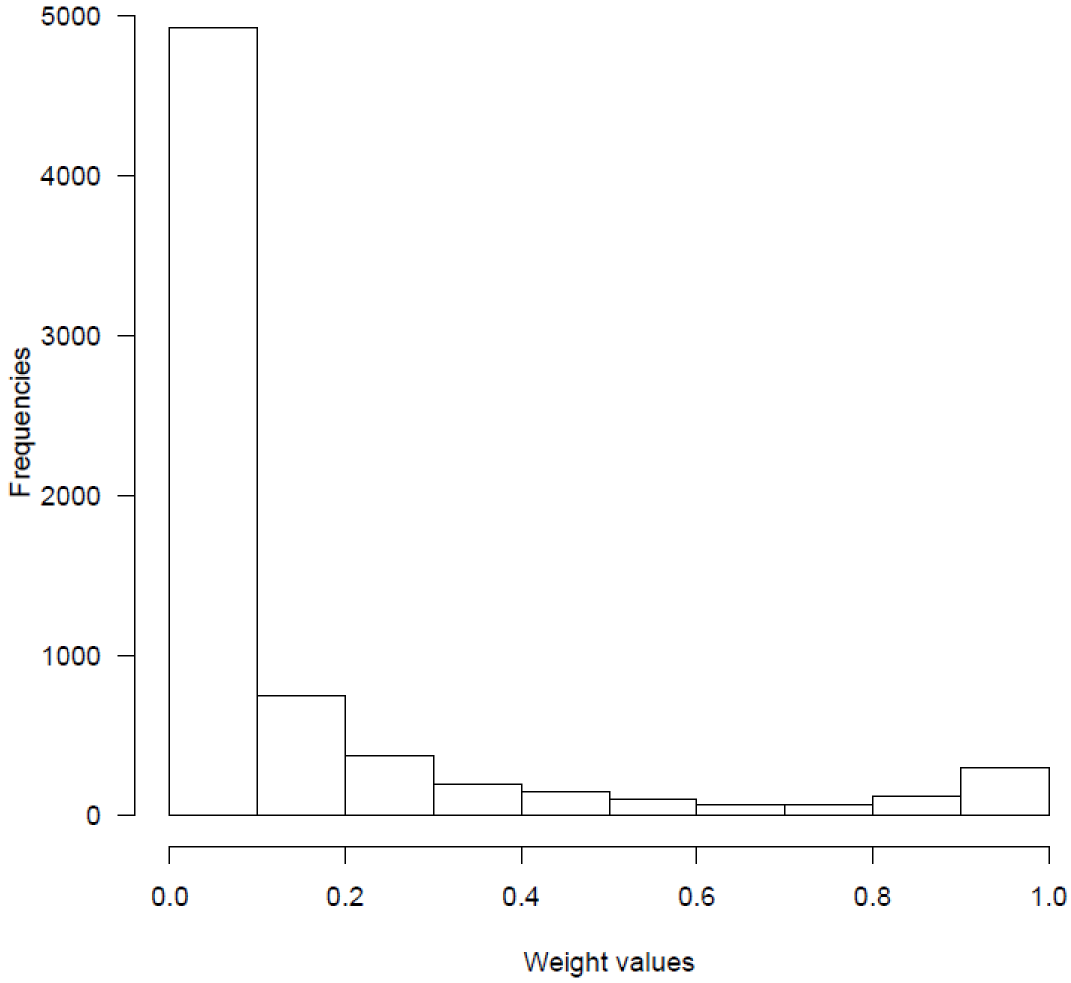


Figure A.1: Histogram of weight values for 199 species

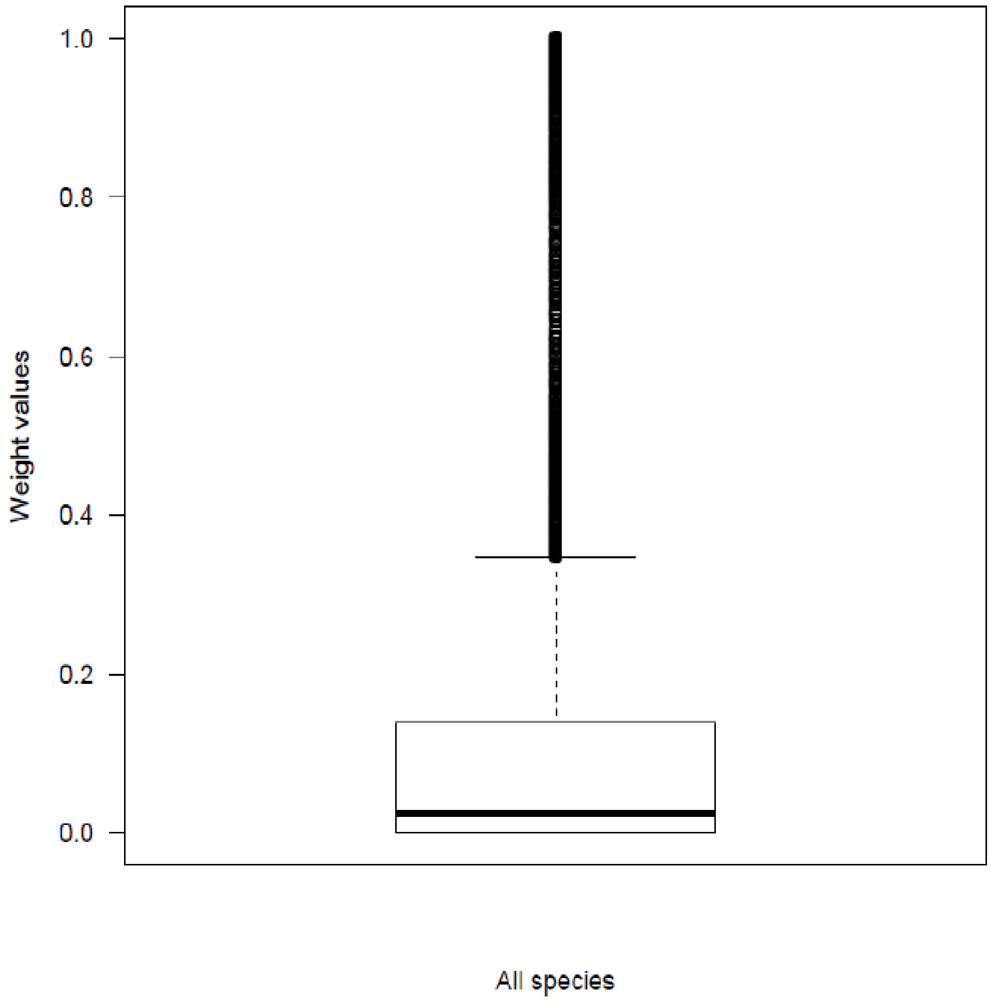


Figure A.2: Boxplot of weight values for 199 species

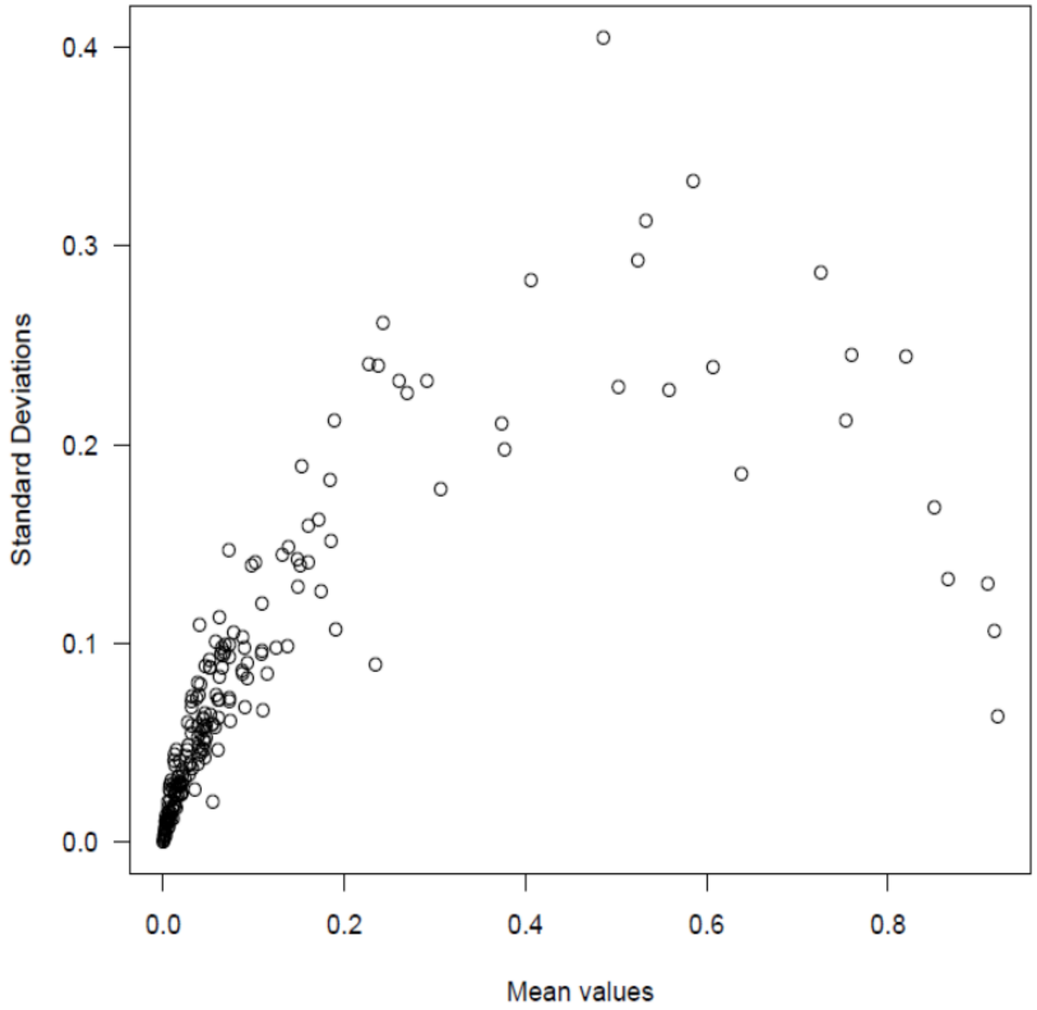


Figure A.3: Relationship between Mean and SD of weight values

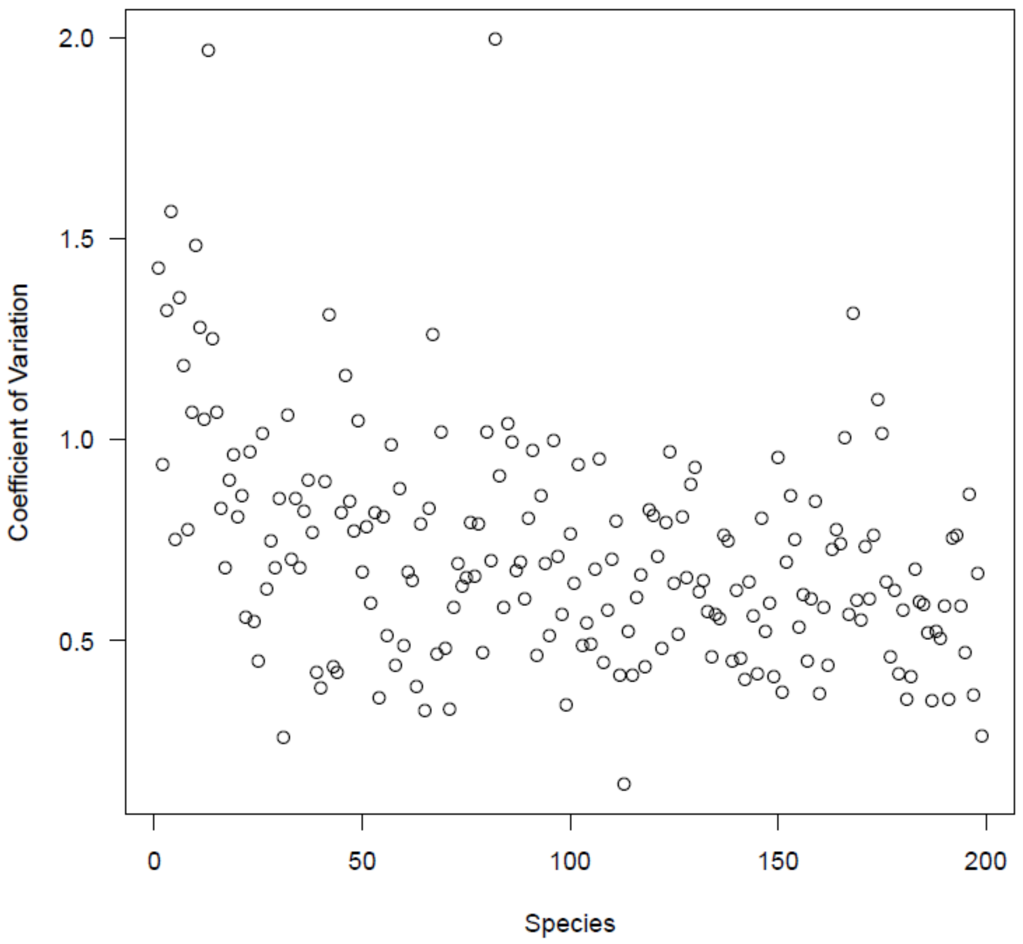


Figure A.4: Scatterplot of CV values for 199 species

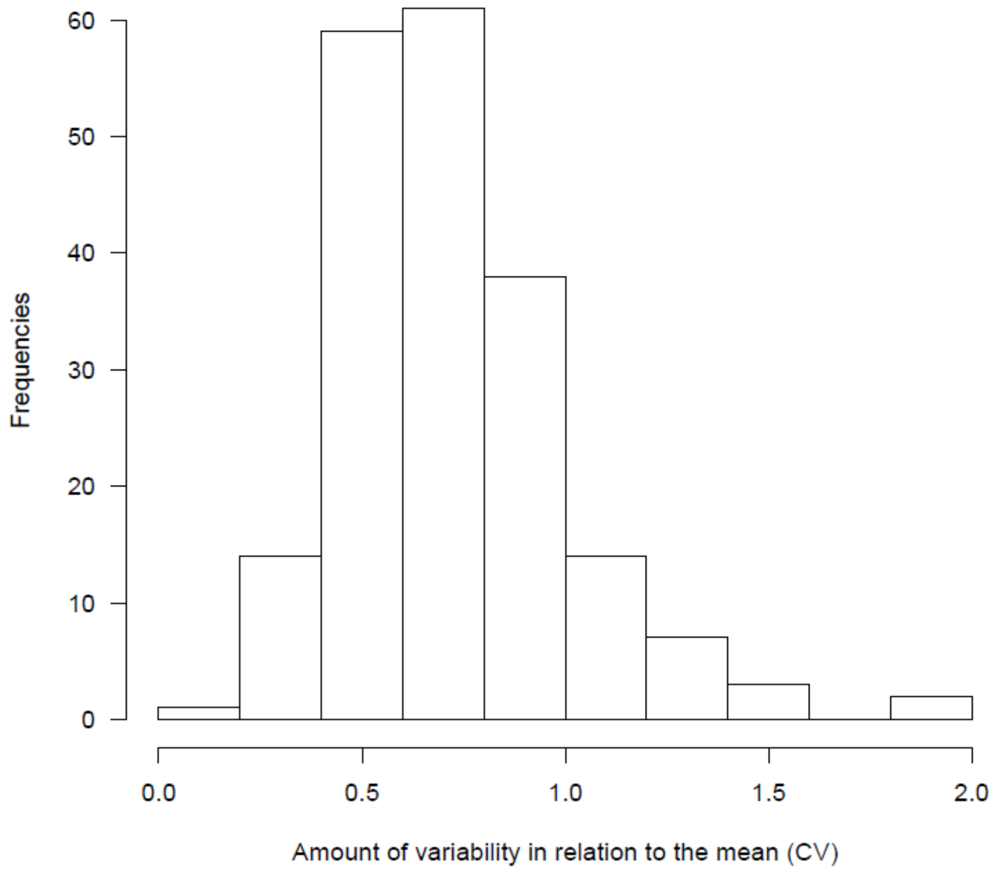


Figure A.5: Histogram of CV values

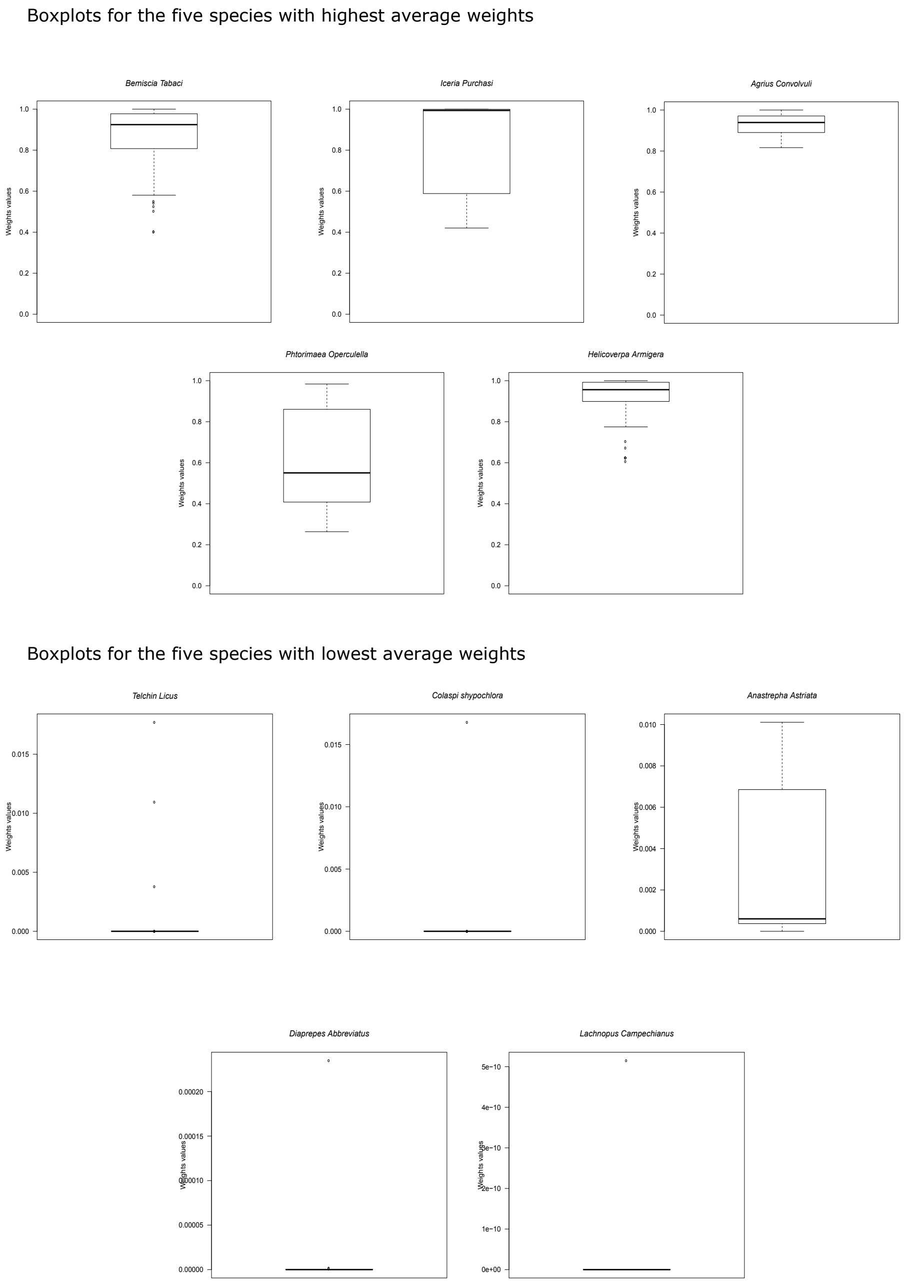


Figure A.6: Boxplots of the weight value for 5 top and 5 bottom average weights species. Mid-line indicates median value and whiskers depict variability outside the upper and lower quartiles

A.2 Ranks variability

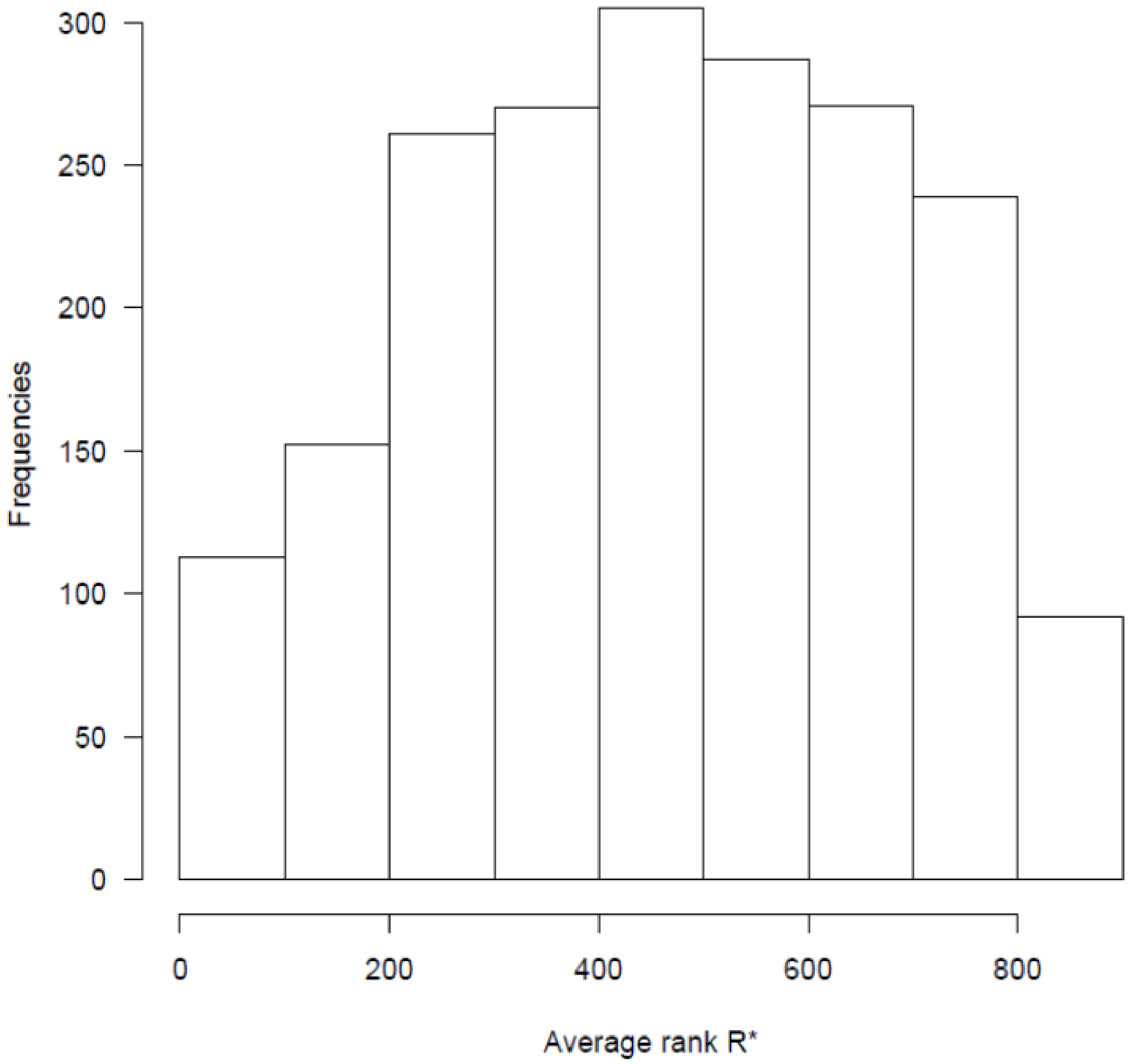


Figure A.7: Values of the R^* for the subset of 199 species

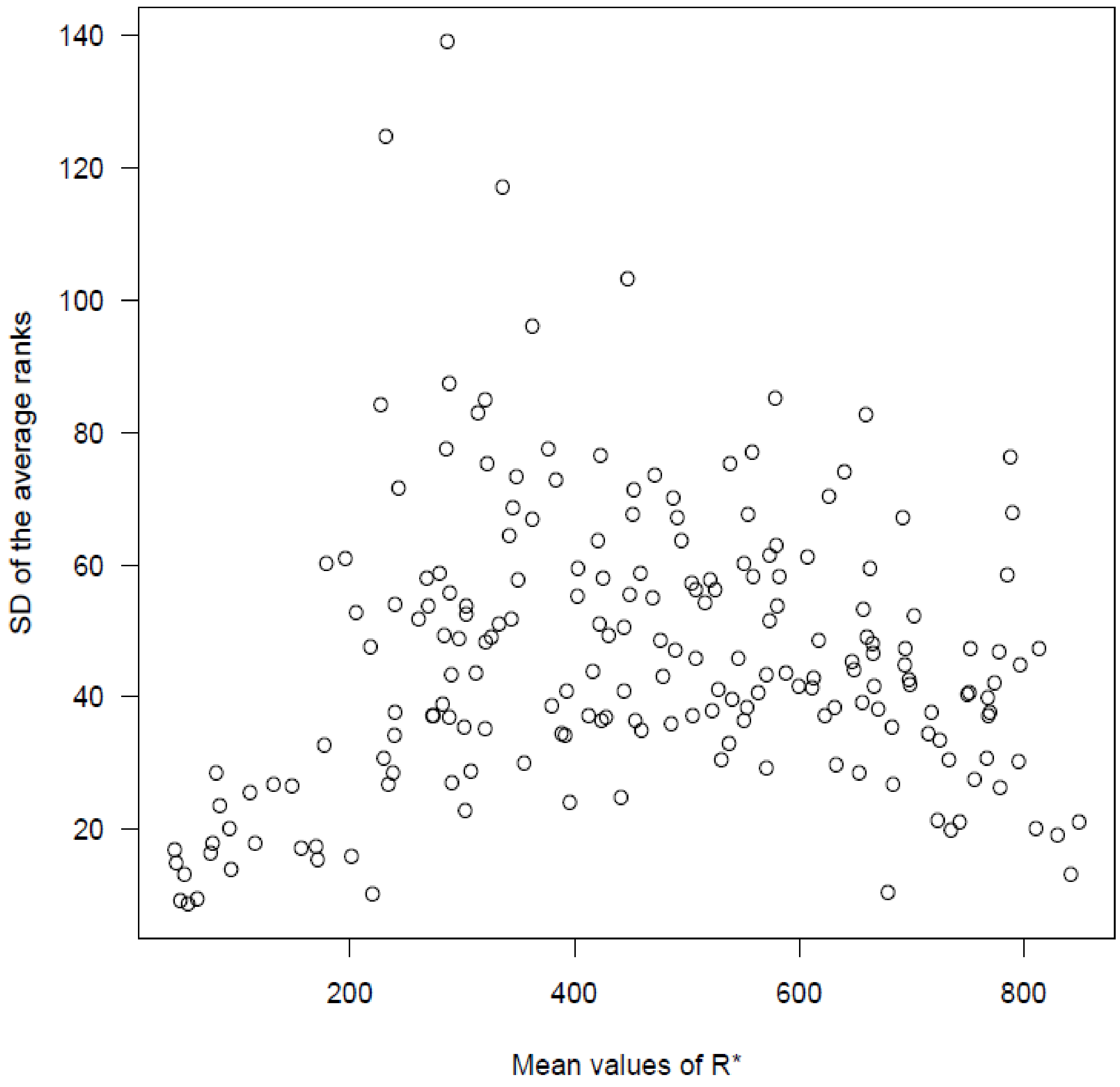


Figure A.8: Relationship between mean and SD of the R^* value

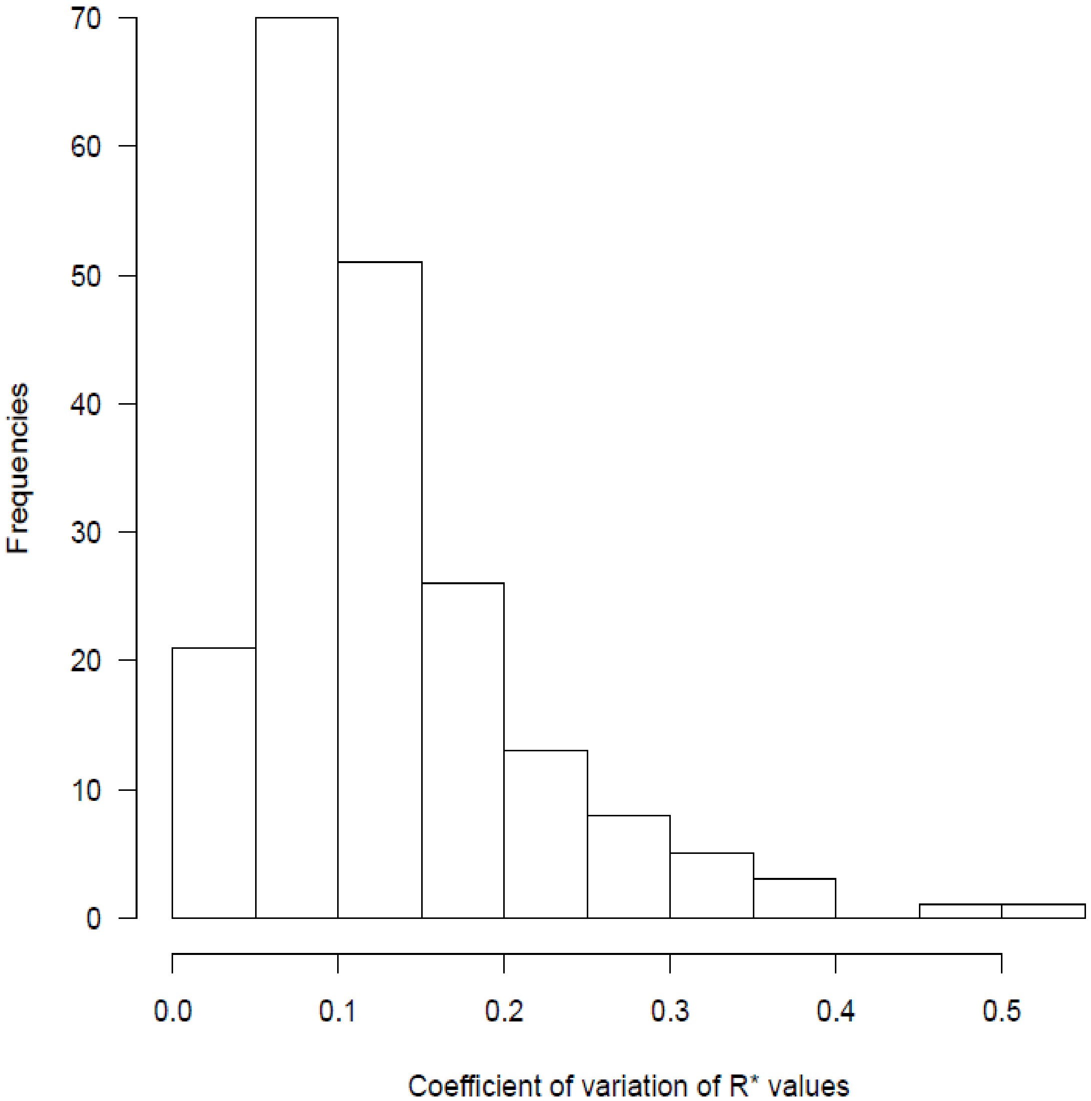


Figure A.9: Histogram of the values of the coefficient of variation of the average rank R^* for the subset of 199 species

A.3 Species prevalence

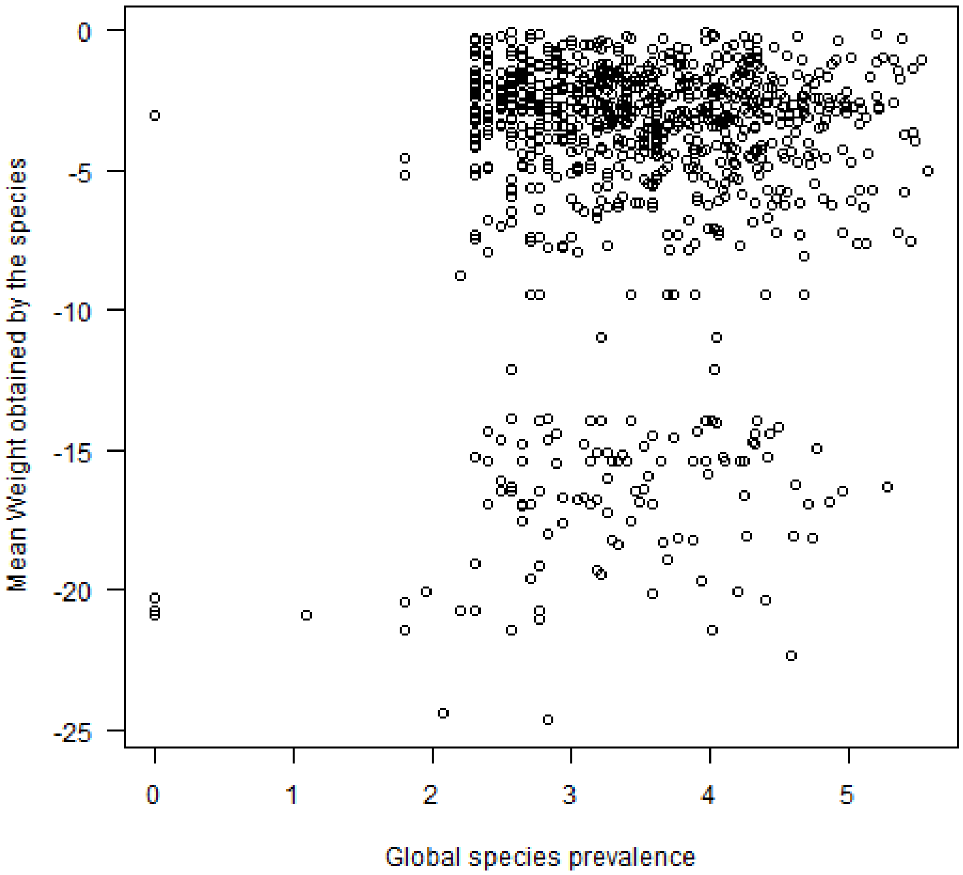


Figure A.10: Species prevalence plotted against species mean weight values. (log-log)

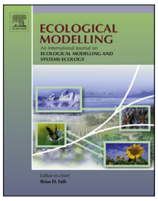
A.4 Related publication

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/

Appendix B

Supplement chapter 3

B.1 Species area relationship

We modelled the global occurence matrix species area relationship (SAR) fitting it into a Gleason log linear model of the form

$$S = -14 + 9.98 * log(area) \tag{B.1}$$

where S is number of species and k is a fitted constant telling the number of species per area unit (Figure B.1). (Dengler, 2009; Gleason, 1922)

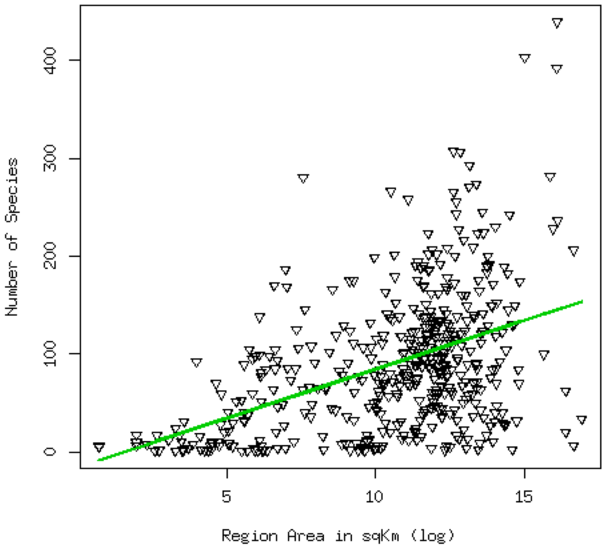


Figure B.1: Species-Area relationship fitted into a log-linear Gleason model. ($k = -14.86, Pr(> |t| = 0.15; slope = 9.98, Pr(> |t| = 2e - 16)$)

B.2 Minimal working example of the computation of ζ

Each neuron of the SOM output map was considered one individual study. Let *neuronA* be one sample neuron of the SOM output map, where region 1, region 2 and region 3 were allocated by SOM PPA. The regional pest profiles for these sample regions were composed of 5 sample species (species 1 to 5). The computation of ζ for *neuronA* is as follows:

```
install.packages('zetadiv')
library(zetadiv)

occurrencemat_neuron_A <- matrix(c(0,1,0,1,1,  0,1,0,1,1,1,1,0,1,0), nrow = 3, nc

datfm_neuron_A <- as.data.frame(occurrencemat_neuron_A)


zetas_Neuron_A <- Zeta.decline(datfm_neuron_A, orders = 1:3)
```

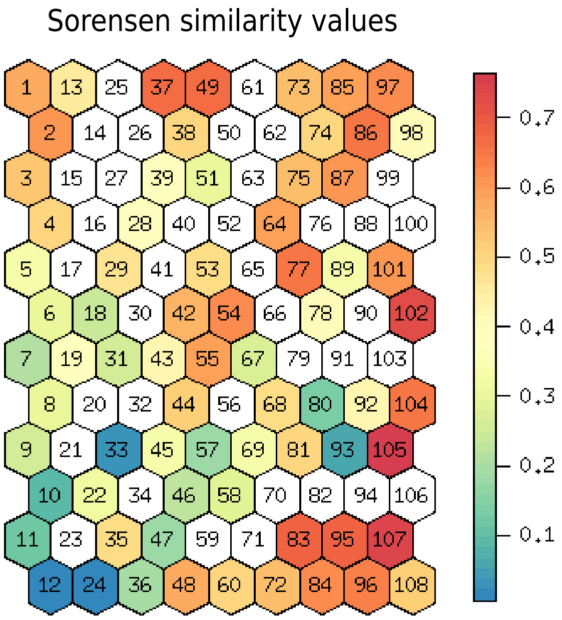


Figure B.2: Sorensen values per neuron of the SOM output map

B.3 Richness

Highest α values	
Region	Richness (α)
USA	439
India	403
China	392
Italy	307
Japan	306
Thailand	293
Australia	282
Indonesia	280
France	274
Spain	271

Table B.1: Regions with highest species richness (α)

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.

Appendix C

Supplement chapter 4

C.1 Computing ζ

```
install.packages('zetadiv')
library(zetadiv)

d2003 = read.table("data2006.txt",header=TRUE)
d2014 = read.table("data2014.txt",header=TRUE)


zetas2003 <- Zeta.decline(d2003)
zetas2014 <- Zeta.decline(d2014)
```

C.2 Significance test for ζ values

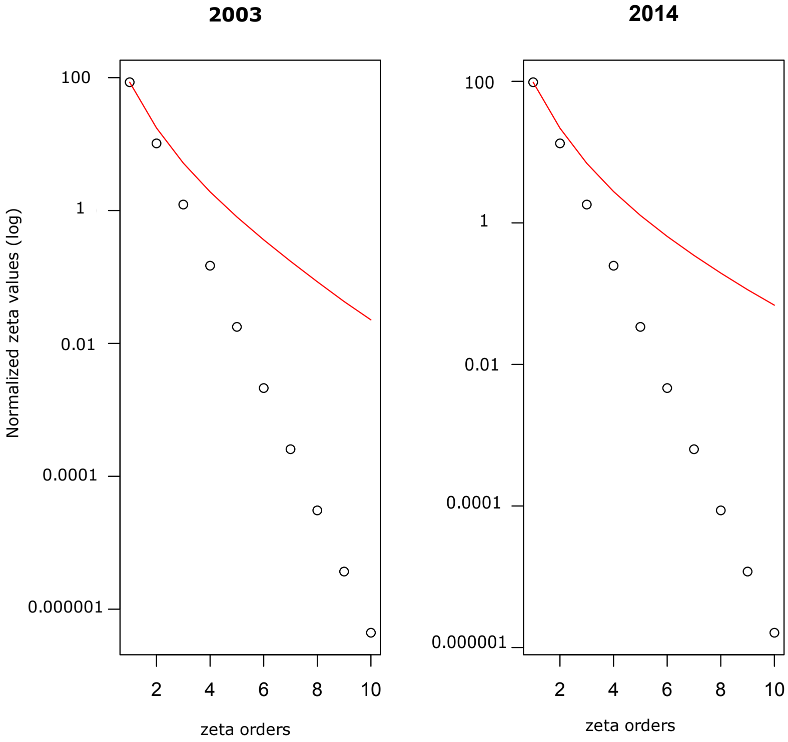


Figure C.1: ζ values computed from randomly generated assemblages (dots) and ζ values computed for the observed data for each year (red lines). Y axes are in logarithmic scale.

C.3 Generating random zeta values for 2014 and 2003 matrices

```
d = read.table("data2003.txt",header=TRUE)
d2 = read.table("data2014.txt",header=TRUE)
nA = nrow(d) # number of sites
M = ncol(d) # number of species
kmax = 10 # max order

theta = sum(d)/nA/M # average presence rate in 2014
theta2 = sum(d2)/nA/M # average presence rate in 2006

zeta_ran = M*theta^(1:kmax) # random zetas 2014
zeta2_ran = M*theta2^(1:kmax) # random zetas 2006

# Plot of zetas against random zetas

par(mfrow=c(1,2))
plot(zeta2_ran,log="y",main="2006",ylab="zeta")
lines(zeta2,col="red")
plot(zeta_ran,log="y",main="2014",ylab="zeta")
lines(zeta,col="red")
```

C.4 Alphas

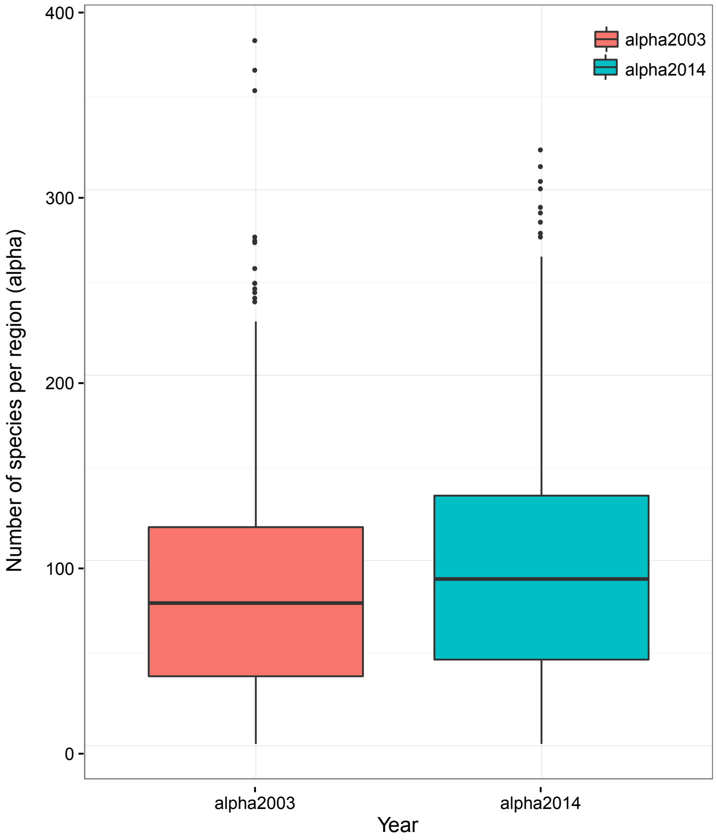


Figure C.2: Values of α diversity in 2003 and in 2014

T