Predicting future invaders and future invasions

Alice Fournier, Caterina Penone, Maria Grazia Pennino, and Franck Courchamp

Invasive alien species are a great threat to biodiversity and human livelihoods worldwide. The most effective way to limit their impacts and costs is to prevent their introduction into new areas. Identifying invaders and invasions before their occurrence would arguably be the most efficient strategy. Here, we provide a profiling method to predict which species—with which particular ecological characteristics—will invade, and where they could invade. We illustrate our approach with ants, which are among the most detrimental invasive species, as they are responsible for declines of numerous taxa, are involved in local extinctions, disturb ecosystem functioning, and impact multiple human activities. Based on statistical profiling of 1,002 ant species from an extensive trait database, we identify 13 native ant species with an ecological profile that matches that of known invasive ants. Even though they are not currently described as such, these species are likely to become the next global invaders. We couple these predictions with species distribution models to identify the regions most at risk from the invasion of these species: Florida and Central America, Brazil, Central Africa and Madagascar, Southeast Asia, Papua New Guinea, Northeast Australia, and many islands worldwide. This framework, applicable to any other taxa, represents a remarkable opportunity to implement timely and specifically shaped proactive management strategies against biological invasions.

Significance

Biological invasions are the second most severe threat affecting biodiversity worldwide with major economic and societal impacts. Yet, the multitude of species with the potential to be invasive makes it extremely difficult to anticipate invasions to develop efficient management plans. Predicting which species are likely to become invaders and where they are likely to invade even before their introduction outside their native range has always been a prime objective of invasion biology. Based on the example of ants, we provide a profiling method to predict future invaders and future invasions, and in this manner, we identify 13 ant species likely to become new invaders and regions at risk for these invasions.
Conservation of Nature (IUCN) list of the “100 of the world’s worst invasive alien species,” a feat equaled by no other taxonomic family (48). In this study, we will follow the definition of invasive species adopted by the Invasive Species Specialist Group of the IUCN, excluding species that are native but expanding, as well as exotic species with no impact.

Here, we have built a profiling tool that identifies which species have the ecological profile that makes them highly likely of being the next invading species. Our approach is to first define an ecological profile of the most invasive ant species and then compare it with the ecological profile of other species based on an extensive dataset of ant species traits, imputed through finely tuned and up-to-date machine-learning algorithms. Our quantitative risk assessment tool responds to two of the most crucial questions of invasion biology today: which species are likely to become invasive and where will they become invasive? This methodology is transposable to other taxa for which there is adequate ecological information on current invaders.

Results

Future Invaders. Our pool of ant species retained for the ecological profiling of invasive species was kept high on account of the setup imputation procedure, which enabled us to estimate missing values in the traits and thus have all trait values, either observed or estimated, for 1,002 species. This was done while keeping the imputation error very low (whole dataset imputation error: 5.97%), thus ensuring robust model estimates and predictions. We used this imputed dataset of four ecological traits to calculate the probability of species being invasive and detect ant species with a profile highly similar to known ant invaders. Our models enabled us to efficiently discriminate between likely and unlikely invasive species [Area Under the receiver operating characteristic Curve (AUC) = 0.95], and it indicated relatively strong relationships between species traits and invasion potential (pseudoR² = 0.53). Our analyses suggest that the four most important traits for predicting ant invasiveness probability are the following: independent colony foundation mode, supercoloniality, generalist nesting type, and a frequent association with disturbances (SI Appendix, Table S3).

Our method also allowed us to distinguish two invasiveness profiles. Species with a probability > 0.70 (both known and predicted invasive species) are hereafter called “superinvasive.” From the 19 known invasive species, 14 have a superinvasive profile, with an invasiveness probability between 0.859 and 0.700 (mean: 0.837, Fig. 1). Four known invasive species have probabilities between 0.132 and 0.115 (mean: 0.157) and are hereafter called “invasive.” This bimodal pattern is likely linked to the colony structure of the species, as all superinvasive species are also supercolonialist while all invasive species are not. Finally, one species recorded by the IUCN as invasive, *Aenomymex octospinosus*, has a significantly distinct profile, with an extremely low invasiveness probability of 0.155 × 1⁻³⁰⁸. Our model also provided us with a list of 13 species that share the trait profile of current invasive species. Among the 13 species, 2 have a superinvasive profile (invasiveness probability between 0.828 and 0.838; mean: 0.830), while 11 have an invasive profile [probabilities from 0.111 to 0.267; mean: 0.163 (Table 1)]. Note that a leave-one-out sensitivity analysis [recoding known invasive species as noninvasive (SI Appendix)] classified all known invasive species as invasive in 100% of our models, thus showing that our models were reliable for predicting invasiveness (SI Appendix, Tables S4 and S5). The remainder of the 970 species that were not identified as invasive (according to the lower fifth centile of the invasive species probability distribution) in more than 90 models of 100 were not considered to be potential future invaders. Targeted studies on some of these species may, however, be relevant to confirm that they are at risk for becoming invaders in the future.

Areas at Risk. Species distribution models allowed us to identify regions of climatic suitability for each of the 11 likely future invaders for which we had sufficient data (SI Appendix, Fig. S11). These predictions were of high quality: True Skill Statistics (TSS) values were between 0.88 and 0.97 and AUC values were between 0.96 and 0.99, indicating good model performances.

The potential invasions are more likely in tropical regions, as illustrated in the cumulated map (Fig. 2). When overlapping all potential distributions, invasion hotspots (i.e., areas with more than eight invasive species) were predicted to occur in Florida and Central America, Brazil, Central Africa and Madagascar, Southeast Asia, Papua New Guinea, Northeast Australia, and many islands worldwide. Areas characterized by temperate and extreme climatic conditions (i.e., ice, hot desert, tundra), especially in the Paleartic and Nearctic (SI Appendix, Fig. S1) seemed less at risk from these novel ant invasive species. This was also the pattern observed for the two identified potential superinvasive species (Fig. 2), although *Lepisiota canescens* could also invade almost all of Africa and South East Asia, and *Technomyrmex difficilis* had a much more restricted potential distribution.

Discussion

Using a large database on ant species traits and an up-to-date phylogeny of ant genera, we developed a robust methodology to quantify the relative importance of ecological traits in shaping their invasiveness probability and used it to both characterize the ecological profile of current invasive ants and identify 13 noninvasive species that share this profile. We distinguished between invasive and highly invasive species based on their probability of invasiveness. We complemented this prediction by identifying regions most at risk for invasion from these species: tropical regions, both mainland and islands. Our predictive approach directly points to species that have not yet invaded and areas that are not yet invaded, but are likely to do so in the future.

Pioneer risk-assessment models are all based on datasets that only comprise information on exotic ant species, i.e., species that have already been moved and detected outside of their native zone (24, 29), or pest species, i.e., species that are already known to have concerning impacts (49, 50). As a result, these models cannot point to novel species of potential risk. In contrast, our screening tool is based on a large and collaborative dataset that contains information for 25 ecological traits for the 19 ant species recognized as invasive by the IUCN (47) as well as over 2,000 native species, of which 1,002 had all information (observed or
inferred) for the traits selected in our model. The fact that this dataset comprises species from 47% of the ant family originating from all of the world’s ecoregions (SI Appendix, Fig. S1) makes our predictive model genuinely compatible with the anticipation of future invasions worldwide.

Unlike most current trait-based predictive models, our methodology includes a robust and accurate way of dealing with missing values, which are commonplace in life-history trait datasets in general. Indeed, the removal of cases with missing values is known to result in reduced statistical power and biased parameter estimates (31, 33, 51). We overcame this issue by bringing together the most recent and accurate imputation algorithms (52, 53) and finely calibrating them to fit our data. We comprehensively examined the performance of tuning parameters for multiple imputation and included the most relevant number of phylogenetic eigenvectors to impute each trait. The crucial issue of how many phylogenetic eigenvectors should be included for trait imputation, which we carefully consider in the present paper, has been rarely discussed in the literature (54). Our analysis, in accordance with previous work, also revealed the existence of some patterns of phylogenetic signal above the species level (SI Appendix, Table S4), supporting the fact that our genus phylogeny was a useful predictor to include when imputing the ant species dataset (20).

After imputing the dataset multiple times, the multimodel inference enabled us to establish strong relationships between species traits and invasion potential and to efficiently discriminate between likely and unlikely invasive species.

Our results highlighted several traits of importance that were previously highlighted as commonly found in invasive ant species: supercoloniality, generalist nesting type, affinity for disturbed environments, and expansion through independent colony founding (37, 55). Moreover, our study goes beyond trait analysis by using it to calculate species invasiveness probability and applying it to a large set of species not yet known to be invasive. This led us to distinguish two invasive profiles—superinvasive and invasive—and detect two potential superinvasive species (L. canescens and T. difficilis) and 11 potential invasive species (Table 1). This result seems in accordance with information available from experts who describe these two superinvasive species as having the characteristics to become future invaders and impact native ecosystems (56–58). However, the 11 other species have, to our knowledge, never been described as problematic, which highlights the proactive potential of our method. There are in fact relatively few studies available on most of them, emphasizing the need to further study and monitor them. Although they have not yet been categorized as invasive, all 13 predicted invaders should be carefully surveyed at points of entry in at-risk regions and become the target of serious monitoring programs. In this regard, our cumulated map highlights the mismatch between the large regions most at risk and those most likely to put proactive measures into place (59).

The fact that one of the known invasive species, A. octospinosus, showed a radically distinct profile from the other invasive species illustrates an additional strength of our methodology: in addition to detecting new invaders, it can also question the established invasive status of certain species. Despite being on the IUCN list of invasive ants (47), A. octospinosus is characterized as invasive by only one scientific paper (60), yet without providing evidence. This leaf-cutting ant species has the potential to destroy agricultural crops and has been observed to damage the fern Cyathea arborea on one island (Guadeloupe) in its native ecosystem. However, its traits (strict monogyny, highly aggressive attitude, independent colony foundation, strict mutualist relationship with fungus species, and long generation time) allow it to slowly spread in Guadeloupe (0.5 km/year), its ability to be transported inadvertently by humans (55), and the climatic and geographic proximity between its native zone and the only island where it has been introduced seem to demonstrate that this species is not invasive, but rather an agricultural pest species that is still currently filling its suitable niche. This species is, however, known to have major impacts on agriculture (32), which may explain why it was proposed and accepted for the IUCN list of invasive species. We suggest that, in the absence of further evidence, its presence on the IUCN list should be questioned, as it does not correspond to the ecological definition of an invasive alien species but rather to that of an agricultural pest. We nevertheless recognize that ant species with a high potential to become invasive may exist with a profile very different from known invasive species. These potential invaders would be detected neither by our model nor by any trait-based model calibrated with invasive species characteristics, which illustrates the complex and idiosyncratic nature of biological invasions.

Our species distribution models allowed us to map the world regions with the most suitable climatic conditions for the two predicted superinvasive species and 8 of the 11 predicted invasive species. These maps showed that future invasions by these predicted invaders are more likely in tropical regions, most of which are at risk for a large number of these species. Tropical islands, already invaded or at risk for invasion by many of the 19 current invasive species, are also generally at high risk of invasion by these potential invaders, emphasizing the importance

Table 1. Predicted invasiveness probabilities, or “invasion profiles,” of 19 invasive species from the IUCN red list (in boldface) and 13 potential future invaders identified with our model

<table>
<thead>
<tr>
<th>Species</th>
<th>P ±</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Superinvasive profiles</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anoplolepis gracilipes</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Linepithema humile</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Monomorium pharaonis</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Myrmica rubra</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Nylanderia pubens</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Paratrechina longicornis</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Tapinoma melaneopelatum</td>
<td>0.86 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Lasius neglectus</td>
<td>0.83 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Solenopsis geminata</td>
<td>0.83 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>0.83 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Technomyrmex albipes</td>
<td>0.83 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Trichomyrmex destructor</td>
<td>0.83 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>L. canescens</td>
<td>0.83 ± 0.12</td>
<td>100</td>
</tr>
<tr>
<td>T. difficilis</td>
<td>0.83 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Pheidole megacephala</td>
<td>0.70 ± 0.05</td>
<td>100</td>
</tr>
<tr>
<td><strong>Invasive profiles</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F. yessensis</td>
<td>0.27 ± 0.03</td>
<td>100</td>
</tr>
<tr>
<td>T. bicaudatum</td>
<td>0.23 ± 0.21</td>
<td>100</td>
</tr>
<tr>
<td>A. spinosa</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Cardiocondylyla emeryi</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Cardiocondylyla minutior</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Cardiocondylyla wroghtoni</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Neivamyrmex pilosus</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Tetramorium similimum</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Monomorium florica</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Solenopsis pappaua</td>
<td>0.16 ± 0.02</td>
<td>100</td>
</tr>
<tr>
<td>Brachyponera chinensis</td>
<td>0.13 ± 0.04</td>
<td>100</td>
</tr>
<tr>
<td>Solenopsis richteri</td>
<td>0.13 ± 0.04</td>
<td>100</td>
</tr>
<tr>
<td>A. trigona</td>
<td>0.13 ± 0.04</td>
<td>99</td>
</tr>
<tr>
<td>T. laevii</td>
<td>0.13 ± 0.03</td>
<td>93</td>
</tr>
<tr>
<td>T. lujae</td>
<td>0.12 ± 0.02</td>
<td>90</td>
</tr>
<tr>
<td>A. octospinosus</td>
<td>0.00 ± 0.00</td>
<td>100</td>
</tr>
</tbody>
</table>

**P**: invasiveness probability (mean: 100 models); "±": invasiveness variability (SD: 100 models); "%": percentage of models. Note that the values of A. octospinosus are P = 0.0001 ± 2.60671E-05.
of biosecurity in these biodiversity-rich ecosystems. It is worth mentioning here that, although occurrence data used for Species Distribution Models (SDMs) were thoroughly checked for issues typical of such data (SI Appendix, Supplementary Material and Methods), some spatiotemporal bias due to the lack of knowledge of the sampling effort remains unavoidable (61). Therefore, the resulting distribution maps should, as usual, be considered with caution.

Our list of predicted invaders crossed with climatic niche modeling offers a worldwide risk assessment of the areas that are the most at risk and informs the relevant countries about the necessity to pay attention to the introduction of ants to prevent invasions. This should be viewed together with the corresponding information and all traits initially downloaded from Antprofiler. We used phylogenetic databases such as plants, fish, and terrestrial vertebrates.

Materials and Methods
To simultaneously identify the most relevant traits and keep the highest number of species, we proceeded in two steps. First, we identified and selected traits that performed well at predicting invasiveness. Second, we optimized our usable dataset by removing species with too many missing values and by estimating the values of the selected traits for the species with few missing values (using all available traits). We then used the dataset with the selected traits, either observed or estimated, to predict invasiveness probability with Generalized Linear Models. Finally, we predicted areas at risk for invasion of species identified as potential invaders, using Species Distribution Models. Our step-by-step process is illustrated in Fig. 3 and is presented below, with further details in SI Appendix.

Trait Data Selection and Imputation. We searched for trait combinations that are associated with invasive species. We extracted ant species traits from our AntProfiler online database (https://antprofiler.ese.u-psud.fr). This database includes information on 2,176 ant species (134 genera) and ecological traits related to their occurrence, morphology, behavior, and invasive status (66). From this database, we selected traits noncorrelated to each other but correlated to the invasive status. The full description of the trait selection process is given in SI Appendix, Supplementary Material and Methods and illustrated in SI Appendix, Fig. S3.

Despite our data collection efforts, the database contained many missing values. We therefore used an imputation technique (missForest) to estimate the missing values (53). Before imputation, we removed the species with too many missing values to keep a dataset with a maximum of ~60% of missing information (following ref. 52). We then explored how missing data could be input into our database to minimize bias and maximize prediction accuracy. To improve the estimation of missing values, we used phylogenetic information and all traits initially downloaded from Antprofiler. We used out-of-bag error estimations to calibrate the models as these were shown to accurately estimate imputation error, and we repeated the imputation 100 times (Fig. 3 and SI Appendix, Supplementary Material and Methods).

We selected a final subset of four traits: colony foundation, colony structure, nesting type, and association with disturbed areas and 1,002 species (of 2,176).

Predictive Model Building.
Potential future invasive species. We modeled the invasive profile using generalized linear models with our four selected traits as predictors and a binomial distribution. We ran 100 models for each of the previously imputed datasets and used model predictions to determine future invasive species. For each model, we identified potentially new invasive ant species as those with a...
predicted invasiveness probability above the lower fifth centile of the 19 known invasive species probability distribution. Each identified species was considered to be potentially invasive if it was selected in at least 90 of the 100 models.

**Areas at risk.** SDMs are based on correlations between environmental variables and geolocalized species records and can be used to delineate potential species distributions (67). We built SDMs for the ant species found to have similar profiles to those of known invasive species to identify the areas that present suitable environmental conditions for these species and thus areas at risk for invasion. A full description of the SDM process is given in SI Appendix, Supplementary Material and Methods. Two metrics were used to evaluate SDM accuracy: the TSS and the AUC.

Ensemble models—i.e., averages of all models used, weighted by the TSS—were run for each of the predicted invaders with sufficient occurrence points to produce individual climatic suitability distribution maps (SI Appendix, Fig. S11). It is noteworthy that three species (Formica yessensis, Aphaenogaster spinosa, and Azteca trigona) had very few occurrence points (24, 33, and 42, respectively), meaning that the resulting distribution maps should be taken with additional caution. For two other species (Technomyrmex laurenti and Technomonymex lujei), the number of occurrences was too low (<20 observations) to ensure reliable results; they were subsequently omitted from the analysis. Finally, we also combined these individual predictions by summing each species potential distribution as binary maps to obtain a cumulative invasion risk map from these future invaders. Binary transformation was based on the threshold that maximized the TSS for each species.

**Data Availability.** All data and scripts used in this study are available at https://github.com/caterinap/Antprofiler.

**ACKNOWLEDGMENTS.** We thank Cleo Bertelsmeier and Sebastien Ollier with whom we worked on different statistical approaches for a preliminary study and who provided useful comments on an earlier version of the manuscript; five anonymous reviewers who provided useful comments that led to improved analyses; Guillaume Guénard and Daniel J. Stekhoven for their advice, which helped improve our imputation methodology; Benoît Guénard and James Wetterer for sharing their ant occurrence data; and Corrie Moreau and Philip Ward who provided information and data on ant phylogeny.


