



Master II Recherche

Lucie LUSIGNAN

Biostatistiques et modelisation

June 2009

## **Modelling fungal and oomycete assemblages at the global scale to assess risk of invasion**



### Supervisors:

Corinne VACHER

Muriel GEVREY

Equipe : Pathologie Forestière

Equipe: Structure des communautés et macroécologie

INRA Bordeaux

Laboratoire EDB, UMR 5174

UMR 1202 BioGeCo

CNRS-Université Paul Sabatier, Toulouse

## Table of Contents

Abstract	2
I) Introduction	3
II) Materials and methods	5
a. Data	5
b. SOM model	6
c. Size of the map	8
d. Cluster analysis	9
e. Component plan	9
f. Connection weights	10
g. Interpretation of SOM output in our case study	10
h. Assessment of the predictive value of the model	10
III) Results	11
a. Size of the SOM map	11
b. Optimal number of clusters	11
c. Distribution of the geographical units in the clusters	13
d. Assessment of the predictive value	15
e. Risks of invasion associated with species recorded as absent	17
IV) Discussion	22
References	26
Appendix	29

## Abstract

Predicting biological invasions is important for anticipating the arrival of novel species. It can help in the development of strategies in the aim to prevent their entry, establishment and spread and to limit their detrimental impacts. Our goal here is to assess risks of invasion for pathogens of crop plants and forest trees, at the global scale. Our analyses are based on presence/absence data for 426 fungal and oomycete species in 259 geographical units. We used a particular type of Artificial Neural Network called Self Organizing Map (SOM) to analyze this complex ecological dataset. SOM was used to group the geographical units according to the similarity in their species assemblages. Then we assumed that geographical units presenting similar species assemblages have similar abiotic and biotic conditions that allow establishment of viable populations of these species. Based on this assumption, risks of invasions were associated to each species in each geographical unit. First, risk values for all species in one geographical unit have been extracted. We focused more in details on Guyana case study. Secondly, risk values for one species in all geographical units were extracted. This was illustrated with *Hemileia vastatrix*, a pathogen of coffee crops. Finally, the efficiency of the model was assessed by investigating whether the species that have been recently recorded in a new geographical unit were associated with a high risk of invasion. This approach can then be used to support additional risk assessment of potential invasive species. It could constitute a management tool and could be a mean for prioritizing species to improve activity of research and conservation, and support quarantine and biosecurity scientists and authorities.

**Keywords:** Invasive species, Self Organizing Map, Fungi, Oomycete, Risk of invasion, Assemblages.

## I) Introduction

Biological invasions are defined as the successful establishment and spread of species outside their native range. To characterize what is common to all invasions, Valery *et al.* (2008) proposed a mechanistic definition: “A biological invasion consists of a species acquiring a competitive advantage following the disappearance of natural obstacles to its proliferation, which allows it to spread rapidly and to conquer novel areas within recipient ecosystems in which it becomes a dominant population”.

Biological invasions are more and more frequent because of the intensification of transport and trade. They often have detrimental ecological and economic consequences. Costs of invasive species were estimated to range from millions to billions of dollars annually (U.S. Congr. Off. Technol. Assess. 1993; Pimentel *et al.* 2000). For example, an estimated 50 000 non-indigenous species established in the United States cause major environmental damage and economic losses that total over an estimated U.S.\$125 billion per years (Pimentel *et al.* 2001). Therefore, biological invasions are considered as a significant component of global environmental change (Vitousek *et al.* 1996).

Some researchers consider that a species must have a major impact on the recipient community/ecosystem in order to be considered invasive (Davis and Thompson 2000; Davis and Thompson 2002; Inderjit 2005). However, this subject is a matter of debate (Valery *et al.* 2008). Other authors consider that the impact of a species on the ecosystem is too difficult to estimate for being taken into account in the definition of a biological invasion (Richardson *et al.* 2000; Daehler 2001; Rejmanek *et al.* 2002). More fundamental reason for them to reject the impact criterion: the characterization of the phenomenon itself should rest solely on criteria relative to its “substance” and not to its effects or consequences, which are circumstantial and variable.

To be successful invaders, species have to pass through a series of filters corresponding to the successive steps of an invasion: transport, establishment, spread and integration (Desprez-Loustau *et al.* 2007).

The transport of species from its native area to a novel area may occur as the result of climatic and tectonic changes as well as through introduction by humans (Vermeij 1996). For cases of invasion unaided by humans, potential means of transport include ocean currents, winds, hitchhiking on the bodies of other organisms, and active locomotion (like walking, swimming or flying) across barriers that have disappeared or are no longer effective. Human-

aided invasion is made possible through the accidental or deliberate transport of crops, game animals, ballast water, soil, agricultural control agents, wool and living food organisms. And by the construction of floating or permanent islands such as ships, planes, buoys, and oil derricks, which can serve as homes or stepping stones for recruits (Vermeij 1996).

After transport, the introduced individuals may form a viable, self-sustaining population in the recipient region. This second stage of the invasion process is called establishment. For being considered as established, an immigrant population should persist by means of local reproduction and recruitment (Vermeij 1996). It is noteworthy that some species may invade so frequently that they appear to maintain local populations when in fact the populations are continually being recharged by dispersers from elsewhere.

Once transport and establishment have occurred, invasive species may spread from continuing long-distance dispersal from foreign sources, and from short-distance dispersal with lateral expansion of the established population (Smith *et al.* 1999; Davis and Thompson 2000). The multiple transatlantic introductions of the western corn rootworm caused by *Diabrotica virgifera virgifera* is an example of this process (Miller *et al.* 2005).

Finally, the immigrant species is considered as integrated in the recipient region when the species in the recipient biota and the invader respond each other ecologically and evolutionary. Integration occurs when, as the invading species forges ecological links with others species in the recipient region, evolution occurs, reflecting the changed selective regime in the recipient community (Vermeij 1996).

Fungi in a broad sense, are among the most ecologically and evolutionary diverse organisms on Earth and are classified within three eukaryotic kingdoms: Fungi, Stramenopila and Protista (Desprez-Loustau *et al.* 2007). Fungi represent an essential component of biodiversity because of the large number of species. Also, they are of considerable economic and social importance through beneficial uses and as causes of plant and animal diseases (Pimentel *et al.* 2001). Fungi have only recently begun to receive consideration in ecology, especially invasion ecology (Desprez-Loustau *et al.* 2007).

The most obvious impacts of fungal invasions are epidemics caused by exotic pathogenic fungi. Non-indigenous fungal pathogens are often highly destructive to species that are closely related to the pathogens' host species. In the native range of the fungus, the pathogens' host species are relatively unaffected, probably due to the host-pathogen co-adaptation (Facon *et al.* 2006). It was been estimated that 65-85% of plants pathogens worldwide are alien in the location where they were recorded (Desprez-Loustau *et al.* 2007). Better known examples include epidemics of potatoes blight, Chestnut blight and Dutch elm

disease. Potatoes blight is caused by three oomycete species pathogens and can cause great crop-production losses (Platt 2008). In the 1840s, potatoes blight caused the death of one million people in Ireland and the emigration of one million more. Dutch elm disease is caused by two fungal species and has resulted in the death of billions of elm trees on two continents (Europe and North America). It is considered as one of the major ecological accidents of the 20<sup>th</sup> century (Heybroek 1993). The chestnut blight fungus, *Cryphonectria parasitica*, virtually eliminated the once-widespread American chestnut tree. In North America, many other native tree species are currently in the process of being removed by introduced pests and pathogens (Orwig 2002). Vitousek *et al.* (1996) claimed that the greatest perturbations to forest ecosystems in the twentieth century resulted from waves of introduced pests and pathogens, and they suggested that nonnative species will continue to be the greatest threat to diversity of eastern North American hardwood forests in the future.

The aim of the present study is to predict risks of invasion for several hundreds of fungal and oomycete species which are pathogens of crop plants and forest trees. Such predictions will allow agronomists and scientists to anticipate the arrival of novel pathogens, by finding a strategy to prevent the invasion or limit its impact. Our predictions are based on a data set including fungal and oomycete species occurrence at the global scale. We used Self Organizing Maps (SOM) model to group together geographical units having similar fungal and oomycetes species assemblages. Our hypothesis was that geographical units with similar fungal and oomycete assemblages share similar biotic and abiotic conditions that allow them to invade the area. Biotic conditions can include crops, forest tree species and garden plants reflecting abiotic one like climate that affect species diversity and abundance (Worner and Gevrey 2006). Then, risks of invasions associated with each species and each geographical area were calculated, by assuming that geographical units presenting similar species assemblages share equivalent conditions that allow establishment of viable population of these species (Gevrey *et al.* 2006).

## II) Materials and methods

### a. Data

Presence/absence data for phytopathogenic fungi and oomycetes were extracted from the Crop Protection Compendium (CABI 2006, 2007). The Crop Protection Compendium is an interactive multimedia encyclopedia with a comprehensive worldwide coverage of pests of

agricultural and horticultural crops (and since 2004, of forest trees). All continents are represented except Arctic and Antarctic regions. Presence/absence data for phytopathogenic fungi were automatically extracted from the 2006 version of the Crop Protection Compendium. Then the data were manually updated with the last available version of the Crop Protection Compendium (2007) and presence/absence data for oomycetes were included. Only the species with detailed information for geographical distribution (i.e. with a full or an outline datasheet) were included in the database. In the Crop Protection Compendium, most of the geographical units correspond to political countries. Large countries are subdivided into provinces or states. Many islands which are far from the country to which they are attached (for example Sardinia) are also considered as separate geographical units. To obtain geographical units which are least dependant of political considerations, we grouped together several geographical units according to the Level-3 world map (<http://www.kew.org/gis/tdwg/>) elaborated by the *International Working Group on Taxonomic Databases* (TDWG). Although large countries are partitioned in the Level-3 map, we chose to keep the presence/absence data related to the large countries - in addition to the data related to their different subunits - because some species have been recorded as present in a large country but recorded as absent in all the states or provinces composing this country. Therefore, we kept the presence/absence data for seven large countries: USA, Canada, Brazil, India, Indonesia, Russia and China. Finally in order to use only robust data, we eliminated the species occurring in less than 2% of the geographical units and then we eliminated the geographical units containing less than 2% of the total number of species (Waite, 2000). The final database gave the distribution of 426 species (369 fungal species and 57 oomycetes species) in 259 geographical units.

## b. SOM model

We used a Self Organizing Map (SOM) to analyze this high dimensional dataset. A SOM is a type of Artificial Neural Networks (ANN) which uses unsupervised learning to produce a low dimensional representation of the input space (Kohonen 1982). ANN are a large class of parallel processing architectures, which can mimic complex and non-linear relationships through the application of many nonlinear processing units called neurons (Lin *et al.* 2009). Usually, the structure of an ANN is hierarchical with neurons grouped in different layer.

The SOM network consists of two layers of neurons: the input and the output layers (Fig. 1). The dataset presented to the network by its input layer is comprised of samples

characterized by descriptors. The input layer is composed of as many neurons as descriptors in the samples forming a Real Vector  $RV$ . The output layer is often represented as a 2D map in a form of rectangular grid with hexagonal lattice composed of neurons also called cells. Each neuron  $j$  of the output layer is connected to each neuron  $i = 1, 2, \dots, p$  of the input layer by a weight. These weights represent the virtual values for each descriptor in each output neurons. Hence, each cell contains a Virtual Vector  $VV_j$ . These output vectors have the same dimension than the input vectors. In order to organize the map, the SOM model compares each  $RV$  to all  $VV$  in term of distance. The distance measure is chosen according to the case study. Each  $RV$  is then assigned to the nearest  $VV$ . This latter is called the best matching unit (BMU).

The algorithm can be summarized as follow in the case of the Euclidean distance:

- Virtual Vectors ( $VV_j$ ) are randomly initialized.
- Real Vectors ( $RV$ ) are sequentially presented to the algorithm:
  - A  $RV_i$  is chosen as an input vector
  - The Euclidean distance between the  $RV$  and each  $VV$  is computed.

$$D(RV_i, VV_j) = \left[ \sum_{i=1 \dots p} (RV_i - VV_j)^2 \right]^{1/2}$$

- The nearest  $VV$  is selected (the Best Matching Unit, BMU)
- Values of the BMU are modified in order to be closer to the value of the  $RV$  that has been assigned to it. Neurons close to the BMU are adjusted towards the  $RV$  using this rule:

$$W_j(t+1) = VV_j(t) + \eta(t) \cdot N(t, r) (RV_k(t) - VV_j(t))$$

Where  $t$  is the number of iterations, and  $RV_i(t)$  is a real vector.  $VV_j(t)$  is a virtual vector that represents the weights between a neuron  $j$  of the output layer and all the neurons of the input one at iteration  $t$ .  $\eta(t)$  is the learning rate that is a decreasing function of iteration  $t$  and  $N(t, r)$  is the neighborhood function with  $r$  representing the distance in the map between the BMU and its neighboring neurons. This function defines the size of the neighborhood of the

BMU to be updated during the learning process. The magnitude of the change decreases with distance from the BMU. This process is repeated for each  $RV$ .

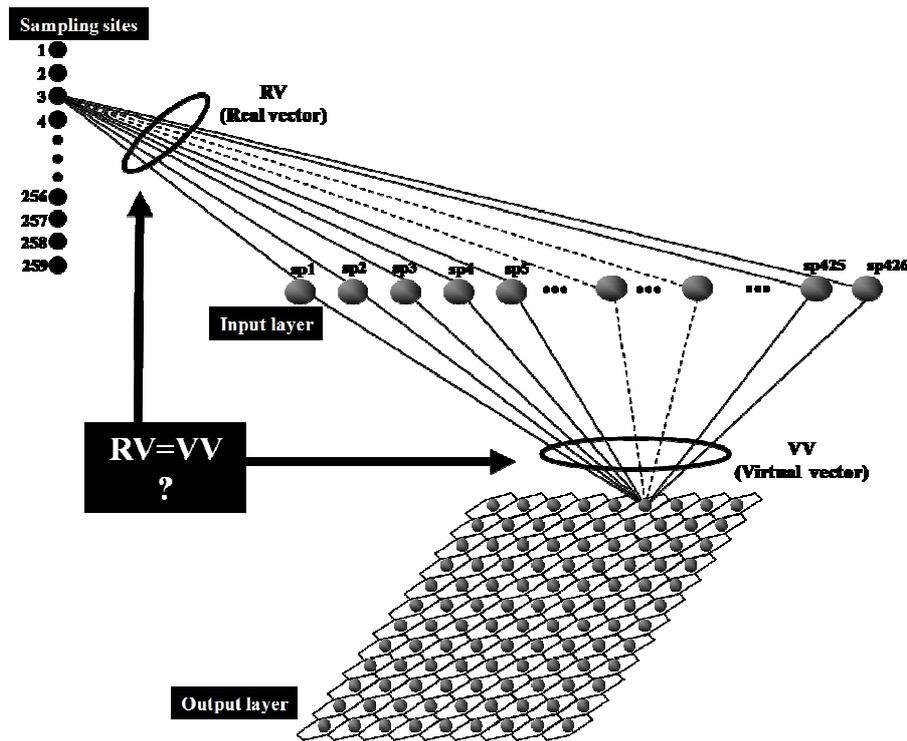


Figure 1: Self-Organizing Map architecture. The input layer is linked to the cells of the output layer by connection called weights that define the virtual vector.

At the end of the training, each sample is assigned to a cell of the map. The virtual values of the descriptors are known for each cell of the map. Therefore, in the final map, real vectors which are similar are mapped close together and dissimilar apart.

### c. Size of the map

The size of the output layer (i.e. the SOM map) has to be defined by the user. According to the Laboratory of Computer and Information Science (CIS) of the Helsinki University of Technology (Espoo, Finland), the number of output neurons should approximately be equal to  $5\sqrt{n}$  where  $n$  is the number of real vectors. There is no theoretical principle to determine the optimum maps size, but quantitative indicators such as quantization and topographic errors have proven to be relevant tools to determine the optimal number of map units (Cereghino and Park 2009). Quantization error is the average distance between each real vector and its BMU, thus it is a measure of map resolution. Topographic error represents the proportion of all real vector for which 1<sup>st</sup> and 2<sup>nd</sup> BMUs are not adjacent, it is a

measure of topology preservation. Quantization and topographic errors, which must be small (Kohonen 2001), are therefore used in this study for adjusting the number of output neurons. We are more confident in the topographic error because it is not dependent on the distance measure used in the algorithm, contrary to the quantization error.

#### d. Cluster analysis

At the end of the training, each cell of the map can be considered as a cluster regrouping similar virtual vectors. However, it is often of interest to define larger clusters by regrouping the neighboring cells that contains similar samples. To define larger clusters, a cluster analysis is then applied to the virtual vectors to determine cluster boundaries on the map. In case of Euclidean distance, a hierarchical cluster analysis is applied with a Ward Linkage method. Results of this analysis can be visualized in the form of a dendrogram.

In order to choose the correct number of clusters, the function *labelstomss* from the *hopach* package of R (Pollard *et al.*, 2009) was used. This function computes silhouettes which measure how well an element belongs to its cluster. The Median Split Silhouette (MSS) is a measure of cluster heterogeneity. Given a partitioning of elements into groups, the MSS algorithm considers each group separately and computes the split silhouette for that group, which evaluates evidence in favor or further splitting the group. If the median split silhouette over all groups in the partition is low, the groups are homogeneous. Then results are compared and the number of clusters for which the median split silhouette is the lower is selected.

#### e. Component plan

We can visualize the contribution of each input variable in each cluster. Each cell of the map is represented by a virtual vector composed by as many weight values as descriptors. It is possible to visualize each descriptor's weight values in each cell of the map. It is represented in color scale and one map is available for each descriptor. Compare to the SOM map, it is possible then to evaluate to which cluster the descriptor contributes the most.

#### f. Connection weights

Neurons of the two layers of the SOM network are linked by connections associated with weights. The virtual vector of each output neuron is composed by a set of weight values (same number as descriptors). Those weights are the coordinates of each output neuron in a multidimensional space.

#### g. Interpretation of SOM output in our case study

In this study, the real values are data of species presence-absence, therefore equal to 0 or 1 (binary data). The virtual values are then comprised between 0 and 1. Samples (i.e. geographical units) that are assigned to a same cell of the map have similar real values and share the same virtual values. The virtual value for one descriptor in a cell can be seen as a “summary” of the real values of all samples assigned to it. The weights of the virtual vector for a given cell have therefore been interpreted as the “probability” of occurrence of the species in geographical units associated with this cell (Gevrey *et al.* 2006).

#### h. Assessment of the predictive value of the model

In order to assess the predictive value of the model in terms of risk of invasion, we established the list of fungal species that invaded a new geographical area after the end of construction of our database. Based on the pest alerts of the NAPPO Phytosanitary Alert System ([http://archives.eppo.org/EPPOReporting/Reporting\\_Archives.htm](http://archives.eppo.org/EPPOReporting/Reporting_Archives.htm)) and the archives of the EPPO reporting service ([http://www.pestalert.org/opr\\_search.cfm](http://www.pestalert.org/opr_search.cfm)), we found 16 cases of novel introductions corresponding to 9 fungal species and 15 geographical units. Places of detection of these novel invasive species were mainly in Europe and in North America. However, Asia and Africa were also represented, with one pest alert each. For each geographical unit, we compared the risks associated with the newly introduced species to those associated with the species recorded as absent (and still absent). We also investigated whether or not the newly introduced species belonged to the top species in terms of risk of invasion, by ranking the absent species according to their associated risks. Finally, we compared the risks and ranks of the fungal species belong to the EPPO A1 List (<http://www.eppo.org/QUARANTINE/listA1.htm>) to those of the other species recorded as absent. Species of the A1 list are species absent from the EPPO region but considered by experts as having a high risk of invasion. We added the soyabean rust, *Phakopsora*

*pachyrhizi*, to the list of A1 species because it has not been detected yet in the EPPO region but is considered at risk in Southern Mediterranean countries ([http://www.eppo.org/QUARANTINE/Alert\\_List/alert\\_list.htm](http://www.eppo.org/QUARANTINE/Alert_List/alert_list.htm)).

### III) Results

#### a. Size of the SOM map

In our study, the input layer comprised 426 neurons (one for each species) connected to the 259 sites (geographical units). So the presence-absence of the 426 species in each site is represented by 259 real vectors.

The optimum cell number given by the equation  $5\sqrt{n}$  is  $c=102.3054$ . The map used in this analysis is composed by 104 neurons organized on an array with 13 columns and 8 rows. Topographic and quantization errors are respectively 0 and 6.009.

#### b. Optimal number of clusters

The smallest value (0.435) of MSS was obtained for seven clusters (Fig. 2). Thereby the 259 geographical units assigned to the map cells were divided into 7 clusters (Fig. 3). The names of the geographical units and the corresponding clusters are given in Table 1.

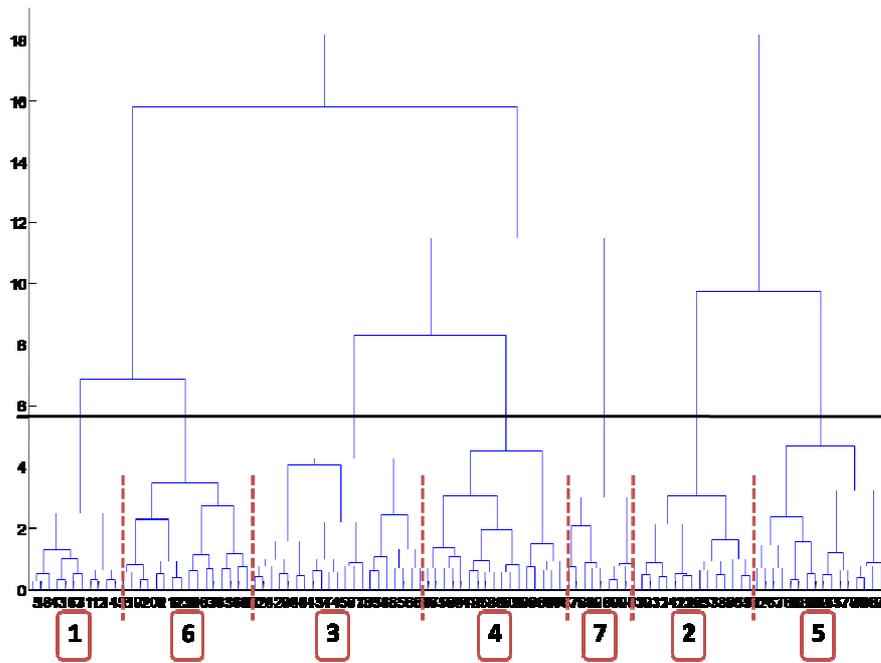


Figure 2: Dendrogram of the hierarchical classification analysis



Figure 3: Self organizing map with numbers corresponding to the clusters defined by the cluster analysis (see dendrogram figure 2). The three-letter codes associated with the cells correspond to the names of the geographical units (see appendix table for name details).

### c. Distribution of the geographical units in the clusters

Then the seven clusters have been plotted on a geographical map of the world using Geographical Information System (GIS) (Fig. 4).

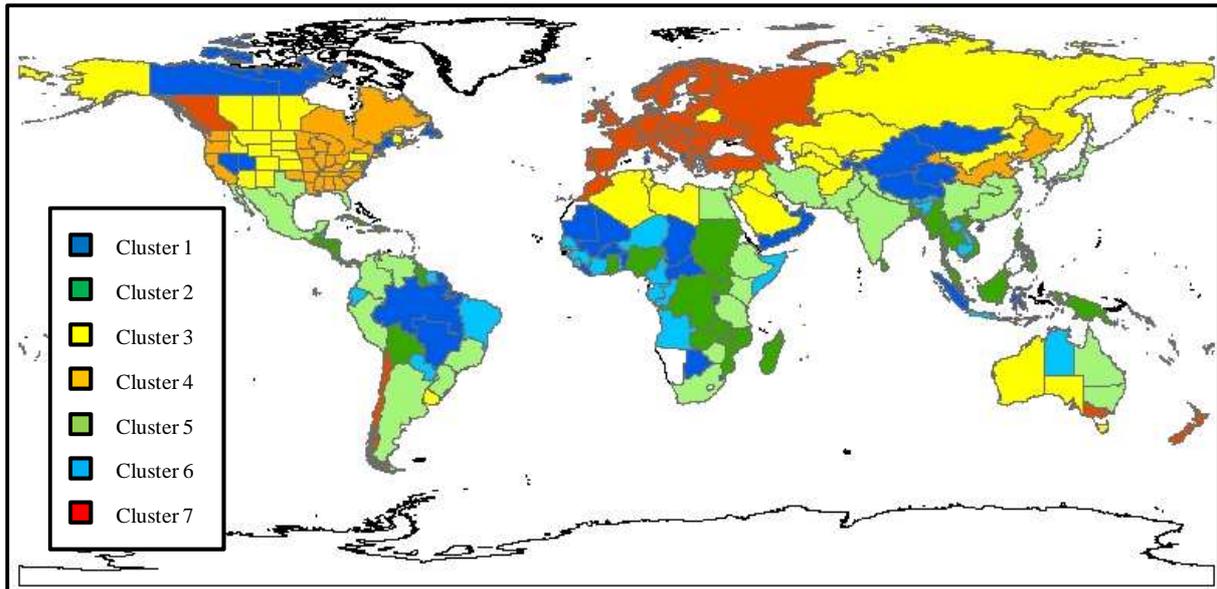


Figure 4: World map showing the 259 geographical units represented by different colors according to the cluster to which they belong.

Cluster 1 represents the geographical units that have a low number of species recorded in the database. Among the 55 geographical units included in this cluster, 47 are geographical units in which there is the least recorded species. Moreover, the average number of species detected in a geographical unit for the entire dataset is 90 against 22.9 in the cluster 1. These geographical units are desert area, colder areas or areas that have probably not been sampled properly (part of Brazil for example). In the same manner, cluster 6 contains geographical units that have low species occurrence values with a mean of 51.34 species per geographical unit. This cluster is constituted by West African countries (Angola, Niger, Cameroon, Gabon...), few South America countries (east part of Brazil, Ecuador, Paraguay and Surinam), few Asian countries (Cambodia, Laos and east part of India) and North Australia. Middle East and West coast of United States and North China are included in the cluster 3. Cluster 4 comprises North Asian countries (as Russia, Kazakhstan and North of China...), Alaska, Middle West of Australia, centre of United States and North African countries (Algeria, Libya, Saudi Arabia, Iraq and Syria). These two clusters are close together in the dendrogram, and near cluster 7. Cluster 7 comprises a majority of European countries (except

Belarus), British Columbia, Chile, New Zealand and a part of Southeast Australia. Large countries as Canada and Russia are also included in this cluster.

Majority of middle and South American countries (Mexico, Argentina, Peru, Uruguay, Columbia and Venezuela), East African countries (Soudan, Congo, Tanzania...), South Asian countries (Iran, India, south China, Borneo...) and East Australia are regrouped in cluster 2 and 5, that are closed in the dendrogram. Large countries as United States, China, Brazil and India are included in the cluster 5.

Table 1a: Names of geographical units associated to their corresponding cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Albania	Bangladesh	Alberta	Alabama	Argentina	Angola	Austria
Andaman Is.	Bolivia	Afghanistan	Arkansas	USA	Assam	Belgium
Azores	Borneo	Algeria	California	Brazil Southeast	Belize	Baltic States
Benin	Costa Rica	Altay	Manchuria	Brazil South	Brazil Northeast	British Columbia
Bermuda	Cuba	Amur	China North-Central	China South-Central	Cambodia	Bulgaria
Burkina	Dominican Republic	Arizona	Delaware	China Southeast	Cameroon	Canada
Botswana	El Salvador	Alaska	Georgia	Colombia	Congo	Chile Central
Burundi	Fiji	Belarus	Illinois	China	Ecuador	Czechoslovakia
Brazil West-Central	Ghana	Inner Mongolia	Indiana	Cape Provinces	East Himalaya	Denmark
Brazil North	Guatemala	Connecticut	Iowa	Egypt	Gabon	Finland
Central African Republic	Guyana	Colorado	Kentucky	Ethiopia	Guinea	France
Chad	Haiti	Cyprus	Louisiana	Florida	Ivory Coast	Germany
Hainan	Hawaii	Idaho	Michigan	Brazil	Jawa	Great Britain
Qinghai	Honduras	Iraq	Minnesota	India	Laos	Greece
Tibet	Indonesia	Kansas	Maryland	Iran	Leeward Is. AB Ant	Hungary
Xinjiang	Jamaica	Kazakhstan	Mississippi	Japan	Marshall Is.	Ireland
Canary Is.	Mauritius	Kirgizistan	Missouri	Kenya	Niger	Italy
Cook Is.	Madagascar	Lebanon-Syria	North Carolina	Korea	Northern Territory	Morocco
Corse	Malawi	Libya	Nova Scotia	Mexico Central	Paraguay	Netherlands
Caroline Is.	Malaya	Manitoba	New Jersey	Nepal	Samoa	Norway
Føroyar	Mozambique	Massachusetts	New York	New South Wales	Senegal	New Zealand
French Guiana	Myanmar	Montana	Ohio	Pakistan	Sierra Leone	Poland
Gambia, The	Nigeria	New Brunswick	Oklahoma	Palestine	Solomon Is.	Portugal
Gulf of Guinea Is.	Nicaragua	North Dakota	Ontario	Peru	Somalia	Romania
Gulf States	New Caledonia	Nebraska	Oregon	Queensland	Surinam	Russia
Iceland	New Guinea	New Mexico	Québec	Taiwan	Togo	Central European Russia

Table 1b: Names of geographical units associated to their corresponding cluster (continued).

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Kriti	Panama	Prince Edward I.	South Carolina	Tanzania	Tonga	Spain
Liberia	Philippines	Pennsylvania	Tennessee	Texas	French Polynesia	Sweden
Lesser Sunda Is.	Puerto Rico	Saskatchewan	Virginia	India	Vanuatu	Switzerland
Maine	Sri Lanka	Saudi Arabia	Washington	Venezuela		Turkey
Madeira	Sudan	South Dakota	Wisconsin	West Himalaya		Ukraine
Mali	Thailand	Sicilia		Zimbabwe		Victoria
Mongolia	Trinidad-Tobago	South Australia				Yugoslavia
Mauritania	Uganda	Tasmania				
Nevada	Vietnam	Turkmenistan				
Newfoundland	Windward Is.	Tunisia				
Nansei-shoto	Zaire	Uruguay				
Niue	Zambia	Uzbekistan				
New Hampshire		Western Australia				
Northwest Territories		West Virginia				
Oman		Wyoming				
Réunion						
Rhode I.						
Rwanda						
Sardegna						
Seychelles						
Sulawesi						
Sumatera						
Swaziland						
Tadzhikistan						
Utah						
Vermont						
District of Columbia						
Yemen						
Yukon						

#### d. Assessment of the predictive value

In order to evaluate the predictive values of the model, risk and rank for alert and A1 list have been extracted. For Alert species, the mean risk is 0.26 and the mean rank 33.83 (Table 2). Two species present risk superior 0.40 in their place of detection. *Mycosphaerella pini* has risk values (0.47, 0.43, 0.64) in the three geographical units where it was actually detected (Belgium, Finland and Netherlands, respectively). *Puccinia horiana* has a high risk of invasion (0.58) in one the places where it was detected (Turkey).

Table 2: Risk, rank and mean rank of Alert species in their place of detection. The two values in blue correspond to the mean of the corresponding columns, risks and mean rank.

Source	Species name	Disease name	Place of detection	Risks	Ranks	Mean rank
EPPO Reporting Service	Gibberella circinata	pitch canker	Italy	0.13	68 / 176	38.64
EPPO Reporting Service	Guignardia citricarpa	citrus black spot	Uganda	0.21	82 / 304	26.97
EPPO Reporting	Monilinia fructicola	brown rot	Czech Republic	0.24	70 / 267	26.22
NAPPO Official Pest Report	Moniliophthora roreri	moniliasis disease of cocoa crops	Chiapas and Tabasco (Mexico)	0.06	185 / 229	80.79
EPPO Reporting Service	Mycosphaerella dearnessii	brown spot needle blight	Czech Republic	0.3	53 / 267	19.85
EPPO Reporting Service	Mycosphaerella pini	Dothistroma blight	Belgium	0.47	41 / 305	13.44
EPPO Reporting Service	Mycosphaerella pini	Dothistroma blight	Finland	0.43	33 / 321	10.28
EPPO Reporting Service	Mycosphaerella pini	Dothistroma blight	Netherlands	0.64	6 / 258	2.33
EPPO Reporting Service	Phytophthora ramorum	sudden oak death syndrome	Lithuania	0.38	52 / 319	16.30
NAPPO Official Pest Report	Puccinia horiana	white rust of chrysanthemum	Connecticut (USA)	0.04	210 / 352	59.66
NAPPO Official Pest Report	Puccinia horiana	white rust of chrysanthemum	Michigan (USA)	0.06	184 / 333	55.26
NAPPO Official Pest Report	Puccinia horiana	white rust of chrysanthemum	British Columbia (Canada)	0.19	97 / 304	31.91
EPPO Reporting Service	Puccinia horiana	white rust of chrysanthemum	Turkey	0.58	24 / 271	8.86
NAPPO Official Pest Report	Puccinia kuehnii	orange rust of sugarcane	Florida (USA)	0.15	122 / 255	47.84
NAPPO Official Pest Report	Uromyces transversalis	gladiolus rust	California (USA)	0.02	165 / 248	66.53
NAPPO Official Pest Report	Uromyces transversalis	gladiolus rust	Florida (USA)	0.21	93 / 255	36.47
				0.26		33.83

Table 3 shows risks associated to the cell n°93 of the map which included among others France, Italy and Spain. Therefore risk values are the same for these three countries. Lists differ according to the absent species on these each geographical units and their risks associated. Values of risk for absent species vary between 0 and 0.16, which is low. It is noted worthy that all A1 species are ranked in the same order in these three geographical units and that eleven A1 species are ranked in the first quarter of the list (i.e. their mean rank is lower than 0.25), despite their low risk values. It is also note worthy that the soyabean rust, *Phakopsora pachyrhizi*, has the highest high risk of invasion among all absent species. This result confirms the predictions of EPPO experts.

Table 3: Risk, rank and mean rank of France, Italy and Spain for species of A1 list.

A1 species	Risk	France		Italy		Spain	
		186 absent species		177 absent species		255 absent species	
		Rank	mean Rank	Rank	mean Rank	Rank	mean Rank
<i>Phakopsora pachyrhizi</i>	0.16	2	1.08	2	1.13	4	1.57
<i>Cronartium comandrae</i>	0.10	10	5.38	10	5.65	17	6.67
<i>Cronartium coleosporioides</i>	0.09	13	6.99	12	6.78	19	7.45
<i>Endocronartium harknessii</i>	0.09	14	7.53	13	7.34	20	7.84
<i>Chrysomyxa arctostaphyli</i>	0.09	18	9.68	15	8.47	23	9.02
<i>Apiosporina morbosa</i>	0.07	29	15.59	27	15.25	37	14.51
<i>Gymnosporangium clavipes</i>	0.07	30	16.13	28	15.82	39	15.29
<i>Puccinia hemerocallidis</i>	0.07	34	18.28	32	18.08	45	17.65
<i>Gymnosporangium juniperi-virginianae</i>	0.06	38	20.43	39	22.03	52	20.39
<i>Alternaria longipes</i>	0.06	42	22.58	specie present in data		specie present in data	
<i>Cronartium comptoniae</i>	0.05	48	25.81	46	25.99	62	24.31
<i>Cronartium fusiforme</i>	0.05	56	30.11	52	29.38	72	28.24
<i>Atropellis piniphila</i>	0.04	58	31.18	55	31.07	76	29.80
<i>Physoderma maydis</i>	0.03	70	37.63	65	36.72	95	37.25
<i>Gibberella circinata</i>	0.02	82	44.09	77	43.50	specie present in data	
<i>Sirococcus clavignenti-juglandacearum</i>	0.02	90	48.39	81	45.76	115	45.10
<i>Phymatotrichopsis omnivora</i>	0.02	91	48.92	82	46.33	116	45.49
<i>Ophiostoma wageneri</i>	0.02	94	50.54	86	48.59	122	47.84
<i>Phellinus weirii</i>	0.02	100	53.76	90	50.85	128	50.20
<i>Cronartium quercuum</i>	0.02	101	54.30	91	51.41	130	50.98
<i>Mycosphaerella gibsonii</i>	0.01	111	59.68	100	56.50	144	56.47
<i>Stegophora ulmea</i>	0.01	118	63.44	106	59.89	154	60.39
<i>Phyllosticta solitaria</i>	0.01	132	70.97	120	67.80	176	69.02
<i>Gymnosporangium globosum</i>	0.01	133	71.51	121	68.36	177	69.41
<i>Melampsora farlowii</i>	0.01	134	72.04	122	68.93	179	70.20
<i>Anisogramma anomala</i>	0.01	135	72.58	123	69.49	180	70.59
<i>Diaporthe vaccinii</i>	0.01	136	73.12	124	70.06	182	71.37
<i>Puccinia pittieriana</i>	0.00	139	74.73	127	71.75	188	73.73
<i>Tilletia indica</i>	0.00	147	79.03	139	78.53	205	80.39
<i>Guignardia citricarpa</i>	0.00	155	83.33	146	82.49	217	85.10
<i>Thecaphora solani</i>	0.00	157	84.41	148	83.62	219	85.88
<i>Atropellis pinicola</i>	0.00	164	88.17	155	87.57	228	89.41
<i>Gymnosporangium yamadae</i>	0.00	177	95.16	170	96.05	245	96.08

e. Risks of invasion associated with species recorded as absent

Each descriptor's weight values can be visualized in each cell of the map, created then an index of risk of invasion of species visible in the form of a color scale on the map. As an example, *Hemileia vastatrix* was chosen (Fig. 5). Thus we can note that this species show high values of risk for cells situated on the bottom left hand side of the map. According to the figure 3, these cells are corresponding to the cluster 2. Indeed geographical units included in this cluster are in majority southern hemisphere countries.

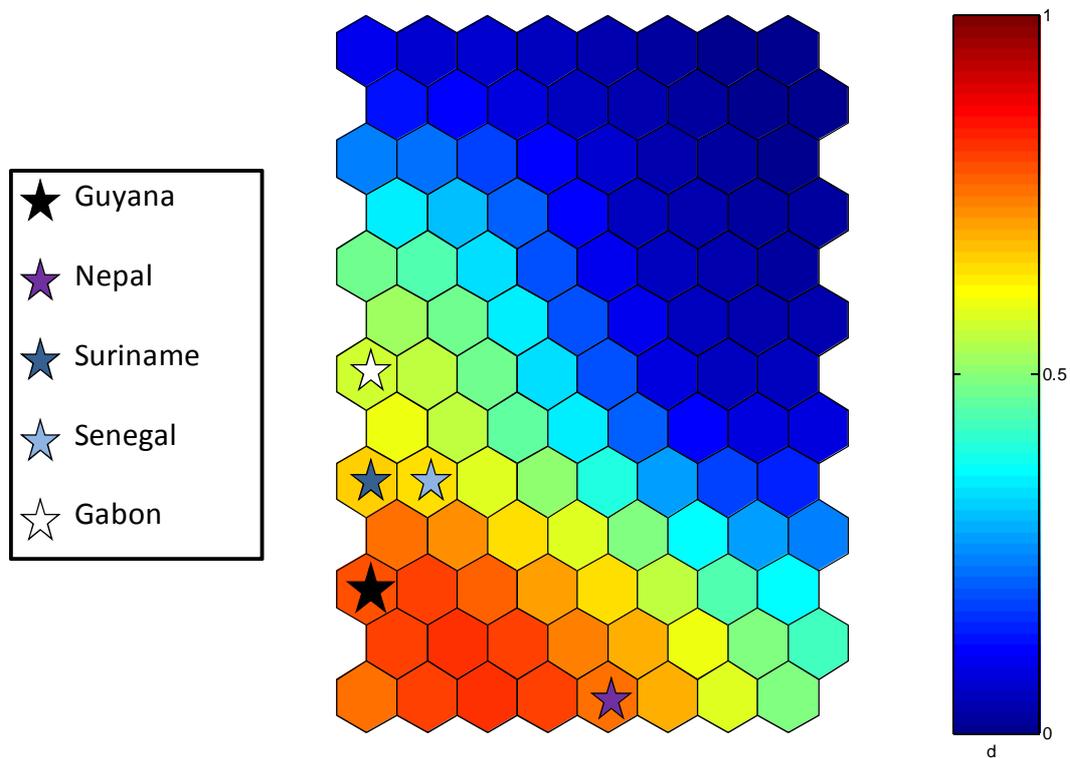


Figure 5: SOM map for *Hemileia vastatrix* and scale of risk values; stars designed cells in which selected geographical units are assigned (Guyana, Nepal, Suriname, Senegal and Gabon).

In the aim to verify and understand values of risk of invasion for *Hemileia vastatrix*, we displayed the risk values for this species on a GIS world map (Fig. 6b). Then we compared the map with the world map of coffee producing countries (Fig. 6a). Geographical units for which *Hemileia vastatrix* has a high risk value are coffee-producing countries except for North and Northeast Australia and Niger. Risk values are important in South America for Guyana (0.80) and Suriname (0.67), in Africa for Gabon (0.57), Senegal (0.65) and Niger (0.57), and in Asia for Nepal (0.75). Cells in which these countries are assigned are represented by stars in the *Hemileia vastatrix* SOM map (Fig. 5).

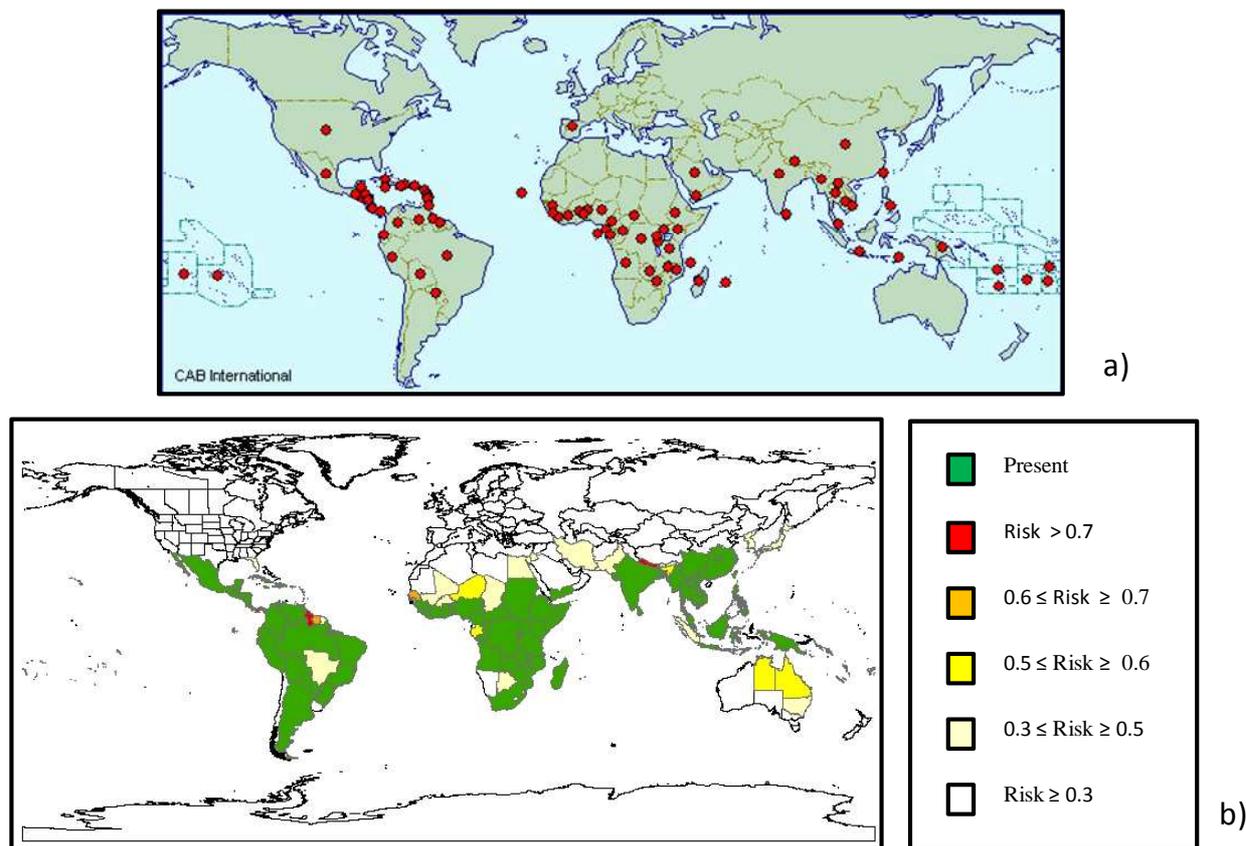


Figure 6: a) World map showing distribution of coffee crop (Cabi, 2007). b) World map showing present distribution of *Hemileia vastatrix* in green and index of risk of invasion for this species.

An alternative way to visualize risk of invasion values is to extract a list of risk values for a given geographical unit. It is particularly interesting when the values of risk for species that are absent in the geographical unit are high.

In this study, we chose to extract the list of species that have highest value of risk in France (Table 4) and Guyana (Table 6). The values for absent species of France are globally low with a maximum of 0.16 for two fungal species *Hemileia vastatrix* and *Phakopsora pachyrhizi*. The first position of *hemileia vastatrix* is surprising given the absence of coffee crops in France, whereas that of *Phakopsora pachyrhizi* was more expected since this species belongs to the EPPO Alert List. France shares the same cell of the map (cell n°93) with others European countries (Table 5). In Guyana risk values are higher. It must be noted that *Hemileia vastatrix* has the third higher risk value (0.80). The species that has the highest risk for Guyana is *Uromyces appendiculatus* (0.89).

Table 4: List of species which have the highest potential risk of invasion in France based on the SOM analysis. Only absent species in France are shown.

Name	Code	Risk
Hemileia vastatrix	HEMIVA	0.16
Phakopsora pachyrhizi	PHAKPA	0.16
Puccinia purpurea	PUCCPU	0.15
Mycosphaerella fijiensis	MYCOFI	0.14
Puccinia polysora	PUCCPY	0.14
Mycosphaerella musicola	MYCOMU	0.13
Glomerella tucumanensis	GLOMTU	0.13
Fusarium oxysporum f.sp. Cubense	FUSACB	0.11
Mycosphaerella henningsii	MYCOMH	0.10
Cronartium comandrae	CRONPO	0.10
Cercospora sorghi	CERCSG	0.10
Elsinoë fawcettii	ELSIFA	0.10
Cronartium coleosporioides	CRONCL	0.09
Endocronartium harknessii	ENDCHA	0.09
Mycosphaerella berkeleyi	MYCOBE	0.09
Sporisorium cruentum	SPHTCR	0.09
Tilletia controversa	TILLCO	0.09
Chrysomyxa arctostaphyli	CHRYAR	0.09
Sclerophthora macrospora	SCPHMA	0.09
Phytophthora colocasiae	PHYTOO	0.09
Ustilago scitaminea	USTISC	0.08
Peronosclerospora sorghi	PRSCSO	0.08
Corticium salmonicolor	CORTSA	0.08
Elsinoë batatas	ELSIBA	0.08
Puccinia arachidis	PUCCAR	0.08
Colletotrichum musae	COLLSU	0.08
Puccinia kuehnii	PUCCKU	0.08
Phellinus noxius	PHELNO	0.07
Apiosporina morbosa	DIBOMO	0.07
Gymnosporangium clavipes	GYMNCL	0.07
Bipolaris sacchari	DRECSA	0.07
Cercospora elaeidis	CERCEL	0.07
Claviceps africana	CLAVAF	0.07
Puccinia hemerocallidis	PUCCHM	0.07
Fusarium oxysporum f.sp. radicle-lycopersici	FUSARL	0.07

Table 5: List of countries assigned to cell 93.

Countries in cell n°93
Austria
Bulgaria
Czech republic
Denmark
France
Germany
Great Britain
Greece
Hungary
Italia
Netherlands
Poland
Romania
Russia
Spain
Switzerland
Turkey
Ukraine
Yugoslavia

Table 6: List of species which have the highest potential risk of invasion in Guyana based on the SOM analysis. Only absent species in Guyana are shown. Oomycete species are highlighted in blue.

Name	Code	Risk
<i>Uromyces appendiculatus</i>	UROMAP	0.89
<i>Setosphaeria turcica</i>	SETOTU	0.82
<i>Hemileia vastatrix</i>	HEMIVA	0.80
<i>Alternaria porri</i>	ALTEPO	0.78
<i>Mycosphaerella fijiensis</i>	MYCOFI	0.71
<i>Alternaria dauci</i>	ALTEDA	0.69
<i>Phaeoisariopsis griseola</i>	PHAIGR	0.67
<i>Lasiodiplodia theobromae</i>	PHYORH	0.62
<i>Leveillula taurica</i>	LEVETA	0.60
<i>Sporisorium sorghi</i>	SPHTSO	0.59
<i>Plasmopara viticola</i>	PLASVI	0.57
<i>Cochliobolus sativus</i>	COCHSA	0.57
<i>Alternaria brassicicola</i>	ALTEBI	0.51
<i>Phakopsora pachyrhizi</i>	PHAKPA	0.50
<i>Colletotrichum musae</i>	COLLMU	0.50
<i>Colletotrichum lindemuthianum</i>	COLLLD	0.49
<i>Didymella bryoniae</i>	DIDYBR	0.48
<i>Phellinus noxius</i>	PHELNO	0.43
<i>Alternaria padwickii</i>	ALTEPD	0.42
<i>Sphacelotheca reiliana</i>	SPHTRE	0.41
<i>Penicillium expansum</i>	PENIEX	0.41
<i>Penicillium italicum</i>	PENIIT	0.41
<i>Sporisorium cruentum</i>	SPHTCR	0.40
<i>Pythium aphanidermatum</i>	PYTHAP	0.40
<i>Penicillium digitatum</i>	PENIDI	0.39
<i>Colletotrichum capsici</i>	COLLCA	0.38
<i>Cercospora nicotianae</i>	CERCNI	0.38
<i>Puccinia graminis</i>	PUCCGR	0.37
<i>Guignardia musae</i>	GUIGMU	0.35
<i>Botryosphaeria ribis</i>	BOTSRI	0.35
<i>Phoma caricae-papayae</i>	ASCOCA	0.35
<i>Cochliobolus carbonum</i>	COCHCA	0.35
<i>Pythium splendens</i>	PYTHSL	0.35
<i>Heterobasidion abietinum</i>	HETEANSI	0.34
<i>Melampsora medusae</i>	METUMU	0.33
<i>Puccinia kuehnii</i>	PUCCKU	0.33
<i>Peronospora hyoscyami</i>	PEROHY	0.32
<i>Ganoderma philippii</i>	GANOPH	0.32
<i>Alternaria longipes</i>	ALTELO	0.32
<i>Puccinia triticina</i>	PUCCRT	0.32
<i>Fusarium oxysporum</i> f.sp. <i>Conglutinans</i>	FUSACO	0.31
<i>Stenocarpella macrospora</i>	DIPDMC	0.31
<i>Peronospora parasitica</i>	PEROPA	0.30
<i>Peronosclerospora sorghi</i>	PRSCSO	0.30
<i>Phakopsora euvitis</i>	PHAKEU	0.30

## IV) Discussion

While classical statistical multivariate analyses have been commonly used in ecology, their limitations appear when large and complex ecological datasets have to be analyzed. Advanced modeling tools are then required in order to extract the maximum of knowledge from the data (Begon *et al.* 1996), as for example the Classification and Regression Trees, or the Artificial Neural Networks (ANN) (Giraudel and Lek 2001; Brosse *et al.* 2001).

ANN are particularly tolerant to noisy data (Hepner *et al.*, 1990), are able to predict non-linear data and are powerful to represent even more complex relationships between variables (Rumelhart *et al.* 1986). Considering our data, this method was clearly appropriate for this study. SOM algorithm assists organization and visualization of data by arranging the distribution of geographical units onto a two-dimensional space, the map. In our study the SOM can indicate countries where a new invasion should alert biosecurity authorities by grouping countries with similar fungal and oomycete species assemblages. One of the advantages provided by this algorithm is its capacity to detect outliers (Lippman 1987) and to confine them on a part of the map without affecting the rest. In our study, cluster 1 and 6 are an example of that. Actually, these two clusters are characterized by the low number of species in geographical units that composed them. The algorithm has gathered these geographical units into two clusters, which are closed on the dendrogram and separate from others. Another advantage of the SOM algorithm is the possibility to give a meaning to the internal parameters which are the connection weights. In our study, these weights are interpreted as an index of risk that the species may invade a new geographical unit if it has the opportunity. Therefore our results highlight countries where a new potential invasion should alert biosecurity authorities. A limitation of the SOM is that the final map depends on the initial conditions. To control the stability of our results, different maps have been computed with several random initializations. It turned out that clusters were stable and that risk values (for species and geographical units of interest) showed minor changes.

As shown by Figure 4, all European countries except Belarus were gathered in the cluster 7. The cluster 7 seems to be relatively stable. As a matter of fact, whatever the map sizes tested, European countries were always contained in the same cluster, meaning that these countries share similar fungal and oomycete species assemblages. Moreover many European countries were assigned to the same cell (Table 5). They consequently have the same risk of invasion values for all species. However, all these countries do not present similar conditions. For example, between Spain and Poland, climatic conditions are different

as well as crops or forest tree species. Several factors are governing the success of a biological invasion. The presence of a host-plant and suitable climate are crucial but others significant factors can be associated as fecundity, environmental tolerance, competitive ability, vulnerability to predation and disease, flexibility and some others. History is also a factor that can affect and explain assemblages. For example, New Zealand's fungal and oomycete species assemblage is in the same cluster as European countries (cluster7). This result is not immediately apparent but if history of New Zealand's agriculture is examined, we realize that many crops and garden plants are originated from Mediterranean regions. Moreover New Zealand presents some similar transitional climates as some Mediterranean countries (Gevrey *et al.* 2006). Risk values for European countries are low and can be explained by the opened borders between European countries, which facilitate transport of fungal and oomycete species. Transports between European countries exist since long times, which may suggest that species have ever invaded this area.

To avoid pathogens invading Europe from adjacent countries, some organizations have proposed measures. For instance, EPPO (European and Mediterranean Plant Protection Organization, [www.eppo.org](http://www.eppo.org)), created three lists of species: A1, A2 and Alert. In Alert list is composed of all species that potentially present a phytosanitary risk for EPPO region. Pests taken into account are insects, nematodes, fungi, bacteria, viruses and invasive plants. Knowledge on risk value for these species can help to evaluate and complete these lists. Results of risk values for Alert species are not relevant because they are quite low, except for two species with a risk superior to 0.4, which are *Mycosphaerella pini* and *Puccinia horiana*. Risk values are low for European countries so these values are not a surprise but rank of species reveal the same information as risk values. Reasons for considering inclusion on the Alert List can be of various natures: pests which are new to science, new outbreaks, reports of spread, etc what is not very definite. Indeed species can be on the list but not presented a risk to invade. Species present in the A1 and A2 list are recommended for regulation as quarantine pest. Within A1 list are species still absent in Europe whereas in A2 list are species already present locally. In this study, risk and range of A1 species are calculated revealing that all species possess a low value of risk and a range covering the entire list. These results cannot bring to interpretation because A1 species do not present difference in term of risk and rank with others absent species.

The index of risk of invasion computed for *Hemileia vastatrix* and representing on the Figure 5 shows that this species has high risk values for southern hemisphere geographic units. Nevertheless, it is a coffeea pathogen that can explain its distribution. The coffee rust pathogen *Hemileia vastatrix* is a 'classic' among plant diseases, being one of the first to be

fully elucidated. The fungus is a co-evolved pathogen of *Coffea* spp. in Africa. It spread rapidly throughout Africa and Asia as the coffee industries of these countries developed; it reached the New World in 1971 and now occurs in virtually all coffee-producing countries. The major effect of coffee rust is to cause premature shedding of leaves; this reduces the photosynthetic capacity of the plant and restricts the growth of new stems on which the next season's crop is borne. The disease can render coffee cultivation uneconomic wherever it reaches epidemic proportions. Total world production of green coffee was estimated by FAO as about 7.2 million metric tons in 2003. Most of this was Arabica coffee, mostly from Latin America (but also eastern Africa and India), Robusta from Africa and Asia and other coffees. Highest risk value for France is 0.16 for *Hemileia vastatrix* and *Phakopsora pachyrhizi*. *Hemileia vastatrix* is a coffee pathogen and *Phakopsora pachyrhizi* causes soyabean rust. The fact that the coffee is not cultivated in France might explain its low value. Therefore, this risk value does not need to keep attention. As far as concerned *Phakopsora pachyrhizi*, the range of this species in the list of risk values for France is interesting. Indeed, this species is not present in EPPO regions but is considered at risk in Southern Mediterranean countries. Even if its absolute risk value is low, having the species in the highest species at risk list for France (and others European countries) confirms the fears of EPPO experts.

Moreover, the example of Guyana is more relevant with *Hemileia vastatrix* which has the third highest risk for this country. Inversely considering this species, Guyana is the country for which *Hemileia vastatrix* has the highest risk value. In the list of risk for this geographical unit, the highest risk value is 0.89 for *Uromyces appendiculatus*. This fungal species affects some of Fabaceae like soyabean or lima bean, and it is present in all border countries of Guyana (Suriname, Brazil and Venezuela). Moreover soyabean is cultivated in Guyana (FAOSTAT, FAO, Rome, 2007). Arguments like the high risk value and soyabean crop present in Guyana should be evaluated seriously by experts and may lead to measures.

In this study, we use information available on the geographical distribution of a wide range of fungal and oomycete species. This type of data is rather rare especially for fungi and oomycetes which are poorly represented in invasive species databases (Desprez-Loustau *et al.* 2007). Moreover for all the 426 species the international distribution is known without country missing. One of the weak points of international dataset is the disparity on sampling. Indeed, all geographical units are not sampled in the same manner and with the same quality. Countries do not invest the same effort, do not have the same budget or neither the same priorities to this kind of investigation. These differences are totally present on our data and consequences are found in our results. Like seen previously, two clusters are characterized by geographical units with low number of species (cluster 1 and 6). It is understandable that

desert areas or cold ones do not receive many species. However, Brazil, a country covered by a great diversity of trees, should not contain such a small number of species. This area is thus a straight example of sampling bias. However, the Crop Protection Compendium contains forest pathogens only since 2004, so this kind of pathogens is certainly less identified than others. Another problem is comparison of larger geographical units with smaller ones. To reduce this bias, larger geographical units have been as far as possible subdivided in regions. Although in our data borders of geographical units are not always clearly determinate what is common in ecological studies.

Parameters of the model were selected in part by default. For example initialization mode of virtual vectors was chose in the way to generate only positive values and sequential algorithm has been used at the expense of batch one. So others intern parameters of this model can vary and it would be interesting to modify them to test variations. Moreover in this analysis only the Euclidean distance was tested. If connection weights can be interpreted as risk whatever the distance, then different distances should be tested to compare results and perhaps find one more adapted to this kind of data. This study can be improved with a deepening of the method and a larger number of species if it is possible, to perform assemblages and therefore increase quality and stability of results.

## References

- Begon, M. Harper, J.L. and Towsend, C.R. (1996). "Ecology: Individuals, Populations and Communities. 3<sup>rd</sup> edn. Blackwell Science, Oxford, Uk.
- Brosse, S., J. L. Giraudel, et al. (2000). Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. 2nd international Conference on Applications of Machine Learning to Ecological Modelling, Adelaide, Australia, Elsevier Science Bv.
- Cereghino, R. and Y. S. Park (2009). "Review of the Self-Organizing Map (SOM) approach in water resources: Commentary." Environmental Modelling & Software **24**(8): 945-947.
- Daelher CC (2001). "Two ways to be an invader, but one is more suitable for ecology." Bulletin of the Ecological Society of America **82**: 101-102.
- David, MA. and Thompson K (2000). "Eight ways to be a colonizer; two ways to be an invader: a propose nomenclature scheme for invasion ecology". Bulletin of the Ecological Society of America **81**: 226-230.
- Desprez-Loustau, M.-L., C. Robin, et al. (2007). "The fungal dimension of biological invasions." Trends in Ecology & Evolution **22**(9): 472-480.
- Facon, B., B. J. Genton, et al. (2006). "A general eco-evolutionary framework for understanding bioinvasions." Trends in Ecology & Evolution **21**(3): 130-135.
- Facon, B., P. Jarne, et al. (2005). "Hybridization and invasiveness in the freshwater snail *Melanoides tuberculata*: hybrid vigour is more important than increase in genetic variance." Journal of Evolutionary Biology **18**(3): 524-535.
- Gevrey, M., S. Worner, et al. (2006). "Estimating risk of events using SOM models: A case study on invasive species establishment." Ecological Modelling **197**(3-4): 361-372.
- Giraudel, J. L. and S. Lek (2000). "A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination." 2nd international Conference on Applications of Machine Learning to Ecological Modelling, Adelaide,

Australia, Elsevier Science Bv.

Hepner, G. F., T. Logan, et al. (1990). "Artificial neural network classification using a minimal training set comparison to conventional supervised classification." Photogrammetric Engineering and Remote Sensing **56**(4): 469-473.

Heybroek, H. M., D. M. Elgersma, et al. (1982). "Dutch elm disease an ecological accident." Outlook on Agriculture **11**(1): 1-9.

Inderjit (2005). "Plant invasions: Habitat invasibility and dominance of invasive plant species." Plant and Soil **277**(1-2): 1-5.

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

Lin, Y. C., G. Liu, et al. (2009). "Prediction of static recrystallization in a multi-pass hot deformed low-alloy steel using artificial neural network." Journal of Materials Processing Technology **209**(9): 4611-4616.

Lippman, R.P. (1987). "An introduction to computing with neural nets." IEEE Acoustics, Speech and Signal processing Magazine **4**: 4-22.

Miller, N., A. Estoup, et al. (2005). "Multiple transatlantic introductions of the western corn rootworm." Science **310**(5750): 992-992.

Orwig, D. A. (2002). "Ecosystem to regional impacts of introduced pests and pathogens: historical context, questions and issues." Journal of Biogeography **29**(10-11): 1471-1474.

Pimentel, D., L. Lach, et al. (2000). "Environmental and economic costs of nonindigenous species in the United States." Bioscience **50**(1): 53-65.

Pimentel, D., S. McNair, et al. (2001). "Economic and environmental threats of alien plant, animal, and microbe invasions." Agriculture, Ecosystems & Environment **84**(1): 1-20.

Platt, R. (2008). "Potato diseases caused by oomycete species pathogens." Cahiers Agricultures **17**(4): 361-367.

Rejmanek M, Richardson DM and al (2002). Biological invasions: politics and the discontinuity of ecological terminology. Bulletin of the Ecological Society of America **83**: 131-133.

Richardson, D. M., N. Allsopp, et al. (1998). Plant invasions - the role of mutualisms. Conference of the INTECOL, Florence, Italy, Cambridge Univ Press.

Rumelhart, D. E., G. E. Hinton, et al. (1986). "Learning representations by back-propagating errors." Nature **323**(6088): 533-536.

Valery, L., H. Fritz, et al. (2008). "In search of a real definition of the biological invasion phenomenon itself." Biological Invasions **10**(8): 1345-1351.

Vermeij, G. J. (1996). "An agenda for invasion biology." Biological Conservation **78**(1-2): 3-9.

Vitousek, P. M., C. M. Dantonio, et al. (1996). "Biological invasions as global environmental change." American Scientist **84**(5): 468-478.

Waite, S (2000). "Statistical Ecology in Practice. A Book to Analysing Environmental and Ecological Field data. Pentice Hall, London.

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

Appendix: Correspondence between names of geographical units and the three letters code.

LEVEL_3_CO	LEVEL_NAME
ABT	Alberta
AFG	Afghanistan
AGE	Argentina
ALA	Alabama
ALB	Albania
ALD	Aldabra
ALG	Algeria
ALT	Altay
AMU	Amur
AND	Andaman Is.
ANG	Angola
ANT	Antarctica
ARI	Arizona
ARK	Arkansas
ARU	Aruba
ASC	Ascension
ASK	Alaska
ASS	Assam
AUT	Austria
AZO	Azores
BAH	Bahamas
BAL	Baleares
BAN	Bangladesh
BEN	Benin
BER	Bermuda
BGM	Belgium
BIS	Bismarck Archipelago
BKN	Burkina
BLR	Belarus
BLT	Baltic States
BLZ	Belize
BOL	Bolivia
BOR	Borneo
BOT	Botswana
BRC	British Columbia
BUL	Bulgaria
BUR	Burundi
BZC	Brazil West-Central
BZE	Brazil Northeast
BZL	Brazil Southeast
BZN	Brazil North
BZS	Brazil South
CAF	Central African Republic
CAL	California
CAY	Cayman Is.
CBD	Cambodia
CHA	Chad
CHC	China South-Central
CHH	Hainan
CHI	Inner Mongolia
CHM	Manchuria
CHN	China North-Central
CHQ	Qinghai
CHS	China Southeast
CHT	Tibet
CHX	Xinjiang
CKI	Cocos (Keeling) I.
CLC	Chile Central
CLM	Colombia
CMN	Cameroon
CNT	Connecticut
CNY	Canary Is.
COL	Colorado
COM	Comoros
CON	Congo
COO	Cook Is.
COR	Corse
COS	Costa Rica
CPP	Cape Provinces
CRL	Caroline Is.
CUB	Cuba
CVI	Cape Verde
CYP	Cyprus
CZE	Czechoslovakia
DEL	Delaware
DEN	Denmark

LEVEL_3_CO	LEVEL_NAME
DJI	Djibouti
DOM	Dominican Republic
EAS	Easter Is.
EGY	Egypt
EHM	East Himalaya
ELS	El Salvador
EQG	Equatorial Guinea
ERI	Eritrea
ETH	Ethiopia
FAL	Falkland Is.
FIJ	Fiji
FIN	Finland
FLA	Florida
FOR	Føroyar
FRA	France
FRG	French Guiana
GAB	Gabon
ECU	Ecuador
GAM	Gambia, The
GEO	Georgia
GER	Germany
GGI	Gulf of Guinea Is.
GHA	Ghana
GIL	Gilbert Is.
GNB	Guinea-Bissau
GNB	Guinea-Bissau
GNL	Greenland
GRB	Great Britain
GRC	Greece
GST	Gulf States
GUA	Guatemala
GUI	Guinea
GUY	Guyana
HAI	Haiti
HAW	Hawaii
HON	Honduras
HUN	Hungary
ICE	Iceland
IDA	Idaho
ILL	Illinois
IND	India
INI	Indiana
IOW	Iowa
IRE	Ireland
IRN	Iran
IRQ	Iraq
ITA	Italy
IVO	Ivory Coast
JAM	Jamaica
JAP	Japan
JAW	Jawa
KAN	Kansas
KAZ	Kazakhstan
KEN	Kenya
KER	Kermadec Is.
KGZ	Kirgizistan
KOR	Korea
KRI	Kriti
KRY	Krym
KTY	Kentucky
KUW	Kuwait
LAO	Laos
LBR	Liberia
LBS	Lebanon-Syria
LBY	Libya
LDV	Laccadive Is.
LEE	Leeward Is. AB Ant
LES	Lesotho
LIN	Line Is.
LOU	Louisiana
LSI	Lesser Sunda Is.
MAI	Maine
MAN	Manitoba
MAS	Massachusetts
MAU	Mauritius
MDG	Madagascar

LEVEL_3_CO	LEVEL_NAME
MDR	Madeira
MDV	Maldives
MIC	Michigan
MIN	Minnesota
MLI	Mali
MLW	Malawi
MLY	Malaya
MNT	Montana
MOL	Maluku
MON	Mongolia
MOR	Morocco
MOZ	Mozambique
MRN	Marianas
MRS	Marshall Is.
MRY	Maryland
MSI	Mississippi
MSO	Missouri
MTN	Mauritania
MXC	Mexico Central
MYA	Myanmar
NAM	Namibia
NBR	New Brunswick
NCA	North Carolina
NDA	North Dakota
NEB	Nebraska
NEP	Nepal
NET	Netherlands
NEV	Nevada
NFK	Norfolk Is.
NFL	Newfoundland
NGA	Nigeria
NGR	Niger
NIC	Nicaragua
NLA	Netherlands Antilles
NNS	Nansei-shoto
NOR	Norway
NRU	Nauru
NSC	Nova Scotia
NSW	New South Wales
NTA	Northern Territory
NUE	Niue
NUN	Nunavut
NWC	New Caledonia
NWG	New Guinea
NWH	New Hampshire
NWJ	New Jersey
NWM	New Mexico
NWT	Northwest Territories
NWY	New York
NZN	New Zealand
OHI	Ohio
OKL	Oklahoma
OMA	Oman
ONT	Ontario
ORE	Oregon
PAK	Pakistan
PAL	Palestine
PAN	Panama
PAR	Paraguay
PEI	Prince Edward I.
PEN	Pennsylvania
PER	Peru
PHI	Philippines
PIT	Pitcairn Is.
POL	Poland
POR	Portugal
PUE	Puerto Rico
QLD	Queensland
QUE	Québec
REU	Réunion
RHO	Rhode I.
ROD	Rodrigues
ROM	Romania
RUC	Central European Russia
RWA	Rwanda
SAM	Samoa

LEVEL_3_CO	LEVEL_NAME
SAR	Sardegna
SAS	Saskatchewan
SAU	Saudi Arabia
SCA	South Carolina
SDA	South Dakota
SEN	Senegal
SEY	Seychelles
SIC	Sicilia
SIE	Sierra Leone
SOA	South Australia
SOL	Solomon Is.
SOM	Somalia
SPA	Spain
SRL	Sri Lanka
SSA	South Sandwich Is.
STH	St.Helena
SUD	Sudan
SUL	Sulawesi
SUM	Sumatera
SUR	Surinam
SVA	Svalbard
SWE	Sweden
SWI	Switzerland
SWZ	Swaziland
TAI	Taiwan
TAN	Tanzania
TAS	Tasmania
TCI	Turks-Caicos Is.
TDC	Tristan da Cunha
TEN	Tennessee
TEX	Texas
THA	Thailand
TKM	Turkmenistan
TOG	Togo
TOK	Tokelau-Manihiki
TON	Tonga
TRT	Trinidad-Tobago
TUB	French Polynesia
TUN	Tunisia
TUR	Turkey
TUV	Tuvalu
TZK	Tadzhikistan
UGA	Uganda
UKR	Ukraine
URU	Uruguay
UTA	Utah
UZB	Uzbekistan
VAN	Vanuatu
VEN	Venezuela
VER	Vermont
VIC	Victoria
VIE	Vietnam
VRG	Virginia
WAK	Wake I.
WAL	Wallis-Futuna Is.
WAS	Washington
WAU	Western Australia
WDC	District of Columbia
WHM	West Himalaya
WIN	Windward Is.
WIS	Wisconsin
WSA	Western Sahara
WVA	West Virginia
WYO	Wyoming
XMS	Christmas I.
YEM	Yemen
YUG	Yugoslavia
YUK	Yukon
ZAI	Zaire
ZAM	Zambia
ZIM	Zimbabwe
ZAI	Zaire
ZAM	Zambia
ZIM	Zimbabwe