Predicting invasion success of forest pathogenic fungi from species traits

Aurore Philibert[1,2], Marie-Laure Desprez-Loustau[3]*, Bénédicte Fabre[4], Pascal Frey[4], Fabien Halkett[4], Claude Husson[4], Brigitte Lung-Escarmant[3], Benoit Marçais[4], Cécile Robin[3], Corinne Vacher[3] and David Makowski[1,2]

[1]INRA, UMR211 INRA-AgroParisTech BP01 78850 Thiverval-Grignon, France; [2]INRA, UMR518 INRA-AgroParisTech 16 rue Claude Bernard 75005 Paris, France; [3]INRA, UMR1202 BIOGECO INRA-Université Bordeaux 1, Equipe de Pathologie forestière, 69 route d'Arcachon, Pierroton, 33610 Cestas, France; and [4]INRA, UMR1136 IAM Nancy-Université, Equipe Ecologie des Champignons Pathogènes Forestiers, 54280 Champenoux, France

Summary

1. Biological invasions are a major consequence of globalization and pose a significant threat to biodiversity. Because only a small fraction of introduced species become invasive, identification of those species most likely to become invasive after introduction is highly desirable to focus management efforts. The predictive potential of species-specific traits has been much investigated in plants and animals. However, despite the importance of fungi as a biological group and the potentially severe effects of pathogenic fungi on agrosystems and natural ecosystems, the specific identification of traits correlated with the invasion success of fungi has not been attempted previously.

2. We addressed this question by constructing an *ad hoc* data set including invasive and non-invasive species of forest pathogenic fungi introduced into Europe. Data were analysed with a machine learning method based on classification trees (Random Forest). The performance of the classification rule based on species traits was compared with that of several random decision rules, and the principal trait predictors associated with invasive species were identified.

3. Invasion success was more accurately predicted by the classification rule including biological traits than by random decision rules. The effect of species traits was maintained when confounding variables linked to residence time and habitat availability were included. The selected traits were unlikely to be affected by a phylogenetic bias as invasive and non-invasive species were evenly distributed in fungal clades.

4. The species-level predictors identified as useful for distinguishing between invasive and non-invasive species were traits related to long-distance dispersal, sexual reproduction (in addition to asexual reproduction), spore shape and size, number of cells in spores, optimal temperature for growth and parasitic specialization (host range and infected organs).

5. *Synthesis and applications.* This study demonstrates that some species-level traits are predictors of invasion success for forest pathogenic fungi in Europe. These traits could be used to refine current pest risk assessment (PRA) schemes. Our results suggest that current schemes, which are mostly based on sequential questionnaires, could be improved by taking into account trait interactions or combinations. More generally, our results confirm the interest of machine learning methods, such as Random Forest, for species classification in ecology.

Key-words: alien, biological invasion, emerging disease, forest pathology, invader, invasiveness, parasitism, pest risk assessment, *Phytophthora*, Random Forest

*Correspondence author. E-mail: loustau@bordeaux.inra.fr

Introduction

Biological invasions, resulting from deliberate and unintentional species transfers, are a major consequence of globalization and pose a significant threat to biodiversity (Levine & D'Antonio 2003; Hulme 2009; Vilà *et al.* 2010). Many authors have ruled out a purely idiosyncratic phenomenon resulting from complex interactions between the history of introduction, the invading species and the recipient community and have sought to identify consistent characteristics of invading species or of invaded communities (Kolar & Lodge 2001; Gasso *et al.* 2009; Sakai *et al.* 2001). This important area of research is particularly relevant to risk assessment and management, as it provides a knowledge-based approach to identification of the most threatening species and the most vulnerable communities (Stohlgren & Schnase 2006). The species traits favouring invasion success (invasiveness) have been extensively investigated in animals and plants. As only a small fraction of introduced species become invasive (Williamson & Fitter 1996), it has been suggested that invaders possess traits favouring their successful establishment, spread and impact on recipient communities. In their recent cross-taxa review, Hayes & Barry (2008) identified climate/habitat match as the only characteristic consistently and significantly associated with invasive behaviour across biological groups. However, some specific attributes were found within groups, such as vigorous vegetative growth, long flowering period and attractiveness to humans for plants (Lloret *et al.* 2005; Pyšek & Richardson 2007).

The fungi constitute a large biological group, but the specific traits correlated with their invasion success have not yet been investigated (Desprez-Loustau *et al.* 2007). Furthermore, very few studies have addressed accidental introductions (Hayes & Barry 2008).

In this study, we aimed to identify traits linked to invasiveness in forest pathogenic fungi. The high environmental, economic and social impacts of introduced pathogens in forest ecosystems (Sache *et al.* 2011) and the failure of current regulatory measures to control them justify efforts to develop predictive approaches of invasion success of these organisms (Brasier 2008). Davis (2003) also suggested that greater emphasis in invasion ecology should be placed on intertrophic interactions between introduced species and long-term residents, because these interactions are more likely to cause extinction than interactions at the same trophic level. Indeed, several tree pathogens, such as *Ophiostoma novo-ulmi*, the agent of Dutch elm disease, or *Cryphonectria parasitica*, the agent of chestnut blight, almost drove their host species to extinction in the area of introduction (Desprez-Loustau *et al.* 2007).

We addressed this question by constructing an *ad hoc* data set for invasive and non-invasive species of forest pathogenic fungi introduced into Europe. Data were analysed with a machine learning method based on classification trees (Random Forest, RF) (Cutler *et al.* 2007; Stohlgren *et al.* 2010). The performance of the RF classification rule based on species traits was compared with that of several random decision rules. This approach made it possible to identify the most important species-level predictors for distinguishing between invasive and non-invasive species, with a view to improving pest risk assessment procedures in particular.

Materials and methods

DATA

Species

Fungal species alien to Europe were recently listed in the European Union (EU)-funded DAISIE project (Desprez-Loustau 2009; Desprez-Loustau *et al.* 2010). Forty of these alien species were forest tree pathogens. Several data sources (European Plant Protection Organisation, EPPO, particularly interception data sets; Plant Protection services, bibliographic data bases) were used to add to the DAISIE data set, to obtain a more comprehensive set of species both alien to Europe (EU27, i.e. the 27 countries included in EU, Norway, Switzerland, and former Yugoslavia) and reported at least once in this area before May 2009. The resulting data set includes a total of 47 species: 40 true fungi (i.e. belonging to the Eumycota kingdom) and seven *Phytophthora* species (belonging to Oomycota, in the Stramenopila kingdom; Fig. 1). We included these seven *Phytophthora* species because, despite the distant relationship of Oomycota to Eumycota in the phylogenetic tree, there is strong morphological, biological and ecological convergence of *Phytophthora* with fungi (filamentous mycelium, production of spores and pathogenicity to plants). Separate analyses were performed with and without *Phytophthora* species.

Response variable

The delimitation of two groups of species defined as 'invasive' or 'non-invasive' in Europe was not straightforward. We overcame the problems associated with different subjective judgements and the use of different terminologies to determine the status of individual species, by basing our classification on objective criteria. The 47 species were assigned to seven classes defining different stages in the introduction–invasion continuum (Table 1). Eventually, species not established in natural habitats were considered to be non-invasive (19 species), and species established in forests, on indigenous tree species (28 species), were considered to be invasive. This approach yielded two groups of comparable size, and the invasion criterion corresponded to the emergence of a new disease.

Explanatory variables

We established an initial list including many traits thought to be potential drivers of the successful transition between different stages of invasion: survival, population build-up, spread and impact. We tried to document these traits for the 47 species by an intensive search in bibliographic data bases, the Crop Protection Compendium (CABI 2009, http://www.cabicompendium.org/cpc/home.asp) and reference books (Viennot-Bourgin 1949; Lanier *et al.* 1976; Sinclair, Lyon & Johnson 2005) as the main data sources. The variables from this initial list that were not documented for a large proportion of species, and/or for which the level of uncertainty was high, were removed, and the definition of the remaining variables was refined to avoid ambiguities. Each species profile was then completed and checked by at least two people. The variables were initially binary, categorical or quantitative. All variables were transformed into binary variables, to ensure a similar weight in analyses and to facilitate interpretation based on biological hypotheses. For example, spore volume and shape were preferred to spore dimensions, included in the species
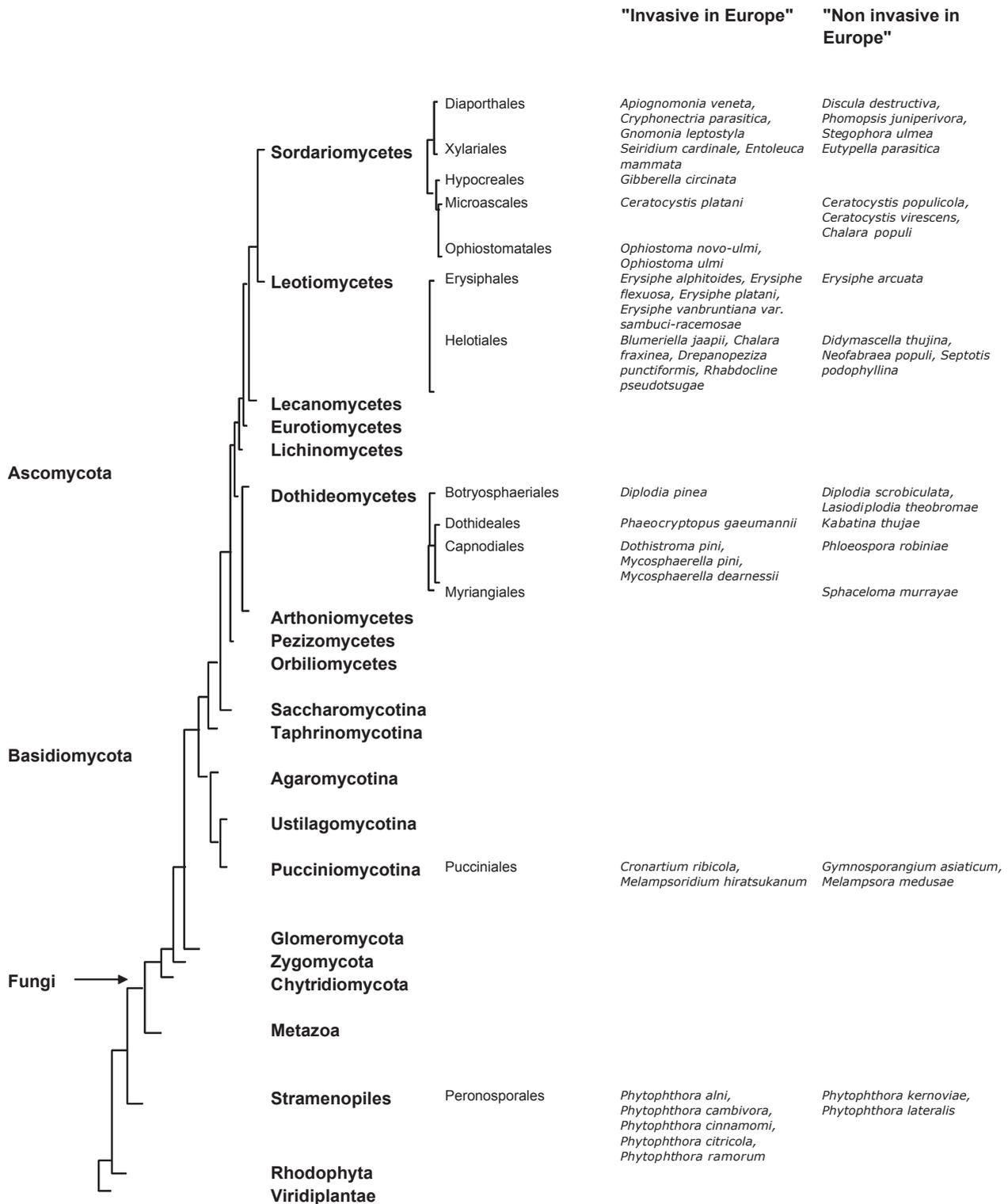
		"Invasive in Europe"	**"Non invasive in Europe"**
Sordariomycetes	Diaporthales	*Apiognomonia veneta, Cryphonectria parasitica, Gnomonia leptostyla*	*Discula destructiva, Phomopsis juniperivora, Stegophora ulmea*
	Xylariales	*Seiridium cardinale, Entoleuca mammata*	*Eutypella parasitica*
	Hypocreales	*Gibberella circinata*	
	Microascales	*Ceratocystis platani*	*Ceratocystis populicola, Ceratocystis virescens, Chalara populi*
	Ophiostomatales	*Ophiostoma novo-ulmi, Ophiostoma ulmi*	
Leotiomycetes	Erysiphales	*Erysiphe alphitoides, Erysiphe flexuosa, Erysiphe platani, Erysiphe vanbruntiana var. sambuci-racemosae*	*Erysiphe arcuata*
	Helotiales	*Blumeriella jaapii, Chalara fraxinea, Drepanopeziza punctiformis, Rhabdocline pseudotsugae*	*Didymascella thujina, Neofabraea populi, Septotis podophyllina*
Lecanomycetes			
Eurotiomycetes			
Lichinomycetes			
Dothideomycetes	Botryosphaeriales	*Diplodia pinea*	*Diplodia scrobiculata, Lasiodiplodia theobromae*
	Dothideales	*Phaeocryptopus gaeumannii*	*Kabatina thujae*
	Capnodiales	*Dothistroma pini, Mycosphaerella pini, Mycosphaerella dearnessii*	*Phloeospora robiniae*
	Myriangiales		*Sphaceloma murrayae*
Arthoniomycetes			
Pezizomycetes			
Orbiliomycetes			
Saccharomycotina			
Taphrinomycotina			
Agaromycotina			
Ustilagomycotina			
Pucciniomycotina	Pucciniales	*Cronartium ribicola, Melampsoridium hiratsukanum*	*Gymnosporangium asiaticum, Melampsora medusae*
Glomeromycota			
Zygomycota			
Chytridiomycota			
Metazoa			
Stramenopiles	Peronosporales	*Phytophthora alni, Phytophthora cambivora, Phytophthora cinnamomi, Phytophthora citricola, Phytophthora ramorum*	*Phytophthora kernoviae, Phytophthora lateralis*
Rhodophyta			
Viridiplantae			

Ascomycota, Basidiomycota, Fungi

Fig. 1. Location of the species in the phylogenetic tree of fungi and pseudo-fungi = Stramenopiles (from James *et al.* 2006). Invasive and non-invasive refer to the classification in Table 1.

descriptions, as more ecologically meaningful. They were computed from the length and width of mitospores (the spores produced by asexual reproduction, present in all 47 species) and were redefined as categorical variables using threshold values. The final data set included 19 binary variables, among which six had missing data (Table 2).

Previous analyses of relationships between species traits and invasion success have identified several factors that may be important sources of bias, phylogeny and residence time in particular (Wilson *et al.* 2007; Hayes & Barry 2008; Pyšek, Křivánek & Jarošík 2009). The potential bias owing to phylogeny was probably limited in our study, because invasive species and non-invasive species were equally

Table 1. Classification of fungal species based on their invasion stage in Europe. Establishment implies repeated reports of the disease-causing organism. Spread was estimated from the distribution of the species in European countries (Desprez-Loustau 2009). Ecological impact relates to the infection of an indigenous host and the severity of the damage caused

Class	Invasion stage	Number of species	Status
1	Introduced, non-established (interceptions)	5	Non-invasive
2	Introduced, non-established in natural habitats (only anthropogenic habitats, such as nurseries, gardens and parks)	14	
3	Introduced, established locally in non-anthropogenic habitats, significant local ecological impact	4	Invasive
4	Introduced, widely spread, low ecological impact	8	
5	Introduced, widely spread, low ecological but significant economic impact	6	
6	Introduced, widely spread, high ecological and medium to high economic impact	8	
7	Cryptogenic (alien or emerging) species, widely spread and with high ecological impact	2	

Table 2. Species traits and confounding variables. All variables except 'date of introduction' are binary (yes/no)

Variable	Definition	Number of missing data	Reasoning for predictor choice
Mitospore characteristics			
Small Size	Volume $< 250 \ \mu L^3$; calculated as $\frac{4\pi}{3} \left[\frac{width}{2}\right]^2 \left[\frac{length}{2}\right]$	0	As for any particle, spore size and shape affect release, transport and deposition, particularly for aerial
Medium Size	Volume between 250 and 2500 μL^3	0	pathogens (Stockmarr, Andreasen & Østergård 2007). Two binary variables related to spore volume were defined, based on the notion that different spore sizes may be associated with different strategies: small spores for higher dispersal, large spores for higher survival and medium spores representing a trade-off between the two strategies. Spore volume in 35 species of common native forest pathogenic fungi in France presented a trimodal distribution, from which the thresholds between classes were derived (M.L. Desprez-Loustau, unpublished data)
Shape	Mitospore length-to-width ratio < 3	0	The threshold value of 3 corresponds to the optimal shape for ascospores defined by a biomechanical approach (Roper *et al.* 2008)
Pigmented		0	Pigmented spores are more resistant to ultraviolet radiation (Durrell 1964), potentially increasing their survival
Pluricellular		0	The cell number of spores has been shown to be a factor associated with survival in stressful environments (Shathele 2009)
Diploid	Haploid vs. non-haploid	0	Ploidy affects fitness components, in particular spore germination (Quintanilla & Escudero 2006)
Existence of a mechanism of Long-Distance Dispersal	Mito- or meiospores vectored by wind, insects or running water (vs. rainsplash only)	0	Species with a long-distance dispersal mechanism are more likely to extend their range rapidly, thus becoming invasive
Life cycle			
Latency duration	Period of latency between mitospore infection and secondary spore production < 30 days	5	Species with a short latency (generation time) are more likely to build up large populations rapidly
Polycyclic	Number of spore production cycles per year	3	Same reasoning as for latency
Sexual reproduction		0	Sexual fruiting bodies are often resistant forms, for overwintering for example, in fungi. Sexual spores are often wind dispersed. Recombination in sexual reproduction yields individuals with novel attributes, possibly increasing virulence to new hosts (Fraser *et al.* 2005)
Selfing possible		10	Selfing may favour expansion, by overcoming the need for a sexual partner (Fraser *et al.* 2005)

Table 2. (Continued)

Variable	Definition	Number of missing data	Reasoning for predictor choice
Parasitic strategy			
Non-obligate parasite	Non-obligate parasites can live independently of their host		A saprophytic lifestyle can allow survival in the absence of suitable hosts
Monogenic	Monogenic parasites need a single host species to complete their life cycle	0	Heteroxenic parasites, which need several host species to complete their life cycle, are less likely to find a local combination of these species in the introduced area
Generalist	Host range including at least two families	0	Species with broad niches ('generalists') are expected to invade more easily than species with narrower niches ('specialists'), because they are more likely to find their resources
Infection of perennial tissues	Bark-wood vs. only expendable parts (leaves and twigs)	0	Parasites of perennial tree parts would be less likely to be affected by extinction
Seed transmission		0	Seed transmission may be linked to higher dispersal rates and impact
Endophytism	Endophytic parasites can live asymptomatically in their host until host conditions favour pathogenicity	4	Endophytic parasites are more likely to be overlooked when plants are introduced
Abiotic niche			
Optimal temperature	Optimal temperature for growth between 20 and 25 °C	12	Optimal growth temperature might be a proxy of climate-matching (the climate of origin is unknown for many species because of their unknown origin).
Climate in the area of origin	Temperate vs. others (boreal, tropical, etc., in Köppen climate maps)	14	We assumed that species with an optimum between 20 and 25 °C were more likely to be adapted to the temperate climate prevailing over much of Europe than species with an optimum below 20 or above 25 °C. Most native species of forest pathogenic fungi in France have their optimum growth temperature in the 20–25 °C range, whereas some species with lower optimum growth temperatures are found mostly in mountainous habitats and those with higher values are found in the Mediterranean area (M.L. Desprez-Loustau, unpublished data)
Confounding variables			
Date of introduction	Estimated as the date of first observation in Europe	0	Proxy of residence time; see reasoning in the main text
Potential habitat	Surface area covered by host species of the same genus: small if < 0·5% of the total EU30 forest area (Köbler & Seufert 2001), large otherwise	0	Species pathogenic on commonly occurring European tree species, such as oaks and pines, would have more opportunity to establish and to spread

frequent in most taxonomic groups (Fig. 1). Residence time was accounted for by using the date on which each species was first observed in Europe, according to previous publications and data bases. We also included a variable related to the potential habitat in the introduced area: the area covered by host tree species (Table 2). Separate analyses were performed with and without confounding variables.

STATISTICAL METHODS

The Random Forest method (RF) was used to determine whether species traits were good predictors of the invasion status of the species and to identify the best such predictors for our data set.

The Random Forest method is a machine learning classification method based on a large number of decision trees (Breiman & Cutler 2003). The goal of classification tree methods is to create a set of classification rules (the branches) from the input variables included in a training set, making it possible to predict the location of new observations (here species) within the groups (the nodes, in this case 'invasive' or 'non-invasive') from the values of the input variables for these observations. The classification rules are built by recursive binary partitioning of the data set. Each binary split, based on an explanatory variable, is selected according to an index measuring the quality of its contribution to the classification, the Gini index (Breiman *et al.* 1984; Therneau & Atkinson 1997). The main improvement of RF over single-tree methods such as CART (Classification and Regression Tree; Breiman *et al.* 1984) lies in the integration of two random selection processes in the construction of each tree, one for data (observations) and the other for variables, increasing the robustness of the method (Ghattas 2000; Prasad, Iverson & Liaw 2006). Data

samples are generated by a bootstrap technique (Efron & Tibshirani 1993). Individuals (species in our case) present in a bootstrap sample are referred to as 'in-bag' data, whereas the remaining individuals form the 'out-of-bag' (OOB) data. A non-pruned classification tree is built from each bootstrap sample with the CART method. For each tree, only a random sample of the explanatory variables is used to determine the best split at each node.

The OOB data can be used to rank the explanatory variables, through the calculation of a criterion called *mean decrease accuracy* (MDA). For each variable, MDA measures the increase in the rate of misclassification resulting from a random permutation of the values of this variable, the values of the other variables remaining unchanged. Thus, variables with a positive MDA improve the classification, and the higher the MDA, the greater the importance of the variable for classification.

The rate of correct classification by RF was calculated by leave-one-out cross-validation. Both the overall rate of misclassification (for both invasive and non-invasive species) and the rate of misclassification of invasive species were calculated. The accuracy of the species classification achieved with RF was compared with that of four reference random decision rules with different probabilities of being invasive (*P*): *P* = 0 (no invasive species), *P* = 0·5 (50% of species invasive), *P* = 1 (100% of species invasive) and *P* = proportion of invasive species in the data set (28/47 for all species, 23/40 for Eumycota). The rates of correct classification associated with these rules were calculated as:

$$\frac{(1 - P) * n_{NI} + P * n_I}{n_I + n_{NI}},$$

for all species, and *p* for invasive species only, where n_I is the number of invasive species and n_{NI} is the number of non-invasive species.

Statistical analyses were carried out with R software (http://www.R-project.org, R Development Core Team 2009) and the RANDOMFOREST package (Liaw & Wiener 2002). The three tuning parameters required by RF were set as follows:

The number of explanatory variables used at each node of each tree was set to the square root of the total number of variables, as described by Breiman (2001).

The total number of trees in the forest, corresponding to the number of bootstrap samples, was set at 100 000, after preliminary trials, as this was found to give robust results.

The number of iterations for the algorithm used to estimate missing values was set at 20.

Results

The proportions of species correctly classified with RF and with the four random decision rules are presented in Tables 3 and 4 for four data sets: the data sets including all species (i.e. Eumycota with *Phytophthora* spp) or true fungi only, with and without confounding variables. The proportion of species classified correctly was higher with RF (62·5–68·1%) than with the random decision rules (40·4–59·6%), for all four data sets (Table 3). The performance of RF was even better when judged by its classification of just the invasive species, with the frequency of correct classification reaching 73·9–82·1% (Table 4). These values were systematically higher than the proportions of well-classified species obtained with random decision rules based on *P* < 1.

The MDA values reported in Figs 2 and 3 were used to rank biological traits in the order of importance for the classification of both invasive and non-invasive species. When no
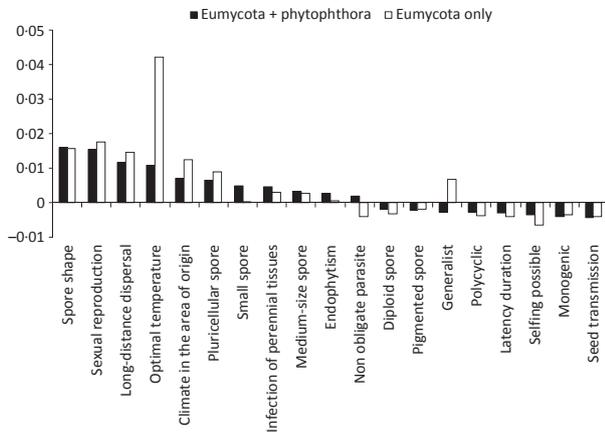
Table 3. Performance of Random Forest (including species traits) and four random decision rules (with *P*, the probability of being invasive) expressed as the percentage of well-classified invasive and non-invasive species. Confounding variables are the date of introduction and potential habitat

Data set	Random Forest	Random decision rule			
		P = proportion of invasive species in the data set	*P* = 0	*P* = 0·5	*P* = 1
Eumycota and *Phytophthora*, without confounding variables	63·8	51·8	40·4	50	59·6
Eumycota, without confounding variables	62·5	51·1	42·5	50	57·5
Eumycota and *Phytophthora*, with confounding variables	68·1	51·8	40·4	50	59·6
Eumycota, with confounding variables	62·5	51·1	42·5	50	57·5

Table 4. Performance of Random Forest (including species traits) and four random decision rules (with *P*, the probability of being invasive) expressed as the percentage of correctly classified invasive species. Confounding variables are the date of introduction and potential habitat

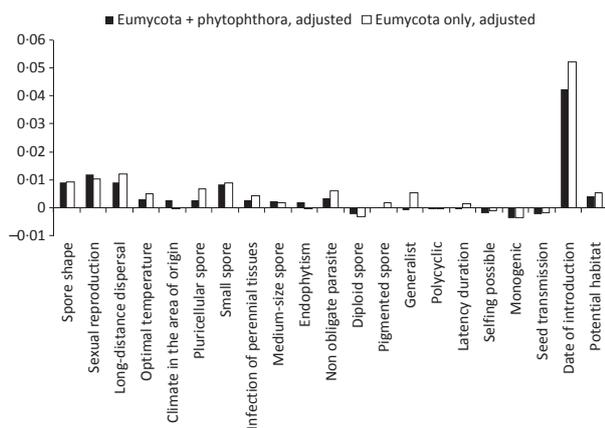| Data set	Random Forest	Random decision rule			
		P = proportion of invasive species in the data set	*P* = 0	*P* = 0·5	*P* = 1
Eumycota and *Phytophthora*, without confounding variables	78·6	59·6	0	50	100
Eumycota, without confounding variables	78·3	57·5	0	50	100
Eumycota and *Phytophthora*, with confounding variables	82·1	59·6	0	50	100
Eumycota, with confounding variables	73·9	57·5	0	50	100

Fig. 2. Importance of the variables for classifying invasive and non-invasive species, as assessed by mean decrease accuracy (MDA) calculated with Random Forest. Data sets including Eumycota + Phytophthora or Eumycota species were considered successively. Variables were sorted in decreasing order of MDA for Eumycota + Phytophthora. Confounding variables were not included in the analysis.

Fig. 4. Importance of the variables for the classification of just the invasive species, as assessed by mean decrease accuracy (MDA), calculated with Random Forest. Data sets including Eumycota + Phytophthora or Eumycota species were considered successively. Confounding variables were not included in the analysis. Variables were sorted as in Fig. 2.

confounding variables were included in the analysis, the traits with the highest MDA were spore shape, sexual reproduction, long-distance dispersal, optimal temperature and climate at the area of origin. Pluricellular spore, spore size and infection of perennial tissues also had a positive MDA for both the Eumycota + *Phytophthora* and Eumycota data sets. The host range variable (generalist) was influential in the analysis of Eumycota but not when *Phytopthora* spp were added. Most of the MDA values were lower when the two confounding variables were added to the biological traits, except for the non-obligate parasite variable which took relatively more importance (Fig. 3). However, even when the confounding variables were included in RF, spore shape, sexual reproduction and long-distance dispersal remained the biological traits with the highest MDA values.

The most important traits for the classification of just the invasive species (Fig. 4) were those already identified in Fig. 2. In addition, the polycyclic and latency duration variables also had a small positive effect on classification. As observed for the classification of all species, inclusion of the confounding variables reduced MDA for most biological traits but the variables with the highest MDA remained the same (Fig. 5).

Overall, three variables, sexual reproduction, spore shape and long-distance dispersal were consistently among the most important traits according to MDA. Several other traits had a high MDA in most classifications: optimal temperature, pluricellular spore, climate in the area at origin. The spore size, type of infected organ and Endophytism variables had a positive MDA in most cases, but with lower values. Some variables had a negative MDA in most cases, consistent with an absence of predictive value for the classification of species as invasive

Fig. 3. Importance of the variables for classifying invasive and non-invasive species, as assessed by mean decrease accuracy (MDA) calculated with Random Forest. Data sets including Eumycota + Phytophthora or Eumycota species were considered successively, with two confounding variables included in the analysis (date of introduction and potential habitat). Variables were sorted as in Fig. 2.
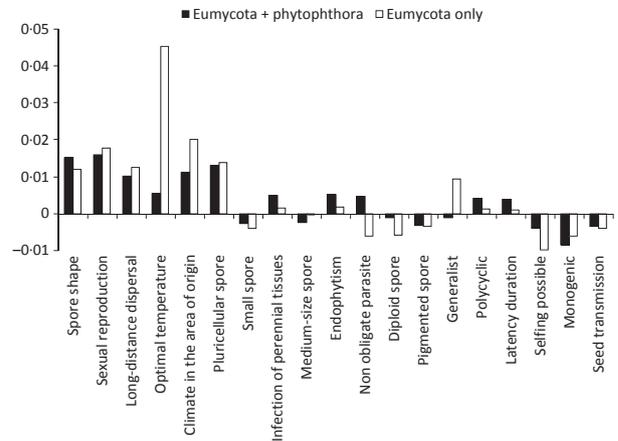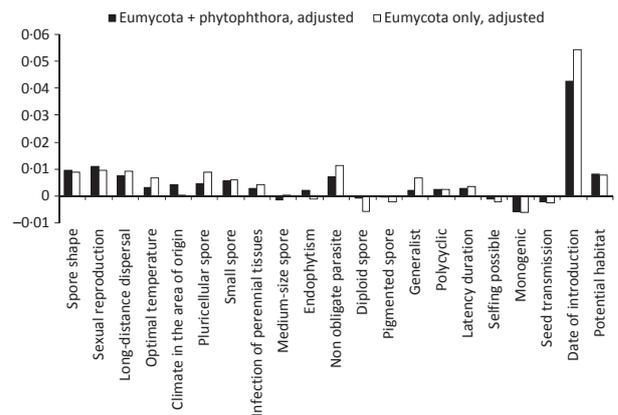

Fig. 5. Importance of the variables for the classification of just the invasive species, as assessed by their mean decrease accuracy (MDA), calculated with Random Forest. Data sets including Eumycota + Phytophthora or Eumycota species were considered successively, with two confounding variables included in the analysis (date of introduction and potential habitat). Variables were sorted as in Fig. 2.

or non-invasive: selfing, monogenic, seed transmission, diploid spore and pigmented spore. The results obtained were consistent for the various data sets, both with and without *Phytophthora* spp., indicating that the same variables were important for differentiating between invasive and non-invasive fungi *sensu stricto* and *sensu lato* (Spearman's rank correlation of MDA values: $r = 0.86$; $P < 2.2 \times 10^{-16}$). The introduction of the two confounding variables into the statistical analysis decreased the MDA values of the biological traits, but the ranks of the traits obtained with and without confounding variables were significantly correlated ($r = 0.90$, $P < 2.2 \times 10^{-16}$ for Eumycota + *Phytophthora*, $r = 0.59$, $P = 0.01$ for Eumycota only).

Discussion

Prediction of the invasiveness of pathogenic fungi is challenging. In addition to the difficulties highlighted for other taxa (Hayes & Barry 2008), fungi have been less extensively studied and are less visible and most introductions were unintentional (Desprez-Loustau *et al.* 2007). The data available for statistical analyses are therefore fewer in number and less accurate. The Random Forest method has proved to be an efficient and reliable classifier, highly suitable for data sets such as ours, in which data are reported for a small number of species but a large number of variables, with some missing data (Cutler *et al.* 2007; Stohlgren *et al.* 2010). Using this method, we showed that certain species traits were useful predictors of invasion success in forest pathogenic fungi alien to Europe.

DATA-RELATED ISSUES

One of the limitations of our study was the small number of species included in the data sets. The use of native species as a control group would have provided a larger number of species, but we thought it inappropriate here. Many comparisons of invasive and non-invasive species have been performed in invasion ecology, but the principal objective in most of these studies was to investigate how the introduced species out-competed the resident species. Here, the determinant factor in the invasion of pathogenic fungi was assumed to be the parasitic behaviour of these species on trees rather than competition with indigenous species at the same trophic level.

Comparisons between established and non-established introduced species appeared more pertinent, but they may be subject to biases in reporting and information quality for the two groups (Keller & Drake 2009; Hayes & Barry 2008). Only in the last few decades have introduced non-established species been listed as interceptions during health inspections at European borders. Furthermore, trait data are usually more readily available for well-known invasive species.

The date of first observation in Europe, identified from accessible publications and data bases, was the best estimate of residence time we could find. Records for most fungal species, particularly for micromycetes, are too sparse for a statistical analysis of temporal trends of invasion and a more accurate estimation of the date of introduction and lag phase, as

proposed for plants (Aikio, Duncan & Hulme 2010). Use of the date of first observation probably results in an underestimation of residence time, because many fungal species may remain inconspicuous and undetected for years, particularly if they are non-invasive.

THE IMPORTANCE OF RESIDENCE TIME AND OTHER CONFOUNDING VARIABLES IN THE PREDICTION OF INVASION STATUS AND THE IDENTIFICATION OF TRAIT PREDICTORS

The key role of residence time in determining invasion success has been highlighted in previous studies and for other organisms (Wilson *et al.* 2007; Desprez-Loustau *et al.* 2010) and was confirmed in our analysis. The date of first observation was the variable with the greatest influence on classification accuracy in all analyses in which it was included. Recently introduced species that are currently assumed to be non-invasive may become invasive in the future. However, residence time did not appear to be a source of bias for the identification of other predictors. Indeed, the inclusion of a proxy for residence time in RF analyses did not eliminate the effects of trait-related variables and had little effect on the rankings of these variables, providing evidence of the reliability of the traits identified in our study. Furthermore, the rates of correct species classification were not increased by adding the confounding variables to the analysis. The performances of RF decision rules derived with and without these variables were very similar and were sometimes better without the confounding variables.

Most fungal clades contained similar numbers of invasive and non-invasive species in our data sets, and the output of analyses was similar for data sets restricted to Eumycota or extended to include *Phytophthora* spp. Most results in our study may therefore apply to different phylogenetic backgrounds for fungi *sensu lato*.

TRAIT PREDICTORS OF INVASION SUCCESS

Classification rules based on species traits systematically outperformed random decision rules, providing a clear demonstration that some traits are useful predictors of invasion success in forest pathogenic fungi alien to Europe. RF analyses made it possible to establish a hierarchy in the importance of these traits. The three most consistently influential variables referred to attributes favouring the ability to spread. This is particular true for the existence of a mechanism of long-distance dispersal mediated by wind, running water or insect vectors rather than rainsplash alone. Interestingly, spore shape, which was a less obvious predictor, also consistently obtained a high MDA. Aerial pathogens are overrepresented among invasive plant pathogenic fungi (Desprez-Loustau *et al.* 2010), therefore also in our data set. As for any particle, spore size and shape affect release, transport and deposition, particularly for aerial pathogens. We used a very simple shape descriptor for which a threshold value corresponding to the optimal drag-minimizing shape for spores ejected in the air could be defined (Roper *et al.* 2008). Our results suggest that

the length-to-width ratio of spores is a potentially useful proxy for fungal dispersal that is both relevant and readily available. Sexual reproduction that was among the most influential variables also has a role in survival and dispersal (Table 2). The importance of dispersal-related traits in our models indicates that natural dispersion after introduction into new areas plays a key role in the invasion process, although human activities are an important factor in the unintentional movement of pathogens. Sexual reproduction may also affect invasion success by generating more virulent variants. This may be of key importance by enabling pathogenic species to adapt to newly encountered host species following their introduction into a new geographic area (McDonald & Linde 2002; Parker & Gilbert 2004). By contrast, traits mostly relating to asexual reproduction, such as the duration of the cycle and the number of sporulation cycles, played a less important role. They did not appear to be important variables for discrimination between invasive and non-invasive species and were of only borderline utility for the prediction of invasive species.

Other traits related to parasitic strategy were among the variables with a positive MDA, such as host range for Eumycota. Generalist species are thought to be better invaders than specialists because they are more likely to find appropriate resources in their new environment (Vall-Ilosera & Sol 2009). However, in the case of parasites, host specialization may not necessarily hinder success in new environments (Parker & Gilbert 2004; Gilbert & Webb 2007). Many successful invaders among forest pathogenic fungi, such as *Cryphonectria parasitica*, are specialist parasites that were able to make a host switch when introduced into new areas (Slippers, Stenlid & Wingfield 2005; Gilbert & Webb 2007). The decrease in MDA for generalist when *Phytophthora* spp were added might indicate a different behaviour: in contrast to fungi, most invasive *Phytophthora* spp in our data set were generalists.

Finally, optimum temperature and climate of origin were among the most important predictors in our analyses. This result highlights the importance of climate-matching, previously identified as the only characteristic consistently associated with invasive behaviour in both animals and plants (Hayes & Barry 2008).

MANAGEMENT APPLICATIONS

The management of forest diseases in general, and of those caused by exotic pathogens in particular, is based principally on preventive measures. It is often difficult to apply eradication measures once a pathogen has been introduced, and such methods are rarely successful (Desprez-Loustau 2009). This may be because the populations of invasive species have already reached intractable levels by the time the first conspicuous symptoms are observed. Identification of the traits characterizing invaders has been proposed as an effective method for targeting efforts to the most likely problematic species, using risk assessment approaches (Keller, Lodge & Finnoff 2007; Keller & Drake 2009). Pest risk assessment (PRA) schemes have been developed by International Plant Protection Organisations as a way of justifying regulatory measures with potentially negative effects on trade. PRA schemes conform to the International Plant Protection Convention standards (ISPM) recognized by the World Trade Organisation (MacLeod *et al.* 2010). Most PRA schemes make use of structured questionnaires, in which experts are asked to score several items on an ordinal scale. Various combinations of verbal and numerical score ratings are used by the European Plant Protection Organisation (EPPO), the North American Plant Protection Organization (NAPPO) and Biosecurity Australia (BA). Until recently, PRA schemes focused principally on crop plants and economic impact, but ISPM11 was recently revised to include environmental aspects. Our results could be used to improve PRA schemes, particularly for pathogens of non-crop plants, such as forest trees, by supporting, ranking and refining the criteria already in use and suggesting new criteria. For example, question 11 of the EPPO scheme 'Does the organism have intrinsic attributes that indicate that it could cause significant harm to plants' could refer explicitly to long-distance dispersal, sexual reproduction and spore attributes, such as spore shape. Our results also demonstrated the potential utility of methods, such as RF, including interactions between variables. This approach is likely to prove more relevant than sequential trait-based approaches, because different ecological strategies, determined by particular combinations of traits, may be associated with invasion success (Küster *et al.* 2008). CART has already been used in risk assessments for invasive species in several studies (Kolar & Lodge 2002; Keller & Drake 2009; Vall-Ilosera & Sol 2009). Random Forest offers a highly reliable method that is increasingly used for risk prediction in clinical research (Sun 2010), but little as yet in ecology (but see Cutler *et al.* 2007; Stohlgren *et al.* 2010), and never, to our knowledge, in PRA.

Conclusion

Trait-based RF analyses predicted invasive species of forest pathogenic fungi in Europe with a success rate of 73·9–82·1%. This success rate is close to that reported for well-studied groups, such as plants, in which risk assessment has been shown to be of economic benefit (Keller & Drake 2009). Expanding the spatial frame of our study, thus including a wider range of species, such as invasive and non-invasive fungi and *Phytophthora* spp in North America, would make it possible to validate and to generalize our results more widely, at least for temperate areas. Nevertheless, these preliminary results are promising and represent a first step towards the construction of accurate trait-based predictive models of invasive plant pathogens. These models could improve PRA, if combined with pathway analyses determining the likelihood of introduction (Brunel 2009).

Acknowledgements

References

Aikio, S., Duncan, R.P. & Hulme, P.E. (2010) Lag-phases in alien plant invasions: separating the facts from the artefacts. *Oikos*, **119**, 370–378.

Brasier, C.M. (2008) The biosecurity threat to the UK and global environment from international trade in plants. *Plant Pathology*, **57**, 792–808.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 15–32.

Breiman, L. & Cutler, A. (2003) Setting up, using, and understanding Random Forests v4.0. http://www.stat.berkeley.edu/~breiman/RandomForests/.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Chapman and Hall, London.

Brunel, S. (2009) Pathway analysis: aquatic plants imported in 10 EPPO countries. *EPPO Bulletin*, **39**, 201–213.

Cutler, D., Edwards, T., Beard, K., Cutler, A., Hess, K., Gibson, J. & Lawler, J. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Davis, M.A. (2003) Biotic globalization: does competition from introduced species threaten biodiversity? *BioScience*, **53**, 481–489.

Desprez-Loustau, M.-L. (2009) The alien fungi of Europe. *Handbook of Alien Species in Europe (DAISIE)*, pp. 15–28. Springer, Dordrecht.

Desprez-Loustau, M.-L., Robin, C., Buée, M., Courtecuisse, R., Garbaye, J., Suffert, F., Sache, I. & Rizzo, D. (2007) The fungal dimension of biological invasions. *Trends in Ecology and Evolution*, **22**, 472–480.

Desprez-Loustau, M.-L., Courtecuisse, R., Robin, C., Husson, C., Moreau, P.-A., Blancard, D., Selosse, M.-A., Lung-Escarmant, B., Piou, D. & Sache, I. (2010) Species diversity and drivers of spread of alien fungi (*sensu lato*) in Europe with a particular focus on France. *Biological Invasions*, **12**, 157–172.

Durrell, L.W. (1964) The composition and structure of walls of dark fungus spores. *Mycopathologia*, **23**, 339–345.

Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Fraser, J.A., Giles, S.S., Wenink, E.C., Geunes-Boyer, S.G., Wright, J.R., Diezmann, S. *et al.* (2005) Same-sex mating and the origin of the Vancouver Island *Cryptococcus gattii* outbreak. *Nature*, **437**, 1360–1364.

Gasso, N., Sol, D., Pino, J., Dana, E.D., Lloret, F., Sanz-Elorza, M., Sobrino, E. & Vila, M. (2009) Exploring species attributes and site characteristics to assess plant invasions in Spain. *Diversity and Distributions*, **15**, 50–58.

Ghattas, B. (2000) Agrégation d'arbres de classification. *Revue de Statistique Appliquée*, **48**, 85–98.

Gilbert, G.S. & Webb, C.O. (2007) Phylogenetic signal in plant pathogen-host range. *Proceedings of the National Academy of Sciences*, **104**, 4979–4983.

Hayes, K. & Barry, S. (2008) Are they any consistent predictors of invasion success? *Biological Invasions*, **10**, 483–506.

Hulme, P.E. (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, **46**, 10–18.

James, T.Y., Kauff, F., Schoch, C.L., Matheny, P. B., Hofstetter, V., Cox, C. *et al.* (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, **443**, 818–822.

Keller, R.P. & Drake, J.M. (2009) Trait based risk assessment for invasive species. *Bioeconomics of Invasive Species: Integrating Ecology, Economics, Policy and Management* (eds R.P. Keller, D.M. Lodge, M.A. Lewis & J.F. Shogren), pp. 44–62. Oxford University Press, New York.

Keller, R.P., Lodge, D.M. & Finnoff, D.C. (2007) Risk assessment for invasive species produces net bioeconomic benefits. *Proceedings of the National Academy of Sciences*, **104**, 203–207.

Köbler, R. & Seufert, G. (2001) Novel maps for tree species in Europe. *Proceedings of the European Symposium on the Physico-Chemical Behaviour of Air Pollutants: "A Changing Atmosphere"*, Torino (Italy) 17–20 September 2001.

Kolar, C.S. & Lodge, D.M. (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology and Evolution*, **19**, 199–204.

Kolar, C.S. & Lodge, D.M. (2002) Ecological predictions and risk assessment for alien fishes in North America. *Science*, **298**, 1233–1236.

Küster, E.C., Kühn, I., Bruelheide, H. & Klotz, S. (2008) Trait interactions help explain plant invasion success in the German flora. *Journal of Ecology*, **96**, 860–868.

Lanier, L., Joly, P., Bondoux, P. & Bellemère, A. (1976) *Mycologie et Pathologie Forestière*. Masson, Paris.

Levine, J.M. & D'Antonio, C.M. (2003) Forecasting biological invasions with increasing international trade. *Conservation Biology*, **17**, 322–326.

Liaw, A. & Wiener, M. (2002) Classification and regression by randomforest. *R News*, **2**, 18–22.

Lloret, F., Médail, F., Brundu, G., Camarda, I., Moragues, E., Rita, J., Lambdon, P. & Hulme, P.E. (2005) Species attributes and invasion success by alien plants in Mediterranean islands. *Journal of Ecology*, **93**, 512–520.

MacLeod, A., Pautasso, M., Jeger, M.J. & Haines-Young, R. (2010) Evolution of the international regulation of plant pests and challenges for future plant health. *Food Security*, **2**, 49–70.

McDonald, B.A. & Linde, C.C. (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annual Review of Phytopathology*, **40**, 349–379.

Parker, I.M. & Gilbert, G.S. (2004) The evolutionary ecology of novel plant–pathogen interactions. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 675–700.

Prasad, A., Iverson, L. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.

Pyšek, P., Křivánek, M. & Jarošík, V. (2009) Planting intensity, residence time, and species traits determine invasion success of alien woody species. *Ecology*, **90**, 2734–2744.

Pyšek, P. & Richardson, D.M. (2007) Traits associated with invasiveness in alien plants: where do we stand? *Biological Invasions* (ed. W. Nentwig), pp. 97–126. Springer-Verlag, Berlin and Heidelberg.

Quintanilla, L.G. & Escudero, A. (2006) Spore fitness components do not differ between diploid and allotetraploid species of *Dryopteris* (Dryopteridaceae). *Annals of Botany*, **98**, 609–618.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. Version 2.9.1. R Foundation for Statistical Computing, Vienna, Austria.

Roper, M., Pepper, R.E., Brenner, M.P. & Pringle, A. (2008) Explosively launched spores of ascomycete fungi have drag-minimizing shapes. *Proceedings of the National Academy of Science USA*, **105**, 20583–20588.

Sache, I., Roy, A.-S., Suffert, F. & Desprez-Loustau, M.-L. (2011) Invasive Plant Pathogens in Europe. *Biological Invasions: Economic and Environmental Costs of Alien Plant, Animal, and Microbe Species*, 2nd edn (ed. D. Pimentel), pp. 227–242. CRC Press, Boca Raton.

Sakai, A.K., Allendorf, F.W., Holt, J.S., Lodge, D.M., Molofsky, J., With, K.A., Baughman, S., Cabin, R.J., Cohen, J.E., Ellstrand, N.C., McCauley, D.E., O'Neil, P., Parker, I.M. and Thompson, J.N. (2001) The Population Biology of Invasive Species. *Annual Review of Ecology and Systematics*, **32**, 305–332.

Shatele, M.S. (2009) Effects of gamma irradiation on fungal growth and associated pathogens. *Research Journal of Environmental Toxicology*, **3**, 94–100.

Sinclair, W.A., Lyon, H.H. & Johnson, W.T. (2005) *Diseases of Trees and Shrubs*. Cornell University Press, Ithaca, New York, USA.

Slippers, B., Stenlid, J. & Wingfield, M.J. (2005) Emerging pathogens: fungal host jumps following anthropogenic introduction. *Trends in Ecology and Evolution*, **20**, 420–421.

Stockmarr, A., Andreasen, V. & Østergård, H. (2007) Dispersal distances for airborne spores based on deposition rates and stochastic modeling. *Phytopathology*, **97**, 1325–1330.

Stohlgren, T.J. & Schnase, J.L. (2006) Risk analysis for biological hazards: what we need to know about invasive species. *Risk Analysis*, **26**, 163–173.

Stohlgren, T.J., Ma, P., Kumar, S., Rocca, M., Morisette, J.T., Jarnevich, C.S. & Benson, N. (2010) Ensemble habitat mapping of invasive plant species. *Risk Analysis*, **30**, 224–235.

Sun, Y.V. (2010) Multigenic modeling of complex disease by random forests. *Advances in Genetics*, **72**, 73–99.

Therneau, T.M. & Atkinson, E.J. (1997) An introduction to recursive partitioning using the rpart routine. *Technical Report*, 61, Mayo Clinic, Section of Statistics, Rochester.

Vall-Ilosera, M. & Sol, D. (2009) A global risk assessment for the success of bird introductions. *Journal of Applied Ecology*, **46**, 787–795.

Viennot-Bourgin, G. (1949) *Les Champignons Parasites des Plantes Cultivées*. Masson, Paris.

Vilà, M., Basnou, C., Pysek, P., Josefsson, M., Genovesi, P., Gollasch, S., Nentwig, W., Olenin, S., Roques, A., Roy, D., Hulme, P.E. & DAISIE partners (2010) How well do we understand the impacts of alien species on ecosystem services? A pan-European, cross-taxa assessment. *Frontiers in Ecology and Environment*, **8**, 135–144.

Williamson, M. & Fitter, A. (1996) The varying success of invaders. *Ecology*, **77**, 1661–1665.

Wilson, J.R.U., Richardson, D.M., Rouget, M., Proches, S., Amis, M.A., Henderson, L. & Thuillier, W. (2007) Residence time and potential range: crucial considerations in modelling plant invasions. *Diversity and Distributions*, **131**, 11–22.