ECOLOGICAL DETERMINANTS OF PLANT INVASION SUCCESS: INTEGRATING MULTIVARIABLE STATISTICAL INFERENCE AND MACHINE LEARNING PREDICTION

[1]Dr. Monali U. Ghurde

[1]Professor, Vidya Bharati Mahavidyalaya Amravati, Maharashtra, Email ID: monali.ghurde25@gmail.com

Abstract

Ecological, biogeographical, functional and phylogenetic processes interact in complex ways to drive biological invasions. The identification of strong invasion success determinants and assessment of predictive modelling techniques are critically needed to enhance ecological risk assessment and management. This study investigated the principal ecological determinants of plant invasion success and compared the predictive performance of multivariable logistic regression and Random Forest classification. A quantitative observational analysis was conducted using 735 cultivated naturalized plant taxa, comprising 300 invasive and 435 non-invasive taxa. Group differences were evaluated using Mann–Whitney U, chi-square, or Fisher's exact tests with Benjamini–Hochberg adjustment. Spearman correlation analysis was followed by multivariable logistic regression and a tuned Random Forest model, with predictive performance assessed using accuracy, sensitivity, specificity, precision, F1 score, and ROC-AUC on an independent test set. Invasive taxa exhibited broader native ranges, greater climatic suitability, and lower phylogenetic distances than non-invasive taxa. Logistic regression identified native-range size, climatic suitability, human food use, and poisonous classification as positive predictors, whereas vegetative propagation and minimum phylogenetic distance were negatively associated with invasion success. Logistic regression achieved higher accuracy (0.735), F1 score (0.655), and ROC–AUC (0.758) than Random Forest, which demonstrated substantial overfitting despite similar test discrimination. These findings indicate that plant invasion success is best explained by integrating environmental compatibility, biogeographical breadth, human use, reproductive strategy, and phylogenetic relationships. Logistic regression was the most robust and interpretable model for ecological prediction and Random Forests gave complementary information about the importance of predictors.

Keywords: *plant invasion, climatic suitability, native-range size, logistic regression, Random Forest.*

1. Introduction

Biological invasions are now considered one of the most significant causes of environmental change globally, as they can change ecosystem structure, decrease native biodiversity, break down ecosystem interactions and inflict significant economic losses. Plant invasions are a growing focus in ecology and conservation science because the spread of non-native plants beyond their natural habitats is now much faster than ever before, due to their spread in trade, agriculture, horticulture, and by human movement (Pyšek et al., 2020). Successful plant invasions have impacts on species replacement, but also on nutrient cycling, ecosystem productivity, habitat quality and the provision of ecosystem services supporting human welfare (Vilà & Hulme, 2017). Plant invasion is a complicated ecological process, which depends on the interactions between species characteristics, environments, history of introduction and biotic communities. Recent work in invasion ecology indicates that no single mechanism adequately explains invasion success across ecosystems. Rather, it is a result of a series of ecological processes which interact, both preceding and following species establishment, during the invasion (Enders et al., 2020). Knowing how these mechanisms interact is thus crucial for enhancing ecological prediction and for developing successful measures to prevent or manage invasions.

Substantial advances have been made over the past decade on the ecological features of successful invasive plants. The importance of multiple biological traits, propagule pressure, environmental suitability, evolutionary history, and ecosystem interactions is demonstrated by contemporary evidence that invasion success relies not on a single trait, but on a combination of traits. Contemporary evidence suggests that invasion success is associated with combinations of functional traits, propagule pressure, environmental suitability, evolutionary history, and interactions with recipient ecosystems, not just a single trait (Gioria et al., 2023). Likewise, from an ecological and evolutionary perspective, adaptation, phenotypic plasticity and evolutionary changes after introduction are factors in establishing and the spread of alien plant species in heterogeneous environments (van Kleunen et al., 2018). Invariably, environmental change also alters invasion dynamics by creating conditions that are likely to benefit invasive species at the expense of native vegetation. The introduced plants may benefit from altered temperature regimes, changes in precipitation, higher atmospheric carbon dioxide levels and higher disturbance levels, allowing for better competitive performance and spread (Liu et al., 2017). The rise of globalization and rapid ecological change has led to a growing integration of multidisciplinary methods that integrate ecological theory, advances in environmental modelling, phylogenetics, and predictive analytics into the study of plant invasion ecology, which helps explain invasion processes (Liu et al., 2022). Propagule pressure is one of the most well-supported hypotheses of invasion success among those proposed. Establishment probability increases with frequency of introductions and with an increase in the number of individuals introduced. Demographic and environmental stochasticity are reduced by frequent introductions and by the number of individuals introduced. Meta-analysis has shown that propagule pressure is significantly and positively associated with invasion success for a wide range of taxa and ecosystems (Cassey et al., 2018). This idea has since found its way into the literature as a major theoretical model for the effect of repeated introductions on the probability of establishment for equivalent ecological circumstances (Jeschke & Starzer, 2018).

Another important factor that affects the success of introduced species establishing self-sustaining populations is environmental suitability. Species with similar climatic needs are more likely to survive, reproduce and spread geographically when transplanted to regions with similar climate. Future climate change projections have demonstrated that the potential for many areas to be more suitable for invasive species could be further enhanced (Adhikari et al., 2019). Distribution modelling has also shown that environmental suitability is a useful tool to identify areas of potential vulnerability to future plant invasion and prioritize management action (McMahon et al., 2021). The invasion potential is also significantly affected by functional traits, which affect resource acquisition, reproduction, competitive ability, and ecological interaction. The ecological opportunities available to introduced plants depend on their characteristics, including their growth form, their reproductive strategies, how they are dispersed and their physiological flexibility. Functional-response approaches have thus been suggested as integrative approaches for linking species traits and invasion mechanisms across environmental contexts (Dick et al., 2017). Invasive species often alter soil characteristics, nutrient cycles, microbial communities and native biodiversity at the ecosystem level, all of which can help further their establishment and ecological effects (Lone et al., 2019). More recent analytical advances have extended beyond traditional statistical methods to include multivariable modelling and machine learning techniques that are able to process multiple ecological predictors simultaneously. Logistic regression has been widely used, as it allows for interpretable estimates of the independent effect of each predictor, controlling for confounding variables. In contrast, machine learning algorithms have complementary strengths such as being able to detect complex nonlinear relationships and provide rankings of the importance of variables for prediction. Combining these analytical approaches offers the potential for a better understanding of ecosystems and predictions of the effects of climate change, while keeping biological interpretation.

Although there have been significant advances in invasion biology, there are still methodological shortcomings. Most of the previous studies have addressed only a single ecological hypothesis (e.g., propagule pressure, climatic suitability, functional traits or phylogenetic relationships), but not all of them have assessed these factors in a single analytical framework. In recent studies that have used predictive modelling, the accuracy of classification has often been the focus of study, while the independent contribution of each variable to the model has been ignored (which is also hard to interpret). Similarly, there has been limited direct comparison of the strengths and weaknesses of multivariable regression and machine learning methods for ecological prediction, limiting understanding of their complementary strengths. Removing these constraints necessitates an integrated approach that is able to link ecological interpretation to predictive modelling to develop reliable predictors of plant invasion success.

In this context, the purpose of this research was to delve into ecological conditions that influence plant invasion success using multivariable statistical inference and machine learning prediction. Specifically, it looked at the relation between

propagule pressure, suitability, functional traits and phylogenetic traits with invasion status, determined independent ecological predictors through multivariate logistic regression, and compared the predictive ability of logistic regression and Random Forest models to assess their usefulness for classification of invasive plant taxa.

2. Methodology

2.1 Research Design

This study used a quantitative observational research design to determine the ecological, biological and phylogenetic factors that were correlated with the success of invasion of cultivated naturalized plant taxa. Descriptive, inferential and predictive statistics were used for comparing invasive and non-invasive taxa. Multivariable logistic regression and machine learning were used to assess the importance of multiple predictors and their ability to discriminate between invasive and non-invasive taxa, in addition to identifying significant ecological characteristics.

2.2 Data Source

Data used were from the public database made available by Dong et al. (2025). The data consist of ecological, cultivation information, life history information, propagation information, and phylogenetic information of cultivated naturalized plant taxa found in China, and the invasion status of each species is the primary outcome variable. The data set was not modified from the original biological data and served as an adequate basis for studying the factors related to successful plant invasion (Dong et al., 2025).

2.3 Study Variables

Invasion status was used as the dependent variable (binary values of 1 and 0 for the invasive and non-invasive taxa, respectively). Propagule pressure indicators, environmental niche characteristics, species traits and phylogenetic distance measures were used as independent variables and represented four ecological groups. Climatic suitability and size of native range were continuous variables, as were maximum plant height and phylogenetic distance metrics. Ecological uses, native-region indicators, life forms, and propagation strategies were categorical variables.

2.4 Data Preprocessing

Before analyzing the data statistically, the dataset was tested for variable structure, missing observations, duplicate data and coding consistency. Categorical variables were filled with the mode; continuous variables were filled with the median value. Duplicate observations were deleted (where applicable), and binary predictors were validated to maintain the 0/1 coding. To minimize multicollinearity and stability of the model, highly correlated predictors and structurally dependent variables were removed before regression modelling.

2.5 Statistical Analysis

Because several variables were not normally distributed, medians and interquartile ranges (IQRs) were used for summarizing continuous variables and Mann-Whitney U tests were used to compare continuous variables between invasion groups. Pearson's chi-square test was used to make comparisons when the expected cell frequencies were small; otherwise, Fisher's exact test was used. For multiple comparisons, p-values were Benjamini-Hochberg false discovery rate (FDR) corrected. Spearman's rank correlation coefficients were used to assess the associations between ecological variables and to select variables with high correlations before multivariable modelling.

2.6 Multivariable Logistic Regression

To determine ecological predictors that are independent of each other, a multivariable binary logistic regression model was built. Before fitting the model, variables with structural dependence, extreme imbalance and high multicollinearity were removed. The degree of multicollinearity was checked by variance inflation factors and those predictors that met the predetermined cutoff were included. Adjusted odds ratios with 95% confidence intervals (CI) were reported for model coefficients; the model calibration was assessed by the Hosmer–Lemeshow goodness-of-fit test.

2.7 Machine Learning Analysis and Model Evaluation

To complement the regression analysis and assess the predictive performance of the model, a Random Forest (RF) classifier was developed. The hyperparameters were optimized with grid search cross-validation using the area under the receiver operating characteristic curve (ROC-AUC). Performance of the model was evaluated using an independent test set and evaluated with accuracy, sensitivity, specificity, precision, F1 score and ROC-AUC. To determine the relative contribution of individual predictors on model performance, and to select the ecological variables that are most important in the prediction of invasion, permutation importance was calculated.

3. Results

3.1 Characteristics of invasive and non-invasive plant taxa

A total of 735 naturalized plant taxa were analysed, and 300 taxa (40.8%) were classified as invasive and 435 taxa (59.2%) as non-invasive. The main ecological, functional, biogeographical and phylogenetic traits (Table 1) are presented in relation to invasion status. There was a higher median number of invasive taxa in the cultivated provinces compared to non-invasive taxa (5.0 vs. 3.0, respectively; adjusted $p = 0.040$). Invasive taxa also exhibited greater climatic suitability and wider distribution ranges than others. The median climatic suitability of invasive taxa was 0.213, compared with 0.199 for non-invasive taxa, and 31.0 regions for invasive taxa versus 20.0 regions for non-invasive taxa for their corresponding native range sizes. The differences were significant after false-discovery-rate correction (adjusted $p < .001$). Generally, the phylogenetic measures were lower for the invasive taxa. The non-invasive group had a mean phylogenetic

distance of 250.948 while the invasive group had a mean phylogenetic distance of 247.838 (adjusted p = 0.003). Additionally, the mean phylogenetic distances were significantly smaller in invasive taxa than in non-invasive taxa, both with minimum and weighted distances. However, maximum plant height, number of botanical gardens and number of economic uses did not significantly vary after adjustment. There were a number of categorical features which were related to invasion. While 85.0% of invasive taxa and 72.2% of non-invasive taxa are used medicinally, 38.0% and 26.7% are poisonous. There were also more common associations with native regions X7 and X8 for invasive taxa. Propagation by seeds was more common among invasive taxa; vegetative propagation and the combined propagation category were more common among the non-invasive taxa.

Table 1. Selected characteristics of naturalized plant taxa according to invasion status

Variable	Measurement	Overall	Non-invasive	Invasive	Adjusted p value
Number of botanical gardens	Median [IQR]	3.000 [0.000, 10.000]	3.000 [0.000, 11.000]	4.000 [1.000, 10.000]	0.072
Number of cultivated provinces	Median [IQR]	4.000 [1.000, 10.000]	3.000 [1.000, 11.000]	5.000 [2.000, 10.000]	0.040
Number of economic uses	Median [IQR]	3.000 [1.000, 5.000]	3.000 [1.000, 5.000]	3.000 [1.000, 6.000]	0.080
Mean climatic suitability	Median [IQR]	0.200 [0.151, 0.251]	0.199 [0.142, 0.234]	0.213 [0.164, 0.266]	<0.001
Native-range size	Median [IQR]	25.000 [10.000, 42.000]	20.000 [7.000, 39.000]	31.000 [14.000, 48.250]	<0.001
Maximum plant height	Median [IQR]	1.500 [0.900, 3.000]	1.500 [0.800, 4.000]	1.650 [1.000, 3.000]	0.628
Mean phylogenetic distance	Median [IQR]	249.104 [235.806, 253.211]	250.948 [237.179, 253.865]	247.838 [232.586, 252.208]	0.003
Minimum phylogenetic distance	Median [IQR]	20.988 [10.305, 42.893]	23.539 [10.887, 45.622]	18.928 [10.003, 36.167]	0.032
Weighted mean phylogenetic distance	Median [IQR]	46.373 [44.159, 47.087]	46.725 [44.408, 47.088]	46.123 [43.683, 46.969]	0.004
Medicinal use	Present, n (%)	569 (77.4%)	314 (72.2%)	255 (85.0%)	<0.001
Poisons	Present, n (%)	230 (31.3%)	116 (26.7%)	114 (38.0%)	0.005
Native region X2	Present, n (%)	200 (27.2%)	133 (30.6%)	67 (22.3%)	0.048
Native region X4	Present, n (%)	146 (19.9%)	102 (23.4%)	44 (14.7%)	0.014
Native region X7	Present, n (%)	365 (49.7%)	178 (40.9%)	187 (62.3%)	<0.001
Native region X8	Present, n (%)	372 (50.6%)	188 (43.2%)	184 (61.3%)	<0.001
Seed propagation	Present, n (%)	694 (94.4%)	400 (92.0%)	294 (98.0%)	0.003
Vegetative propagation	Present, n (%)	434 (59.0%)	294 (67.6%)	140 (46.7%)	<0.001
Both propagation types	Present, n (%)	393 (53.5%)	259 (59.5%)	134 (44.7%)	<0.001

Note: Continuous variables were compared using the Mann–Whitney U test. Binary variables were compared using Pearson's chi-square test or Fisher's exact test where appropriate. P values were adjusted using the Benjamini–Hochberg procedure.

The distributions of the most strongly differentiated continuous predictors are shown in Figure 1. The median size of the native range of invasive taxa was larger than that of non-invasive taxa (Figure 1A), with the upper range distribution being larger as well (Figure 1B). For the invasive taxa, climatic suitability was shifted upwards, but there was still a high degree of overlap between the two groups (Figure 1B). Figure 1C shows that invasive taxa had a lower median

phylogenetic distance than non-invasive taxa. The boundaries of the boxes are the interquartile range, the lines inside the boxes are the medians, and the dots show the individual plant taxa. The Benjamini–Hochberg procedure was used to obtain adjusted p-values.

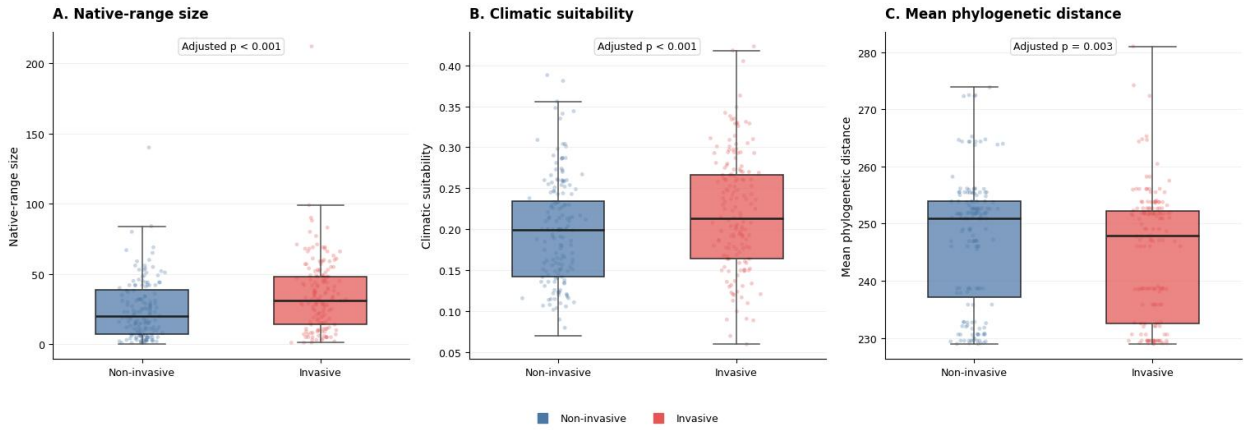

Figure 1. Distribution of leading ecological predictors by invasion status: (A) native-range size, (B) climatic suitability, and (C) mean phylogenetic distance.

3.2 Correlation structure among ecological predictors

The Spearman correlation structure between the main continuous and count predictors and invasion status is presented in Figure 2. Positive coefficients mean that variables moved in the same direction; negative coefficients mean they moved in opposite directions. Propagule-pressure indicators, climatic suitability, size of native range, plant height, phylogenetic distances and invasion status are part of the matrix. Native-range size ($\rho = 0.202$) and climatic suitability ($\rho = 0.167$) were the strongest positive correlations with invasion status. Positive correlations were somewhat less pronounced with the number of cultivated provinces, botanical gardens and economic uses ($\rho = 0.082$, 0.070, and 0.067, respectively). Mean phylogenetic distance, weighted mean phylogenetic distance, and minimum phylogenetic distance were, on the contrary, negatively related to invasion status ($\rho = -0.121$, -0.116, and -0.087, respectively). Both the number of cultivated provinces and the mean and weighted mean phylogenetic distance showed significantly high correlation with the number of botanical gardens ($\rho = 0.935$; $\rho = 0.992$, respectively). These relationships were helpful in excluding one variable from each pair of highly correlated variables prior to the modelling of the multivariable. Moderate correlations were detected between the economic uses and the cultivated provinces ($\rho = 0.538$), economic uses and botanical gardens ($\rho = 0.508$) and economic uses and maximum plant height ($\rho = 0.482$).

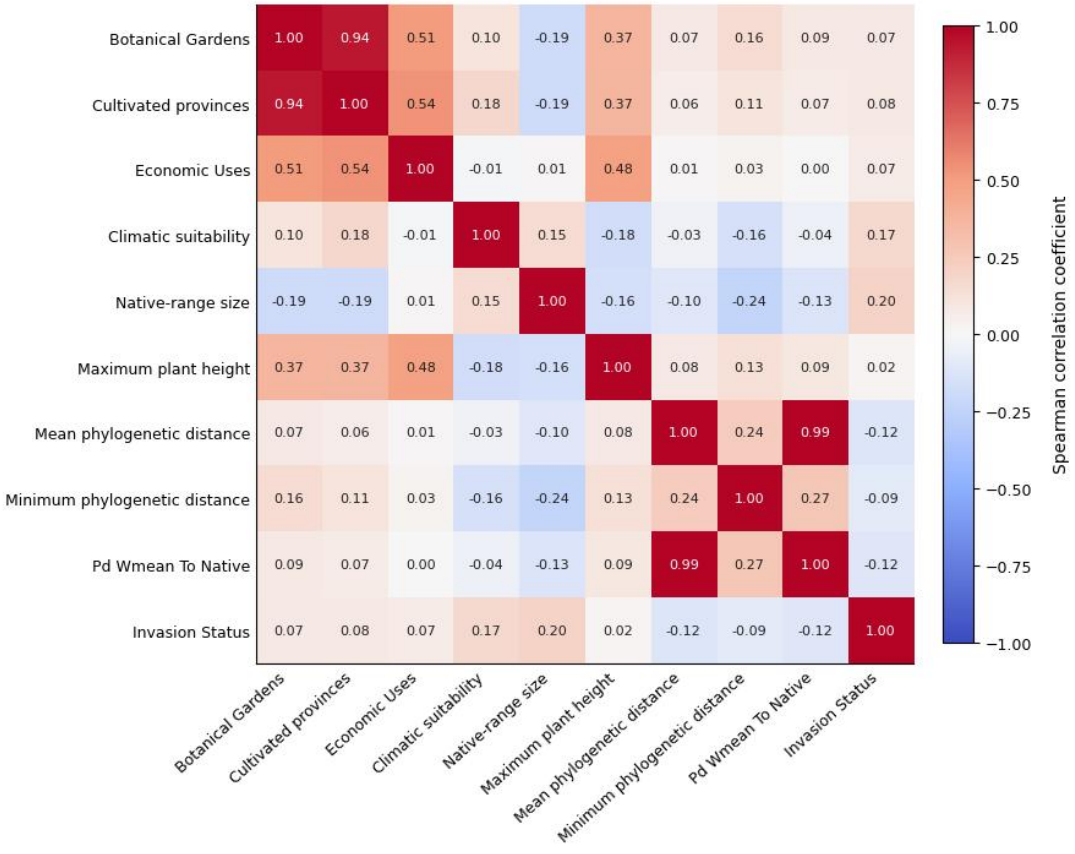

Figure 2. Spearman correlations among ecological predictors and invasion status.

3.3 Multivariable predictors of invasion success

After eliminating rare, structurally redundant and highly collinear predictors, the final logistic regression model contained 27 predictors. The model achieved successful convergence and was statistically significant compared to the intercept-only model (likelihood-ratio $p < 0.001$). The pseudo-R^2 of McFadden was 0.145, suggesting that the ecological and biological variables included explained a significant portion of the variation in the invasion status. The relationships between the retained predictors and invasion success are reconfirmed in Table 2 following adjustment for other predictors. There were six variables that remained at the 5% level of significance. The number of native-range units was positively correlated with invasion success, with an increase of one native-range unit representing an increase of 1.9% in odds of invasiveness (adjusted odds ratio [AOR] = 1.020, 95% confidence interval [CI]: 1.007–1.032; p = 0.002). Invasion status was also positively correlated with climatic suitability (AOR = 82.925, 95% CI: 3.443–1997.380; p = 0.006); however, this was due to the limited fractional range of this predictor. Taxa used as human food had 1.81 times greater adjusted odds of being invasive than taxa without this use (AOR = 1.806, 95% CI: 1.017–3.207; p = 0.044). The odds of invasion were also higher for poisonous-use (AOR = 1.605, 95% CI: 1.011–2.547; p = 0.045). In contrast, vegetative propagation was associated with decreased odds of invasive (AOR = 0.528; 95% CI: 0.348–0.801; p = 0.003). There was also a very slight negative relationship with invasion success for minimum phylogenetic distance (AOR = 0.992, 95% CI: 0.985–1.000; p = 0.046).

Table 2. Multivariable logistic regression predicting plant invasion success

Predictor	Coefficient	Adjusted odds ratio	95% confidence interval	p value
Native-range size	0.019	1.020	1.007–1.032	0.002
Vegetative propagation	−0.638	0.528	0.348–0.801	0.003
Climatic suitability	4.418	82.925	3.443–1997.380	0.006
Human food	0.591	1.806	1.017–3.207	0.044
Poisons	0.473	1.605	1.011–2.547	0.045
Minimum phylogenetic distance	−0.008	0.992	0.985–1.000	0.046
Woody species	0.476	1.609	0.989–2.620	0.056
Gene source	−0.522	0.593	0.346–1.018	0.058
Materials	−0.502	0.605	0.356–1.029	0.064
Native region X4	−0.566	0.568	0.309–1.042	0.068
Native region X8	0.417	1.518	0.933–2.471	0.093
Medicinal use	0.414	1.513	0.909–2.520	0.111
Seed propagation	0.617	1.853	0.711–4.832	0.207
Cultivated provinces	0.025	1.025	0.986–1.066	0.212
Invertebrate food	0.451	1.570	0.765–3.219	0.219
Native region X2	−0.313	0.731	0.401–1.333	0.306
Native region X6	−0.797	0.451	0.089–2.285	0.336
Mean phylogenetic distance	−0.009	0.992	0.975–1.009	0.336
Maximum plant height	−0.015	0.985	0.954–1.017	0.361
Social use	−0.227	0.797	0.422–1.503	0.483
Native region X3	−0.241	0.786	0.352–1.756	0.557
Native region X1	−0.200	0.819	0.371–1.806	0.621
Environmental use	−0.094	0.910	0.576–1.438	0.687
Fuel use	−0.117	0.890	0.429–1.843	0.753
Native region X7	−0.029	0.972	0.555–1.701	0.920
Short-lived herb	−0.024	0.976	0.582–1.639	0.928
Native region X5	−0.030	0.971	0.368–2.559	0.952

Note: Adjusted odds ratios were estimated from the multivariable binary logistic regression model. Long-lived herbs served as the reference life-form category. Variables removed before modelling included rare, mathematically redundant, or highly correlated predictors.

3.4 Predictive performance and variable importance

The same independent test set (147 taxa) was used for evaluation of the logistic regression and Random Forest models. The test-set results of both models are presented in Table 3. Logistic regression achieved an accuracy of 0.735, sensitivity of 0.617, specificity of 0.816, precision of 0.698 and F1 score of 0.655. Its ROC AUC was 0.758, with a bootstrap 95% CI of 0.677–0.834. Random Forest achieved an accuracy of 0.707 and a ROC AUC of 0.752 (95% CI: 0.668–0.828). It performed slightly better at 0.851, but not as well at 0.500, compared with logistic regression. The two models obtained the same precision value of 0.698, while the Random Forest model got 0.583. The overlapping 95% CIs suggested generally similar discrimination, and the overall test-set classification was better with logistic regression.

Table 3. Test-set performance of plant invasion prediction models

Model	Accuracy	Sensitivity	Specificity	Precision	F1 score	ROC AUC (95% CI)
Logistic regression	0.735	0.617	0.816	0.698	0.655	0.758 (0.677–0.834)
Random Forest	0.707	0.500	0.851	0.698	0.583	0.752 (0.668–0.828)

Note: Sensitivity represents the proportion of invasive taxa correctly classified, while specificity represents the proportion of non-invasive taxa correctly classified. Confidence intervals for ROC AUC were estimated using bootstrap resampling.

The model results from the comparisons and the predictor results as a rank based on permutations are given in Figure 3. Random Forest performed slightly better than logistic regression in terms of specificity, while logistic regression outperformed RF in terms of accuracy, sensitivity, F1 score and ROC AUC (Figure 3A). Vegetative propagation resulted in the greatest decrease in Random Forest test ROC AUC following permutation, with climatic suitability, size of native range, cultivated provinces and native region X8 following in order. Permutation Importance shows the decrease in test ROC AUC with the predictors randomized in each iteration; error bars are calculated as the standard deviation of the permutations.

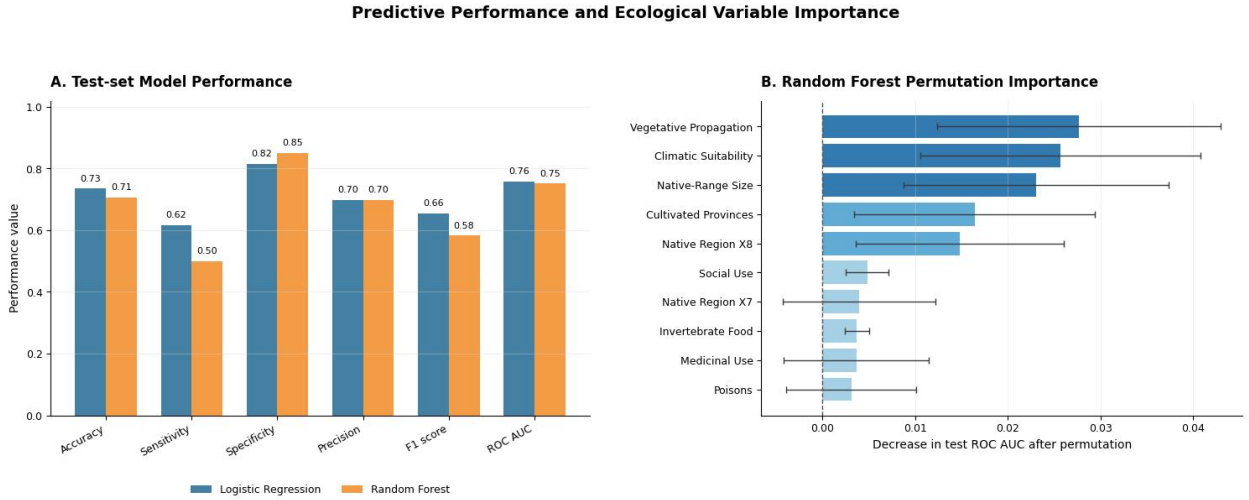

Figure 3. Predictive performance and ecological variable importance: (A) test-set performance of logistic regression and Random Forest models and (B) Random Forest permutation importance.

The training discrimination (ROC AUC) was close to perfect with the Random Forest model (0.9996 compared to 0.7521 for test discrimination), with a training–testing difference of 0.2475 and out-of-bag accuracy was 0.6701. Such results showed significant overfitting with the hyperparameter selection process cross-validated. Logistic regression was the more stable and interpretable base model for the identification of ecological drivers of invasion success, and was therefore used as the primary model.

4. Discussion

The results suggest that a combination of biogeographical breadth, climatic compatibility, human use, reproductive strategy and phylogenetic relatedness has an impact on the success of plant invasion. Native-range size was one of the most consistent predictors through descriptive, regression and machine learning analyses. Species with wider distributions in their native habitats were more prone to be invasive, which indicates that wide native range distributions are associated with ecological tolerance, dispersal ability, or persistence in a heterogeneous environment. Invasion status also had a significant positive association with climatic suitability, indicating that climatic matching is important during establishment and later invasion. Human-use characteristics also contributed additional explanatory value. Higher adjusted odds of invasion for species employed for food or for those deemed as poisonous. Associations could be a consequence of cultivation re-occurrence, intentional introduction, increased exposure to human-modified landscapes or ongoing propagule supply. Based on these results, it is thus suggested that utilitarian value could be used as an indirect measure for introduction effort and exposure opportunity. At the unadjusted descriptive level, medicinal use was higher for invasive taxa, but it did not stand out in the multivariable model, suggesting that it may have been confounded by other ecological or cultivation-related factors.

After adjustment, vegetative propagation was negatively correlated with invasion success. This finding does not mean that clonal reproduction is harmful in general, but rather it may have a varying impact depending on the ecological situation, species characteristics, and the local/long-distance spread balance. Seeding was not used to a greater degree of discrimination in either of the groups and was relatively uncommon. The extent of the negative relationship between the invasion status and phylogenetic distance was small, suggesting that some ecological traits of the invaded flora might be pre-adapted for invasion by more closely related taxa. The modest effect size suggests that phylogenetic relatedness is a part of a larger invasion process. Random Forest had a slightly better test accuracy, test sensitivity, test F1 score and test ROC-AUC than the logistic regression model. This was a minor benefit, but yielded more reliable estimates of independent ecological effects and greater stability between training and testing data. High training discrimination, but significantly lower test discrimination, suggests that the Random Forest is overfitting. This comparison shows that there is no apparent advantage to having more complex models in the present data set, because they are not better at generalizing. In conclusion, the results confirm the use of interpretable statistical models as the key analysis method, while machine learning can be used as a secondary tool to prioritize the different predictors and identify potentially non-linear patterns. These interpretations follow in the course of the manuscript, and correspond to the analytical results.

The significance of the size of native range is consistent with the worldwide pattern that widely distributed alien plants

are likely to have wide environmental tolerances and high potential for establishing in other regions and expanding their ranges. Geographic range size is thus a distributional trait as well as an informative measure of the invasiveness potential (Pyšek et al., 2017). This relationship was also found in European floras, with an increase in geographic range and habitat occupancy correlated with an increase in dimensions of invasiveness (Fristoe et al., 2021). The positive effect of climatic suitability is consistent with ecological niche modelling studies that have shown that the risk of invasion increases where the conditions in the recipient environment are similar to the climatic niche of alien plants. Due to species-specific responses and regional conditions, this suitability can be enhanced or reduced in the future in response to climate change (Gong et al., 2020). Habitat modelling using machine learning has also been demonstrated to enhance the ability to combine environmental variables and future climate scenarios to better identify plant invasion-susceptible landscapes (Sittaro et al., 2023).

The phylogenetic results partially validate the relatedness hypothesis as a factor for the susceptibility of a community. Experimental data show that it is possible that a decline in phylogenetic diversity can make a system more vulnerable to invasion under high propagule pressure, implying that invasion success can be affected by interaction between the evolutionary structure of the invader and the intensity of the invasion (Ketola et al., 2017). The observed negative relationship with minimum phylogenetic distance could therefore be due to environmental filtering/ecological similarity, not phylogenetic distinctiveness. Many differences in the functional attributes of invasive versus non-invasive taxa have been documented. There are varying effects of invasive species on aspects of resource use, growth and reproduction when compared to native or non-invasive alien plants, but this varies between environments (Mathakutha et al., 2019). This negative correlation seen in vegetative propagation could be this context dependence. Reproductive or defensive traits also show geographical variation due to genetic differentiation, plastic response to the environment, and changes in their importance along environmental gradients (Bhattarai et al., 2017). Ideally, the identified role of climatic suitability needs to be viewed within the context of the wider ecological networks. Mutualistic, competitive and trophic interactions may be disrupted by altered climates in ways that can't be predicted from climate variables, which may enable establishment success in generalist species (Fontúrbel et al., 2021). This might be an explanation for the moderate test discrimination the models had.

The results apply to invasion risk screening. Both native-range size and climatic suitability were consistently correlated with invasion status and were among the top-ranked variables in predictive analyses and can be used in preliminary assessments. Human use categories could also be used for further screening, identifying those taxa that have been repeatedly introduced and dispersed by humans. The integrated analytical framework provides a methodological benefit, too. Unlike the Random Forest, which gives information about the importance of the predictors, logistic regression gives interpretable effect estimates for ecological inference. These two approaches can be combined for enhancing decision-making capability without solely relying on predictive accuracy. Risk assessment programmes could thus profit from models that help to minimize discrimination, ensure interpretability and ecological plausibility.

Some restrictions are noted. Invasion status was treated as a binary outcome, and did not reflect abundance, rate of spread, or ecological effects. A few variables were not included, though were potentially relevant, including residence time, disturbance intensity, and local habitat conditions, and introduction frequency. Simple distribution-based methods were applied to the imputation of missing values, which may have led to a loss of variation in a small number of cases.

More studies are needed to confirm the ability of the identified predictors with other regional datasets and other taxonomic groups. Longitudinal studies may look at the naturalization to the invasion, and spatially explicit models may include land use, climate change, and introduction pathways. Other machine learning algorithms can be tested as well, but future modelling should emphasize tests for external validation and ecological interpretability, rather than only testing the model's ability to perform.

5. Conclusion

Plant invasion success was determined by a complex interplay of biogeographical, climatic, functional, human-use and phylogenetic factors, and not solely by any single mechanism. Ecological tolerance and environmental compatibility were the most prevalent positive factors, with broader native ranges and climatic suitability also being good indicators for establishment and spread. Cultivation history and repeated human-mediated exposure (as in the use of human food or its classification as a poison) also increased invasion likelihood, indicating that history of cultivation and repeated exposure may be factors in invasion risk. However, vegetative propagation and higher minimum phylogenetic distance were linked to lower invasion odds. The comparative modelling approach demonstrated that logistic regression had slightly higher and more stable test performance than the Random Forest model, and had an improved ecological interpretation. Random Forest proved to be valuable for finding important predictors and presented evidence for overfitting. The results show the importance of using an inferential and predictive approach in risk assessment for invasion. The screening and prioritization frameworks should thus take into account native-range size, climatic suitability, human use, reproductive strategy and phylogenetic proximity. A validation of these ecological predictors in independent regions and taxonomic groups is important for further relying on their transferability.

References

1. Adhikari, P., Jeon, J. Y., Kim, H. W., Shin, M. S., Adhikari, P., & Seo, C. (2019). Potential impact of climate change on plant invasion in the Republic of Korea. Journal of Ecology and Environment, 43(1), 36.
2. Bhattarai, G. P., Meyerson, L. A., Anderson, J., Cummings, D., Allen, W. J., & Cronin, J. T. (2017). Biogeography of a plant invasion: genetic variation and plasticity in latitudinal clines for traits related to herbivory. Ecological Monographs, 87(1), 57-75.
3. Cassey, P., Delean, S., Lockwood, J. L., Sadowski, J. S., & Blackburn, T. M. (2018). Dissecting the null model for biological invasions: A meta-analysis of the propagule pressure effect. PLoS biology, 16(4), e2005987.

4. Dick, J. T., Alexander, M. E., Ricciardi, A., Laverty, C., Downey, P. O., Xu, M., ... & Shaw, R. H. (2017). Functional responses can unify invasion ecology. Biological invasions, 19(5), 1667-1672.
5. Dong, B.-C., Dong, R., Yang, Q., Kinlock, N., Yu, F.-H., & van Kleunen, M. (2025). Data from: Predicting invasion success of cultivated naturalized plants in China [Dataset]. Dryad. https://doi.org/10.5061/dryad.2ngf1vj08
6. Enders, M., Havemann, F., Ruland, F., Bernard-Verdier, M., Catford, J. A., Gómez-Aparicio, L., ... & Jeschke, J. M. (2020). A conceptual map of invasion biology: Integrating hypotheses into a consensus network. Global Ecology and Biogeography, 29(6), 978-991.
7. Fontúrbel, F. E., Nespolo, R. F., Amico, G. C., & Watson, D. M. (2021). Climate change can disrupt ecological interactions in mysterious ways: using ecological generalists to forecast community-wide effects. Climate Change Ecology, 2, 100044.
8. Fristoe, T. S., Chytrý, M., Dawson, W., Essl, F., Heleno, R., Kreft, H., ... & Van Kleunen, M. (2021). Dimensions of invasiveness: Links between local abundance, geographic range size, and habitat breadth in Europe's alien and native floras. Proceedings of the National Academy of Sciences, 118(22), e2021173118.
9. Gioria, M., Hulme, P. E., Richardson, D. M., & Pyšek, P. (2023). Why are invasive plants successful?. Annual Review of Plant Biology, 74(1), 635-670.
10. Gong, X., Chen, Y., Wang, T., Jiang, X., Hu, X., & Feng, J. (2020). Double-edged effects of climate change on plant invasions: Ecological niche modeling global distributions of two invasive alien plants. Science of the total environment, 740, 139933.
11. Jeschke, J. M., & Starzer, J. (2018). Propagule pressure hypothesis. Invasion Biology-Hypothesis and Evidence. CABI Invasives Series, Oxfordshire, UK: CABI, Wallingford.
12. Ketola, T., Saarinen, K., & Lindström, L. (2017). Propagule pressure increase and phylogenetic diversity decrease community's susceptibility to invasion. BMC ecology, 17(1), 15.
13. Liu, Y. J., Huang, W., Yang, Q., Zheng, Y. L., Li, S. P., Wu, H., ... & Ding, J. Q. (2022). Research advances of plant invasion ecology over the past 10 years.
14. Liu, Y., Oduor, A. M., Zhang, Z., Manea, A., Tooth, I. M., Leishman, M. R., ... & Van Kleunen, M. (2017). Do invasive alien plants benefit more from global environmental change than native plants?. Global Change Biology, 23(8), 3363-3370.
15. Lone, P. A., Dar, J. A., Subashree, K., Raha, D., Pandey, P. K., Ray, T., ... & Khan, M. L. (2019). Impact of plant invasion on physical, chemical and biological aspects of ecosystems: A review. Tropical Plant Research, 6(3), 528-544.
16. Mathakutha, R., Steyn, C., le Roux, P. C., Blom, I. J., Chown, S. L., Daru, B. H., ... & Greve, M. (2019). Invasive species differ in key functional traits from native and non-invasive alien plant species. Journal of Vegetation Science, 30(5), 994-1006.
17. McMahon, D. E., Urza, A. K., Brown, J. L., Phelan, C., & Chambers, J. C. (2021). Modelling species distributions and environmental suitability highlights risk of plant invasions in western United States. Diversity and Distributions, 27(4), 710-728.
18. Pyšek, P., Hulme, P. E., Simberloff, D., Bacher, S., Blackburn, T. M., Carlton, J. T., ... & Richardson, D. M. (2020). Scientists' warning on invasive alien species. Biological Reviews, 95(6), 1511-1534.
19. Pyšek, P., Pergl, J., Essl, F., Lenzner, B., Dawson, W., Kreft, H., ... & van Kleunen, M. (2017). Naturalized alien flora of the world: species diversity, taxonomic and phylogenetic patterns, geographic distribution and global hotspots of plant invasion.
20. Sittaro, F., Hutengs, C., & Vohland, M. (2023). Which factors determine the invasion of plant species? Machine learning based habitat modelling integrating environmental factors and climate scenarios. International Journal of Applied Earth Observation and Geoinformation, 116, 103158.
21. van Kleunen, M., Bossdorf, O., & Dawson, W. (2018). The ecology and evolution of alien plants. Annual review of ecology, evolution, and systematics, 49(1), 25-47.
22. Vilà, M., & Hulme, P. E. (Eds.). (2017). Impact of biological invasions on ecosystem services (Vol. 12). Cham: Springer.