12-2015

The effectiveness of a single regional model in predicting non-native woody plant naturalization in five areas within the Upper Midwest (United States)

Philip M. Dixon
Iowa State University, pdixon@iastate.edu

Janette R. Thompson
Iowa State University, jrrt@iastate.edu

Mark P. Widrlechner
Iowa State University, isumw@iastate.edu

Emily J. Kapler
Iowa State University, ekapler@iastate.edu

The effectiveness of a single regional model in predicting non-native woody plant naturalization in five areas within the Upper Midwest (United States)

Abstract
Numerous predictive models have been developed to determine the likelihood that non-native plants will escape from cultivation and potentially become invasive. Given the substantial biological and economic costs that can result from the introduction of a new invasive plant and the unending pressures of world trade and transport, the creation and implementation of effective predictive models are becoming increasingly important. One key question in the development of such models focuses on the geographic scope at which models can best be developed and applied. We have developed models to predict woody-plant naturalization in five local areas within the Upper Midwest (United States). Herein, we consider whether naturalization can be reasonably predicted from a single model for the entire region or whether local models are required for each specific area. We develop a random forest model to predict the probability of naturalization in the region and compare out-of-sample prediction errors between the regional and local models. The regional model makes better predictions of the probability of naturalization for those species observed to naturalize but worse predictions for those not currently observed to naturalize. This model development process has given us an opportunity (not previously addressed in the literature) to examine the strengths and weaknesses of local and regional approaches, with the ultimate intent of optimizing geographic scope.

Keywords
invasion, geographic scope, model-based decision making, non-native plants, risk analysis, woody plants

Comments
This is a post-peer-review, pre-copyedit version of an article published in *Biological Invasions*. The final authenticated version is available online at: 10.1007/s10530-015-0976-2. Posted with permission.

Dixon, P.M., Thompson, J.R., Widrlechner, M.P, & Kapler, E.J. 2015. The effectiveness of a single regional model in predicting non-native woody plant naturalization in five areas within the Upper Midwest (United States). *Biological Invasions* 17: 3531-3545. DOI 10.1007/s10530-015-0976-2

The effectiveness of a single regional model in predicting non-native woody plant naturalization in five areas within the Upper Midwest (United States)

Phillip M. Dixon • Janette R. Thompson • Mark P. Widrlechner • Emily J. Kapler

P. M. Dixon, Department of Statistics, Iowa State University, 2121 Snedecor Hall, Ames, IA 50010, United States; E-mail: pdixon@iastate.edu

J. R. Thompson (corresponding author), Department of Natural Resource Ecology and Management, Iowa State University, 339 Science II, Ames, IA 50011, United States; E-mail: jrrt@iastate.edu

M. P. Widrlechner, Departments of Horticulture and Ecology, Evolution, and Organismal Biology, Iowa State University, 360 Bessey Hall, Ames, IA 50011, United States; E-mail: isumw@iastate.edu

E. J. Kapler, Department of Ecology, Evolution, and Organismal Biology, Iowa State University, 105B Bessey Hall, Ames, IA 50011; E-mail: ekapler@iastate.edu

Abstract

Numerous predictive models have been developed to determine the likelihood that non-native plants will escape from cultivation and potentially become invasive. Given the substantial biological and economic costs that can result from the introduction of a new invasive plant and the unending pressures of world trade and transport, the creation and implementation of effective predictive models are becoming increasingly important. One key question in the development of such models focuses on the geographic scope at which models can best be developed and applied. We have developed models to predict woody-plant naturalization in five local areas within the Upper Midwest (United States). Herein, we consider whether naturalization can be reasonably predicted from a single model for the entire region or whether local models are required for each specific area. We develop a random forest model to predict the probability of naturalization in the region and compare out-of-sample prediction errors between the regional and local models. The regional model makes better predictions of the probability of naturalization for those species observed to naturalize but worse predictions for those not currently observed to naturalize. This model development process has given us an opportunity (not previously addressed in the literature) to examine the strengths and weaknesses of local and regional approaches, with the ultimate intent of optimizing geographic scope.

Keywords: invasion, geographic scope, model-based decision making, non-native plants, risk analysis, woody plants

Introduction

Interest in the spread and impacts of non-native and potentially invasive trees, shrubs, and vines has increased markedly in the last decade (e.g., Brooks et al. 2004; Eldridge et al. 2012; Hejda et al. 2009; Pyšek et al. 2012; Rejmánek 2014; Richardson and Rejmánek 2011; Richardson et al. 2014). Historically, the native ranges of woody plant species have naturally expanded and contracted, but human actions have greatly accelerated their global expansion (Rejmánek et al. 2013). Although many human-assisted plant introductions have been beneficial (Smith et al. 1999), others have led to naturalization (the ability of a plant to propagate and sustain a population outside of cultivation) or invasion (the ability of a naturalized plant to aggressively colonize and displace native plants) into new locations (Widrlechner et al. 2013). A recent update of a global database of trees and shrubs indicates that 751 species of woody plants (434 trees and 317 shrubs) are now considered invasive, alien species somewhere in the world (Rejmánek and Richardson 2013).

Ideally, risk-assessment models designed to predict naturalization, invasiveness, and impacts of non-native plant introductions should inform environmental policy development and decision making (Addison et al. 2013; Barney et al. 2013; Burgman and Yemshanov 2013). The use of such models to screen plant introductions could reduce the adverse ecological and economic consequences associated with potential plant invasions (Kapler et al. 2012), which can be quite substantial (Keller et al. 2007; Pimentel 2011). Efforts to develop and validate predictive risk-assessment models have expanded rapidly in recent years (Burgman and Yemshanov 2013; Gordon et al. 2008a, b; Hulme 2012; Widrlechner et al. 2009, 2013). To date, however, implementation (either regulatory or voluntary) of predictive risk-assessment models as an integral part of screening protocols has been limited (Addison et al. 2013; Jefferson et al. 2004; OIPC 2013). This may be because such models are not readily shared with entities that would use them, or because they are time-consuming and/or difficult to use (e.g., hard to find required plant trait information for each species in question) (Gordon et al. 2008a; Jefferson et al. 2004). A notable exception is the incorporation of the Australian Weed Risk Assessment Protocol into the national regulatory frameworks of Australia and New Zealand (Gordon et al. 2010).

Models to assess the risk of naturalization or invasiveness have been based upon plant life-history characteristics, geographic and climatic analyses, or a combination of the two. Typical plant life-history characteristics include population fitness, degree of fecundity, height, flowering time, generation time,

mode of pollination, and rate of population expansion (Grotkopp et al. 2004, 2010; Pyšek et al. 2014; Rejmánek 2011; Rejmánek et al. 2013). Geographic and climatic characteristics include information about a species' native range and its adaptation to climatic variables (Hui et al. 2014; Kriticos and Randall 2001; Mgidi et al. 2007; Richardson and Thuiller 2007; Widrlechner 2001; Widrlechner and Iles 2002). Risk-assessment models using a combination of these two types of characteristics have been presented by Castro-Diez et al. (2011), Kapler et al. (2012), Pyšek et al. (2014), Rejmánek et al. (2013), and Widrlechner et al. (2004, 2009, 2013). Additional characteristics that have been considered include propagule pressure, lag time, intensity of cultivation, and ecological amplitude in native habitats (Bucharova and van Kleunen 2009; Hulme 2012; Kowarik 1995; Křivánek et al. 2006; Larkin 2012; McGregor et al. 2012; Pyšek et al. 2015; Rouget and Richardson 2003; Wilson et al. 2007). To facilitate clear and rapid determinations through screening protocols, models should be based on a limited set of plant-specific characteristics, ones for which adequate information is readily available (see Jefferson et al. 2004; Gordon et al. 2008a, b), and which have the closest correspondence to naturalization.

Many different statistical methods have been used to develop predictive models, including logistic regression (Koop et al. 2012), discriminant analysis (Reichard 1994; Rejmánek and Richardson 1996), classification and regression trees (Jarošík 2011; Maillet and Lopez-Garcia 2000; Reichard and Hamilton 1997), analytic hierarchy models (Ou et al. 2008), and hierarchical Bayesian frameworks (Diez et al. 2012). Recent extensions of classification and regression-tree methods include the random forest (Cutler et al. 2007; Kapler et al. 2012; Keller et al. 2011; Philibert et al. 2011; Widrlechner et al. 2013) and the boosted regression tree (McGregor et al. 2012). These extensions are more accurate and more powerful than earlier approaches (Jarošík 2011; Kapler et al. 2012; McGregor et al. 2012; Widrlechner et al. 2013). Also, these newer techniques identify the traits that are the most important predictors of invasiveness (Archer and Kimes 2008; Kapler et al. 2012; Widrlechner et al. 2013).

A third, less studied characteristic of a risk-assessment model is its geographic scope: should a model be built to predict naturalization potential for an entire nation, for an ecoregion, or for a specific state or province? Geographic scope is especially important for woody plants because their natural adaptation and distribution patterns are often closely related to variation in climates, soils, and photoperiod or precipitation regimens (Thompson et al. 2000; Widrlechner 1994). The geographic scope of predictive

models should not be confused with assessment of the extent of invasions (e.g., Kaplan et al. 2014) or with the spatial scale/granularity of data collected to support such models (e.g., Marcer et al. 2012).

Using political boundaries to define the scope of a predictive model aligns it with the jurisdictional area of a regulatory entity likely to use it (e.g., in the United States at the level of a state), but a model developed within such boundaries may not be as sound as one developed for areas defined by important biological, edaphic, or ecological factors (Widrlechner et al. 2009, 2013). One must also keep in mind that models developed with a much larger geographic scope, such as the continental model of Reichard and Hamilton (1997), have been shown to have lower classification rates (ability to classify species into "accept" or "reject" categories vs. "needing further study") and accuracy (the proportion of classified species that are misclassified based on *a priori* knowledge of outcomes) when applied to smaller areas, most likely because of the importance of local variation in adaptation patterns of woody plants (Kapler et al. 2012; Widrlechner et al. 2004, 2009).

We are not aware of published studies intentionally designed to compare models that predict invasiveness at different geographic scopes. To help fill that gap, we compare the predictive accuracy of models developed at two geographic scopes: local models developed for five areas no larger than the state of Iowa (ca. 145,000 km^2) and a regional model developed for multiple areas (total ca. 370,000 km^2) within the Upper Midwest of the United States (Figure 1). These models build on our previously developed models which assessed the likelihood of naturalization of non-native woody plants in Iowa, and parts of Minnesota, Missouri, Illinois, Indiana, Michigan, and Wisconsin (Kapler et al. 2012; Widrlechner et al. 2004, 2009, 2013). Modeling approaches we used previously included decision trees, CART (classification and regression tree) models, models combining these techniques, and random forest models. Our previous models incorporated species' life-history characteristics and geographic criteria related to species' native range. One key geographic criterion in our previous models was the geographic risk value, calculated from the ratio of the number of species native to a geographic subdivision known to naturalize in a given area to the total number of species cultivated, native to that same geographic subdivision (Widrlechner et al. 2004). This was then used to calculate a range-wide geographic risk value (G-value) by averaging the unweighted ratios for each species based on all the geographic subdivisions in its native range.

Developing local models required generating and testing different models for each area and evaluating their operating characteristics, such as false positive and false negative rates. We found that random

forest models typically made the best predictions for each individual area (Kapler et al. 2012). However, because the creation of large numbers of local models is neither efficient nor realistic, and continental models lack power and accuracy when applied to smaller geographic areas, development of models with regional scope could provide a "middle ground" with better power and accuracy, and potentially more frequent use, thereby making them more cost-effective. Further, comparisons of relative variable importance for random forest models, based on five distinct data sets from the Upper Midwest, revealed very similar patterns among variable weighting for the five locally-derived models (Widrlechner et al. 2013). In fact, one of the key variables predicting the probability of naturalization, the geographic risk value, is highly correlated in data sets for Iowa and two areas around Chicago (Widrlechner et al. 2009). Together, these results pointed toward the potential for a regional approach to model development.

The work reported here evaluates whether a single, regional model for the Upper Midwest, created based on species characteristics and naturalization status for that larger area (Figure 1), could make predictions as accurately as local models developed for each of the five areas from previous studies. We did this by comparing the accuracy of the regional random forest model to that of random forest models derived from localized data sets for each area studied within the region. We considered random forest models because previously we found that random forest models gave the overall best predictions for each local area (Kapler et al. 2012).

Methods

Kapler et al. (2012) and Widrlechner et al. (2004, 2009, 2013) compiled lists of non-native woody plants currently cultivated in five study areas of the Upper Midwest (Figure 1): southern Minnesota, Iowa, northern Missouri, and two areas around Chicago that we refer to as Chicago A (northeastern Illinois and southeastern Wisconsin) and Chicago B (northwestern Indiana and southwestern Michigan). Because of the known lag time between woody plant introduction and naturalization (Kowarik 1995; Larkin 2012), only species with a record of cultivation in their respective areas for at least thirty years were included. The naturalization status of each species in these study areas was determined by using herbarium specimens, local floras, and consultation with local floristic experts (e.g., Kapler et al. 2012; Widrlechner et al. 2009).

Information on traits previously associated with invasiveness and used in prior risk-assessment models (e.g., Reichard and Hamilton 1997) were obtained from published and online sources (Widrlechner et al. 2013). These traits included seven life-history characteristics (evergreen foliage, fleshy bird-dispersed fruits, group invasive in North America, invades outside North America, quick maturity, quick vegetative spread, and requires germination pretreatment) and whether or not the species is native to North America. Local experts reviewed species information for each of the five study areas to ensure its accuracy for that area.

The final trait used to predict naturalization status is the range-wide geographic risk factor, or G-value. This value summarizes the tendency of species with similar native ranges to naturalize in the study area. G-values were computed from information on the native range of each species following the procedures described in Widrlechner et al. (2004). Briefly, for the cultivated woody plants in each study area, we calculated the proportion of species native to a geographic subdivision (i.e., countries, provinces/states or subdivisions of provinces/states) that have naturalized in the study area. This is the P-value for that region. The G-value is the unweighted average of the P-values for all geographic subdivisions that comprise a species' native range. A species with a G-value close to zero is a species with a native range that includes very few naturalizing species, while one with a G-value close to one is a species with a native range from which nearly all evaluated species that are cultivated have naturalized there. For example, *Abies concolor* (Gord. & Glend.) Lindl. ex Hildebr. is endemic to the western United States, a region where our list of native cultivated species included no naturalizers. It received a G-value of zero in our regional analysis. In contrast, *Caragana arborescens* Lam. is native to northeastern Asia, and many introduced species from that region naturalize in the Upper Midwest. It received a G-value of 0.57 in our regional analysis.

In order to make a comparison between the original study areas and a regional study area, a regional list of non-native woody plants was assembled from the five local study areas by omitting any species native somewhere within the discontinuous region (Figure 1). These species were also omitted from their original study areas. For example, *Tsuga canadensis* (L.) Carrière was included in our previous Iowa, Missouri, and Chicago A studies. It was omitted from all data sets for the current study because it is native to southwestern Michigan (part of the Chicago B study area). A species was declared naturalizing in the region if it naturalized anywhere within the region, based on the presence of at least two independent naturalization events documented by herbarium vouchers. When a species was included

in more than one local list, we checked the life-history traits for consistency. If a trait varied across local lists, the regional value of that trait was recorded as 'varies'. This regional species and trait matrix is provided as a file in Supplemental Material 2.

The randomForest function in the R (R Core Team 2011) randomForest package (Liaw and Wiener 2002) was used to construct a regional random forest model to predict naturalization status from eight traits and the G-value. We used default settings for all options except the number of trees in the forest, which was increased to 2000. New random forests, one for each local study area (five total), were constructed by using the subset of species that were known to be cultivated in each local area. These local models used the G-values generated from the regional data set combined with life-history trait information determined for the local area. Sample sizes for each new local model are given in Table 1. These are slightly smaller than the sample sizes reported previously (Kapler et al. 2012; Widrlechner et al. 2009, 2013) because species native elsewhere in the region are omitted. The importance of each variable to the random forest predictions was quantified by the mean decrease in the Gini measure of node impurity (Hastie et al. 2009), as calculated by the importance function in the randomForest package.

Each random forest model predicted the probability of naturalization, or P[nat], for each species. We primarily considered out-of-sample predictions based on leave-one-out cross-validation (Harrell 2001), where each species is omitted from the model used to calculate its P[nat]. We also calculated in-sample predictions, where P[nat] for a species is calculated from a model built including that species, and out-of-bag predictions from the random forest. We quantified the error in a set of predictions in two ways. One is to classify each species as "naturalizing" or "not" by comparing P[nat] to a threshold and then calculate the misclassification, false positive, and false negative rates. Unlike our previous evaluations of models (e.g., Kapler et al. 2012), where there were three potential classification outcomes (accept, reject, or require further study), we defined error here based on two outcomes (accept, because P[nat] is small, or reject, because P[nat] is large). We also quantified the prediction error for each species, defined as 1-P[nat] when the species is known to naturalize, and P[nat] when the species is not known to naturalize. Unlike misclassification probabilities, the prediction error avoids the need to choose a threshold P[nat] above which a species is classified as "naturalizing." Further, unlike area-under-the curve and similar summaries across all possible thresholds, the prediction error separately quantifies the error for species that naturalize and for those that do not. Error magnitudes were compared between the regional and each of the five local models by using a Wilcoxon signed-rank test for paired data (Sprent and Smeeton 2007).

Results

The importance of variables in the regional random forest model is similar to the importance in each local model (Figure 2). The most important variable is the G-value, a measure of the relative likelihood of naturalization based on the naturalization histories of woody plants that share a species' native range, followed by whether the species matures quickly, and whether it is invasive outside of North America. The remaining five variables generally have small variable importance values (typically < 0.1), both in the regional model and in each local model.

For each of the five local models, the in-sample misclassification rate (Table 2) is substantially smaller than either the out-of-bag misclassification rate or the leave-one-out misclassification rate. In contrast, the in-sample misclassification rate for the regional model is approximately the same as the regional out-of-bag and leave-one-out misclassification rates. For all six models, the out-of-bag estimate of the misclassification rate is similar to the leave-one-out cross-validation estimate (Table 2).

Across the five data sets, between 13% and 25% of the species are classified differently by the local and regional models (Table 3). There is no consistent tendency for models at one extent or the other to be more likely to predict naturalization. For species in southern Minnesota, the regional model is significantly more likely to classify a species as a naturalizer (Table 3). But, the differences between local and regional models are smaller in the other four study areas, and are neither consistently in favor of one model nor significantly different from zero (Table 3).

When the empirical proportion of naturalizing species within an entire data set is used as the threshold to classify a species as a potential naturalizer, the false negative and false positive rates vary between 19% and 49% (Table 4). Across the five study areas, there is no consistent trend for the local model to perform better than the regional model, or vice-versa, on either type of error. Pooling over the five study areas, the estimated false negative rate for the local model is 2.4% larger than that for the regional model, but the difference is not significant ($p = 0.57$), using McNemar's test for paired binary data (Sprent and Smeeton 2007).

When screening for potentially naturalizing species, a false negative error (i.e. concluding that a species will not naturalize when in fact it will) is much more serious than a false positive error. The false negative rate can be reduced by using a lower threshold of naturalization probability, P[nat], to classify a species as a potential naturalizer. When the threshold is set at 0.10, the false negative rates across the

five study areas vary from 5% to 13% (Table 4). In all five study areas, the local model has a lower (or similar) false negative rate than does the regional model. Pooling over the five study areas, the estimated false negative rate for the local model is 3.6% lower than that for the regional model, but the difference is not significant ($p = 0.26$, using McNemar's test).

When computed for all species, the median out-of-sample prediction error for the local model is significantly smaller than that for the regional model in southern Minnesota and has no evident pattern in the other four study areas (Table 5). When computed separately for naturalizing and non-naturalizing species, the regional model has a smaller median error for naturalizing species in all study areas (Table 6). This difference is significant ($p < 0.05$) in two study areas and weakly significant in a third study area (Table 6). Conversely, the local model has the smaller median out-of-sample error for non-naturalizing species, but this is statistically significant only in southern Minnesota (Table 6).

For most species in southern Minnesota, both naturalizing and non-naturalizing, the predicted probability of naturalization is larger from the regional model than from the local model (Figure 3). In the other four regions, there is no consistent difference between the two models; instead the pattern is one of heterogeneity among species, especially in northern Missouri. In northern Missouri, the regional model has a much higher predicted probability of naturalization for some species, compared to the local model, but a much smaller probability for others (Figure 3). The two probabilities are more similar to each other in Chicago B (Figure 3), which is the region with the lowest median prediction error. The patterns in Iowa and Chicago A are intermediate between those for northern Missouri and Chicago B (Supplemental Appendix, Figure A1).

In all five regions, there are some naturalizing species for which both models predict a low probability of naturalization (Figure 3). When analyzing model performance, it can be instructive to examine these and other unexpected outcomes on an individual basis, especially those that generate errors. In Tables 7 and 8, we present two lists of species that generated the most extreme errors, resulting in their misclassification. Table 7 includes those five naturalizing species with the lowest P[nat] values for each of our six models. Accounting for duplication, there are 18 taxa on that list. Conversely, Table 8 includes those five non-naturalizing species with the highest P[nat] values for each of the six models. There are 22 taxa on the second list.

Of the 18 naturalizing species, the ones most consistently generating extreme errors are *Berberis thunbergii* DC, *Catalpa speciosa* (Warder ex Barney) Warder ex Engelm., and *Maclura pomifera* (Raf.) C.K. Schneid. (Table 7). Ten species are idiosyncratic, included in only one model. And of the 22 non-naturalizing species, the ones most consistently generating extreme errors are *Salix caprea* L. and *Prunus cerasifera* Ehrh. (Table 8). Seventeen of the species in this second group are unique.

Discussion

As previously noted for independently created local models in the Upper Midwest (Widrlechner et al. 2013), the G-value was always the most important contributor to model performance. This is not surprising since G integrates ecogeographic adaptation based on native ranges with the naturalization histories of an array of sympatric woody plants cultivated in the target region, and has also been related to the occurrence of climatic analogs between native ranges and regions of intentional introduction (Widrlechner and Iles 2002). The G-value or similar metrics have not been widely used in models to predict naturalization or invasion. Although Castro-Díez et al. (2011) examined the role that climatic adaptation in native habitats plays in naturalization of *Acacia*, they did not extend their analysis to examine sympatric species or relationships in environmental conditions between native and introduced sites that are quantified by the G-value or a similar metric. Our results suggest that incorporating the G-value has the potential to significantly improve predictions of naturalization.

The two traits with moderate contributions to the regional model and at least one local model are rapid reproductive maturity and invades outside North America. Rapid reproductive maturity is commonly linked to plant invasion (Myers and Bazely 2003; Rejmanek and Richardson 1996), and while invasion outside of North America is not a specific trait per se, invasion history in other parts of the world has often been valuable in identifying potential "repeat offenders" (Gordon et al. 2012; Reichard and Hamilton 1997). Bird- (or animal-) dispersed fruits and ease of vegetative spread are biological traits that are often linked to plant invasion (Myers and Bazely 2003; Rejmanek and Richardson 1996) but were less important for the species evaluated here.

We found that species' status as native to North America contributed little to predicting naturalization, potentially because similar, but more geographically-specific, information is carried in the G-value. For species native to North America, we calculated G-values from regional and state information on the native range and whether those species were invasive in the Upper Midwest. Species native to North

America are not equally likely to naturalize in the region. For example, species native to the southeastern US are more likely to naturalize in the region than are species native to the western US and Canada. The G-value captures these regional differences, while the binary (yes/no) trait "native to North America" does not.

As might be expected, the in-sample misclassification rate is an over-optimistic estimate of the error rate for new observations, and the extent of that over-optimism depends on the sample size. The sample sizes for the five local models range from 85 to 131, and their in-sample error rates are much smaller than the error rate for the regional model developed with 230 species. Although there has been some concern that the out-of-bag error rate is a less precise and over-pessimistic estimate of the mis-classification rate (Efron and Tibshirani 1993), we found no practical difference between the two rates.

Across all five study areas, the regional model made better predictions of naturalizing species than did any local model. This came at the cost of larger errors when making predictions for non-naturalizing species. Potential reasons that predictions might differ between the local and regional model include: differences in the trait values between the local and regional data set, a larger proportion of naturalizing species in the regional data set, and a larger set of species used to develop the regional model.

With regard to trait values, some biological traits (e.g., quick maturity and rapid vegetative spread) are not constant across the entire region, which required a new category, "Varies," in the regional data set. For example, *Castanea mollissima* matures quickly in northern Missouri but not further north in Iowa or Chicago A, so it was coded as "Varies" in the regional data set. In total, quick maturity was coded as "Varies" for five species, and fast vegetative spread was coded as "Varies" for seven species, out of the 231 in the regional data set. When predicting the probability of naturalization, the northern Missouri random forest model considers *C. mollissima* with the rest of the species with quick maturity, the Iowa or Chicago A models consider *C. mollissima* with the rest of species without quick maturity, while the regional model considers *C. mollissima* as a member of the small group of species coded as "Varies." Although *C. mollissima* has not been reported to naturalize anywhere in the region, species with quick maturity are more likely to naturalize, so the predicted probability of naturalization for *C. mollissima* is much higher in the Missouri local model (P[nat] =0.65) than in either the Iowa or Chicago A local models

(P[nat] = 0.054 and 0.0085, respectively). Three of the five species with quick maturity coded as "Varies" do naturalize in the region, so the regional prediction (P[nat] = 0.54) is also high and contributes to the larger median error for non-naturalizing species in the regional model.

The variable that most frequently changes between the local and regional data sets is the response variable: whether or not a species naturalizes in the model area. Thirty-four of the 231 species are coded as not naturalizing in one or more local areas but are known to naturalize in the region. These are species that show no evidence of naturalizing in one local area but do naturalize in other local areas. For example, *Alnus glutinosa* does not naturalize in southern Minnesota but does in both Chicago areas, so is recorded as naturalizing in the region. These species that change in naturalization status represent only between 8.5% and 21.5% of the non-naturalizing species in any local area, but they do contribute substantially to the prediction error rates. Not surprisingly, the median P[nat] predicted by the regional model for the species that change in status is substantially larger than the median P[nat] for those species that are not known to naturalize anywhere within the region (supplemental Table A1). Across all five study areas, the average of the median P[nat] is 0.46 for the changing species and 0.13 for the species that do not naturalize. More surprisingly, the same pattern is seen in predictions from the local models even though both groups of species are coded as "does not naturalize" in the local model (supplemental Table A2). Across all five study areas, the average median P[nat] from the local model is 0.26 for the changing species and 0.075 for the species that do not naturalize. It is intriguing that the local model, in which both groups of species are coded as "does not naturalize," can still distinguish, at least as a group, between species that have naturalized elsewhere in the region and those that have not.

The single most important trait in the random forest models, the G-value of a species, is not identical in the regional and local models. The regional G-value is computed from a larger collection of species than for any single local G-value, and it is based on regional naturalization, not local naturalization response. Even so, the local and regional G-values are generally well correlated with each other. The two values are highly correlated for species in the Iowa and Chicago A models (correlation = 0.91 and 0.92, respectively), and slightly less so for species in the southern Minnesota, northern Missouri, and Chicago B models (correlation = 0.82, 0.68, and 0.83, respectively).

In Table 6, we list the 18 naturalizing species that generated the most extreme errors in our models. These were generated by finding the five naturalizing species with the lowest predicted P[nat] (i.e., the largest errors), in each local model and in the regional model. Among the 18 naturalizing species, the

most frequently reported error generator, *Berberis thunbergii*, is a native of Japan, which has relatively low G-values, but this species' extensive cultivation (Lubell et al. 2008), its adaptation to seasonal moisture deficit stresses, and tolerance of high deer populations (Widrlechner et al. 2013) are evidently not well accounted for within our existing models. The other two taxa with frequent and large errors, *Catalpa speciosa* and *Maclura pomifera*, are native to North America just to the south of our target region. Their native ranges have relatively low G-values, with few naturalizing species (Widrlechner and Iles 2002). Their records of naturalization, which likely resulted as a result of extensive historic cultivation (Droze 1977; Smith and Perino 1981), may now be confounded with gradual processes of natural range expansion, especially in times of global climate change.

Looking more broadly at the 18 naturalizing species in Table 6, they can be divided into three general groups. There are two small groups, one of three species, the two above-mentioned taxa plus *Lonicera sempervirens* L., native just to the south of the target region, and another also of three species, *Philadelphus coronarius* L., *Spiraea prunifolia* Siebold & Zucc., and *Rhamnus utilis* Decne., native to Eurasia, that each have very few reported naturalization events, all quite limited in extent. The remaining 12 taxa include an array of Eurasian species that are either special cases, like *B. thunbergii*, or are naturalizing only in certain parts of the region, but not in others. We suspect that most of these observed geographic limitations are driven by strong climatic or edaphic barriers that could slowly be weakened during the course of future evolution or by climate change.

Among the 22 non-naturalizing species (Table 7), only two were frequently reported: *Salix caprea* and *Prunus cerasifera*. Although these two species have been widely cultivated in Europe since pre-Colonial times (Rehder 1947) and were likely introduced to North America centuries ago, *P. cerasifera* is almost exclusively represented in cultivation by fruit-tree rootstocks (Okie 1987) and by ornamental forms with dark red leaves (Dirr 2009), which may place this taxon at a disadvantage in natural settings. For *S. caprea*, the situation may be quite different. It has naturalized to the south and east of our study area (USDA-NRCS 2014), and it may just be a matter of time before it does so in the southeastern portions of our region. We suspect that many of the other 20 taxa included in this list may still be in a lag phase (Crooks 2005; Kowarik 1995; Larkin 2012) prior to naturalization or could be undocumented or incipient naturalizers already present in neighboring areas (Widrlechner et al. 2013). We suggest that land-use managers and field botanists in the Upper Midwest give special attention to these taxa to search for their presence away from cultivation and potential spread.

If a manager needs to make predictions for a local area, is it necessary to generate a local model for the area of interest, or is a regional model adequate? Prior success with random forest models applied to local data sets (e.g. Kapler et al. 2012; Widrlechner et al. 2013) suggested that this approach to predicting naturalization could be useful at a larger scope. We find that the importance of predictor variables is similar for local and regional scopes and that their overall prediction errors are also similar. The regional model does tend to have a smaller prediction error for the species of most concern: those that have naturalized. We believe this results from an interplay between the size of data set, biological characteristics of the species, and details of individual species. As suggested by others (e.g., Hulme 2012; Kueffer et al. 2013; Pyšek et al. 2012), integrating multiple modeling approaches at various scopes to account for species' traits, environmental variation, and differences in habitat invasibility can lead to greater predictive accuracy. Incorporating additional traits, such as habitat diversity in the native range (Pyšek et al. 2015), seed mass, seed production, specific leaf area, height at maturity, length of flowering/fruiting periods, and seed bank persistence may also increase predictive accuracy, but this has to be balanced against the difficulty of obtaining such information for a species being evaluated.

Our results support the possibility of developing and implementing regional models to screen non-native plant introductions, which is promising given that a single regional model is more general and requires less work to produce than does the crafting of multiple local models. A significant challenge for future work will be to define the upper limit of geographic scope that maintains regional ecological homogeneity.

Acknowledgements

Journal paper of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA, and supported by Hatch Act, McIntire-Stennis, and State of Iowa funds. We acknowledge additional financial support from USDA-ARS through the Floriculture and Nursery Research Initiative. We thank Donald Robinson, Michael Dosmann, and two anonymous reviewers for their comments on an earlier version of the manuscript.

References

Addison PFE, Rumpff L, Sana Bau S, Carey JM, Chee YE, Jarrad FC, McBride M, Burgman MA (2013) Practical solutions for making models indispensable in conservation decision-making. Divers Distrib 19:490-502

Archer KJ, Kimes RV (2008) Empirical characterization of random forest variable importance measures. Comp Stat Data Anal 52:2249-2260

Barney JN, Tekiela DR, Dollete ESJ, Tomasek BJ (2013) What is the "real" impact of invasive plant species? Front Ecol Environm 11: 322-329

Brooks M, D'Antonio C, Richardson D, Grace J, Keeley J, DiTomaso J, Hobbs R, Pellant M, Pyke D (2004) Effects of invasive alien plants on fire regimes. BioScience 54:677-688

Bucharova A, van Kleunen M (2009) Introduction history and species characteristics partly explain naturalization success of North American woody species in Europe. J Ecol 97:230-238

Burgman MA, Yemshanov D (2013) Risks, decision and biological conservation. Divers Distrib 19:485-489

Castro-Díez P, Godoy O, Saldaña A, Richardson DM (2011) Predicting invasiveness of Australian acacias on the basis of their native climatic affinities, life history traits and human use. Divers Distrib 17:934-945

Crooks JA (2005) Lag times and exotic species: The ecology and management of biological invasions in slow-motion. EcoScience 12:316-329

Cutler DR, Edwards TC Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783-2792

Diez JM, Hulme PE, Duncan RP (2012) Using prior information to build probabilistic invasive species risk assessments. Biol Invasions 14:681-691

Dirr MA (2009) Manual of woody landscape plants (6th ed) Stipes Publishing, Champaign, IL

Droze WH (1977) Trees, prairies, and people: a history of tree planting in the plains states. Texas Woman's University, Denton, TX

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York

Eldridge DJ, Maestre FT, Maltez-Mouro S, Bowker MA (2012) A global database of shrub encroachment effects on ecosystem structure and functioning. Ecology 93:2499

Gordon DR, Flory SL, Cooper AL, Morris SK (2012) Assessing the invasion risk of *Eucalyptus* in the United States using the Australian Weed Risk Assessment. Int J For Res. doi:10.1155/2012/203768

Gordon DR, Mitterdorfer B, Pheloung P, Ansari S, Buddenhagen C, Chimera C, Daehler C, Dawson W, Denslow J, Jaqualine TN, LaRosa A, Nishida T, Onderdonk, DA, Panetta D, Pyšek P, Randall R, Richardson D, Virtue J, Williams P (2009) Guidance for addressing the Australian Weed Risk Assessment questions. Plant Prot Q 25:56-74

Gordon DR, Onderdonk DA, Fox AM, Stocker RK (2008a) Consistent accuracy of the Australian weed risk assessment system across varied geographies. Divers Distrib 14:234-242

Gordon DR, Onderdonk DA, Fox AM, Stocker RK, Gantz C (2008b) Predicting invasive plants in Florida using the Australian weed risk assessment. Invasive Plant Sci Manag 1:178-195

Grotkopp E, Erskine-Ogden J, Rejmánek M (2010) Assessing potential invasiveness of woody horticultural plant species using seedling growth rate traits. J Appl Ecol 47:1320-1328

Grotkopp E, Rejmánek M, Rost TL (2004) Evolution of genome size in pines (*Pinus*) and its life-history correlates: Supertree analyses. Evolution 58:1705-1729

Harrell FE Jr. (2001) Regression modeling strategies. Springer, New York

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd ed. Springer, New York

Hejda M, Pyšek P, Jarošík V (2009) Impact of invasive plants on the species richness, diversity, and composition of invaded communities. J Ecol 97:393-403

Hui C, Richardson DM, Visser V, Wilson JRU (2014) Macroecology meets invasion ecology: performance of Australian acacias and eucalypts around the world revealed by features of their native ranges. Biol Invasions 16:565-576

Hulme PE (2012) Weed risk assessment: a way forward or a waste of time? J Appl Ecol 49:10-19

Jarošík V (2011) CART and related methods. In: Simberloff D, Rejmánek M (eds) Encyclopedia of biological invasions, University of California Press, Berkeley, pp 104-108

Jefferson L, Havens K, Ault J (2004) Implementing invasive screening procedures: The Chicago Botanic Garden model. Weed Tech 18:1434-1440.

Kaplan H, van Niekerk A, Le Roux JJ, Richardson DM, Wilson JRU (2014) Incorporating risk mapping at multiple spatial scales into eradiation management plans. Biol Invasions 16:691-703

Kapler EJ, Widrlechner MP, Dixon PM, Thompson JR (2012) Performance of five models to predict the naturalization of non-native woody plants in Iowa. J Environ Hort 30:35-41

Keller R, Kocev D, Džeroski S (2011) Trait-based risk assessment for invasive species: High performance across diverse taxonomic groups, geographic ranges and machine learning/statistical tools. Divers Distrib 17:451-461

Keller RP, Lodge D, Finnoff DC (2007) Risk assessment for invasive species produces net bioeconomic benefits. Proc Nat Acad Sci (USA) 104:203-207

Koop AL, Fowler L, Newton LP, Caton BP (2012) Development and validation of a weed screening tool for the United States. Biol Invasions 14:273-294

Kowarik I (1995) Time lags in biological invasions with regard to the success and failure of alien species. In: Pyšek P, Prach K, Rejmánek M, Wade M (eds*)* Plant invasions: General aspects and special problems, SPB Academic Publishing, Amsterdam, pp 15–38

Kriticos DJ, Randall RP (2001) A comparison of systems to analyze potential weed distributions. In: Groves RH, Panetta FD, Virtue JG (eds) Weed risk assessment, CSIRO Publishing, Collingwood, Australia, pp 61-79

Křivánek M, Pyšek P, Jarošík V (2006) Planting history and propagule pressure as predictors of invasion by woody species in a temperate region. Conserv Biol 20:1487-1498

Kueffer C, Pyšek P, Richardson DM (2013) Integrative invasion science: model systems, multi-site studies, focused meta-analysis and invasion syndromes. New Phytol 200:615-633

Larkin DJ (2012) Lengths and correlates of lag phases in upper-Midwest plant invasions. Biol Invasions 14:827-838

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18-22

Lubell JD, Brand MH, Lehrer JM (2008) AFLP identification of *Berberis thunbergii* cultivars, inter-specific hybrids, and their parental species. J Hort Sci Biotech 83:55-63

Maillet J, Lopez-Garcia C (2000) What criteria are relevant for predicting the invasive capacity of a new agricultural weed? The case of invasive American species in France. Weed Res (Oxford) 40:11-26

Marcer A, Pino J, Pons X, Broton, L (2012) Modelling invasive alien species distributions from digital biodiversity atlases. Model upscaling as a means of reconciling data at different scales. Divers Distrib 18:1177-1189

McGregor KF, Watt MS, Hulme PE, Duncan RP (2012) What determines pine naturalization: species traits, climate suitability or forestry use? Divers Distrib 18:1013-1023

Mgidi TN, Le Maitre DC, Schonegevel L, Nel JL, Rouget M, Richardson DM (2007) Alien plant invasions—incorporating emerging invaders in regional prioritization: A pragmatic approach for Southern Africa. J Environm Manage 84:173-187

Myers JH, Bazely D (2003) Ecology and control of introduced plants: Evaluating and responding to invasive plants. Cambridge University Press, Cambridge

Okie WR (1987) Plum rootstocks. In: Rom RC, Carlson RF (eds) Rootstocks for fruit crops, John Wiley, New York, pp 321-360

OIPC (Ohio Invasive Plant Council) (2013) Ohio invasive plant assessment protocol. Accessed online at http://www.oipc.info/AssessmentDocsPublic/Ohio_Invasive_Plant_Assessment_Rev071513.pdf, December 2014

Ou L, Lu C, O'Toole D (2008) A risk-assessment system for alien plant bio-invasion in Xiamne, China. J Environm Sci 20:989-997

Philibert A, Desprez-Loustau ML, Fabre B, Frey P, Halkett F, Husson C, Lung-Escarmant B, Marçais B, Robin C, Vacher C, Makowski D (2011) Predicting invasion success of forest pathogenic fungi from species traits. J Appl Ecol 48:1381-1390

Pimentel D (2011) Biological invasions: Economic and environmental costs of alien plant, animal and microbe species, 2nd ed. CRC Press, New York

Pyšek P, Jarošík V, Hulme PE, Pergl, J, Hejda M, Schaffner U, Vilà M (2012) A global assessment of invasive plant impacts on resident species, communities, and ecosystems: the interaction of impact measures, invading species' traits, and environment. Glob Change Biol 18:1725-1737

Pyšek P, Jarošík V, Pergl J, Moravcová L, Chytrý M, Kühn I (2014) Temperate trees and shrubs as global invaders: the relationship between invasiveness and native distribution depends on biological traits. Biol Invasions 16:577-589

Pyšek P, Manceur AM, Alba C, McGregor KR, Pergl J, Štajerová K, Chytrý M, Danihelka J, Kartesz J, Klimešová J, Lučanová M, Moravcová L, Nishino M, Sádlo J, Suda J, Tichý L, Kühn I (2015) Naturalization of central European plants in North America: species traits, habitats, propagule pressure, residence time. Ecology 96:762-774

R Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rehder A (1947) Manual of cultivated trees and shrubs hardy in North America, 2nd ed. Macmillan, New York

Reichard SH (1994) Assessing the potential of invasiveness in woody plants introduced in North America. Ph.D. Dissertation. University of Washington, Seattle, WA.

Reichard SH, Hamilton CW (1997) Predicting invasions of woody plants introduced into North America. Conserv Biol 11:193-203

Rejmánek M (2011) Invasiveness. In: Simberloff D, Rejmánek M (eds) Encyclopedia of biological invasions. University of California Press, Berkeley, CA, pp 379-385

Rejmánek M (2014) Invasive trees and shrubs: where do they come from and what we should expect in the future? Biol Invasions 16:483-498

Rejmánek M, Richardson DM (1996) What attributes make some plant species more invasive? Ecology 77:1655-1661

Rejmánek M, Richardson, DM (2013) Trees and shrubs as invasive alien species – 2013 update of the global database. Divers Distrib 19:1093-1094

Rejmánek M, Richardson DM, Pyšek P (2013) Plant invasions and invasibility of plant communities. In: van der Marel E, Franklin J (eds) Vegetation ecology. John Wiley, New York, pp 387-424

Richardson DM, Hui C, Nuñez MA, Pauchard A (2014) Tree invasions: patterns, processes, challenges and opportunities. Biol Invasions 16:473-481

Richardson DM, Rejmánek M (2011) Trees and shrubs as invasive alien species – a global review. Divers Distrib 17:788-809

Richardson DM, Thuiller W (2007) Home away from home: Objective mapping of high-risk source areas for plant introductions. Divers Distrib 13:299-312

Rouget M, Richardson DM (2003) Inferring process from pattern in plant invasions: A semi-mechanistic model incorporating propagule pressure and environmental factors. Am Nat 162:713-724

Smith C, Lonsdale W, Fortune J (1999) When to ignore advice: Invasion predictions and decision theory. Biol Invasions 1:89-96

Smith JL, Perino JV (1981) Osage orange (*Maclura pomifera*): History and economic uses. Econ Bot 35:24-41

Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods (4th ed). CRC Press, Boca Raton, FL

Thompson RS, Anderson KH, Bartlein PJ (2000) Atlas of relations between climatic parameters and distributions of important trees and shrubs in North America (2 vols). USGS Professional Paper 1650

Widrlechner MP (1994) Environmental analogs in the search for stress-tolerant landscape plants. J Arboric 20:114-119

Widrlechner MP (2001) The role of environmental analogs in identifying potentially invasive woody plants in Iowa. J Iowa Acad Sci 108:158-165

Widrlechner MP, Iles J (2002) A geographic assessment of the risk of naturalization of non-native woody plants in Iowa. J Environ Hort 20:47-56

Widrlechner MP, Kapler EJ, Dixon PM, Thompson JR (2013) The importance of geographic and biological variables in predicting the naturalization of non-native woody plants in the Upper Midwest. J Environ Hort 31:124-131

Widrlechner MP, Thompson JR, Iles, JK, Dixon PM (2004) Models for predicting the risk of naturalization of non-native woody plants in Iowa. J Environ Hort 22:23-31

Widrlechner MP, Thompson JR, Kapler EJ, Kordecki K, Dixon PM, Gates G (2009) A test of four models to predict the risk of naturalization of non-native woody plants in the Chicago Region. J Environ Hort 27:241-250

Wilson JRU, Richardson DM, Rouget M, Proches S, Amis, MA, Henderson L., Thuiller W (2007) Residence time and potential range: crucial considerations in modeling plant invasions. Divers Distrib 13:11-22

Table 1. Numbers of non-native and naturalizing species for five study areas in the Upper Midwest. Numbers were reduced in the data set for the regional study as a result of species native within the region, but outside of the original study area (see text).

Study area and data set	Total non-native plants (#)	Naturalized non-native plants (#)	Non-naturalized non-native plants (#)
Southern Minnesota			
Original study	94	23	71
Regional study	85	20	65
Iowa			
Original study	129	42	87
Regional study	113	35	78
Northern Missouri			
Original study	126	39	87
Regional study	115	38	77
Chicago A			
Original study	135	38	97
Regional study	120	35	85
Chicago B			
Original study	142	40	102
Regional study	131	37	94
Entire region	231	74	157

Table 2. Comparison of misclassification rates based upon in-sample predictions, out-of-bag estimates from the randomForest function and leave-one-out cross-validation. In all three cases, a species is classified as naturalizing if the predicted P[nat] exceeds the average proportion of naturalizing species in that data set.

Data set	In sample (%)	Out-of-bag (%)	Cross-validation (%)
Regional	24.7	24.7	21.1
Southern Minnesota	4.7	30.6	29.4
Iowa	8.8	24.4	23.9
Northern Missouri	4.3	26.7	27.0
Chicago A	2.5	27.8	30.0
Chicago B	3.8	20.6	20.6

Table 3. Results for classifying species as a potential naturalizer (Y) or non-naturalizer (N), using the proportion of naturalizing species in the data set as the classification threshold. Counts are reported for each combination of local-model result and regional-model result (as local / regional). Percentages of total species are shown for the two cases where local and regional predictions differ. P-values are for the test of whether the local and regional models are equally likely to classify as species as a potential naturalizer, based on McNemar's test for paired binary outcomes.

Data set	Y / Y (#)	Y / N (#)	N / Y (#)	N / N (#)	P-value
Southern Minnesota	29	2 (2.4%)	19 (22%)	35	0.0002
Iowa	33	9 (7.9%)	12 (11%)	59	0.66
Northern Missouri	35	16 (14%)	11 (9.6%)	53	0.44
Chicago A	39	12 (10%)	7 (5.8%)	62	0.36
Chicago B	39	7 (5.3%)	10 (7.6%)	75	0.63

Table 4. False positive and false negative rates for local and regional models, based on two classification thresholds. One threshold is the proportion of naturalizing species in that data set. The second threshold is 10%, which is more conservative in the sense of more likely to classify a species as a naturalizer. The false positive rate (FPR) is the proportion of non-naturalizing species that are mis-classified as potential naturalizers. The false negative rate (FNR) is the proportion of naturalizing species that are misclassified as not naturalizing.

Data set	Threshold (%)	FPR: Local	FPR: Regional	FNR: Local	FNR: Regional
Southern Minnesota	23.5	0.28	0.49	0.35	0.20
Iowa	31.0	0.22	0.27	0.29	0.31
Northern Missouri	33.0	0.29	0.26	0.24	0.32
Chicago A	29.2	0.31	0.21	0.29	0.20
Chicago B	28.2	0.19	0.20	0.24	0.19
Southern Minnesota	10.0	0.49	0.63	0.05	0.10
Iowa	10.0	0.47	0.58	0.057	0.11
Northern Missouri	10.0	0.52	0.61	0.10	0.13
Chicago A	10.0	0.46	0.54	0.11	0.11
Chicago B	10.0	0.43	0.49	0.11	0.16

Table 5. Median out-of-sample error in P[nat] for local and regional models to predict P[nat]. The p-values are based on the Wilcoxon signed-rank tests comparing error magnitudes for the two models.

Data set	Local model error	Regional model error	P-value
Southern Minnesota	0.17	0.24	0.028
Iowa	0.20	0.18	0.39
Northern Missouri	0.22	0.22	0.84
Chicago A	0.18	0.17	0.35
Chicago B	0.13	0.16	0.51

Table 6. Median out-of-sample error rate in P[nat] for local and regional models to predict P[nat], when species are classified by their naturalization status. The p-values are based on the Wilcoxon signed-rank tests comparing error magnitudes for the two models.

Study area	Non-naturalizing			Naturalizing		
	Local Model	Regional Model	P-value	Local Model	Regional Model	P-value
Southern Minnesota	0.08	0.23	< 0.001	0.62	0.26	< 0.001
Iowa	0.08	0.15	0.61	0.43	0.31	0.049
Northern Missouri	0.11	0.17	0.45	0.48	0.31	0.18
Chicago A	0.09	0.12	0.94	0.50	0.32	0.07
Chicago B	0.08	0.09	0.18	0.37	0.28	0.32

Table 7. Species that naturalize in each study area and produce false negatives. The values shown are the probabilities of naturalization for each model that includes the species. Values in (parentheses) are for species that have no record of naturalizing in that area, so that entry is not a false negative. A blank for a local model indicates a species was absent in that model. A small predicted P[nat] is a large error, since a perfect prediction would be P[nat] = 1. Bold text indicates the species generated one of the five largest errors in that model. Species with large errors in the most models are listed first.

	Data set					
Species	IA	Chicago A	Chicago B	MO	MN	Regional
Berberis thunbergii	0.226	**0.128**	**0.004**	**0.047**	**0.122**	0.074
Catalpa speciosa	0.198	**0.098**	**0.061**		0.310	**0.050**
Maclura pomifera	**0.045**	0.153	**0.164**	0.607		**0.068**
Euonymus alatus	0.278	**0.046**	0.268	**0.090**		0.237
Euonymus fortunei		**0.054**	**0.062**	0.160	(0.064)	0.090
Lonicera sempervirens	**0.133**	(0.011)	(0.048)		(0.046)	**0.032**
Philadelphus coronarius	(0.002)	(0.048)	**0.007**		(0.078)	**0.004**
Viburnum opulus	0.647	0.800	0.468	**0.018**	**0.052**	0.762
Acer platanoides	**0.097**	0.412	0.242	(0.044)	0.374	0.382
Acer tataricum	0.269	**0.082**	(0.242)	0.341	0.192	0.210
Ailanthus altissima	**0.133**	0.156	0.249	0.958		0.678
Caragana arborescens		(0.441)	(0.099)		**0.162**	0.654
Koelreuteria paniculata	(0.079)	(0.405)	(0.270)	**0.129**		0.319
Lonicera morrowii	0.351	0.890			**0.153**	0.158
Quercus acutissima		(0.013)		0.398		**0.046**
Rhamnus utilis	**0.158**					0.255
Rosa multiflora	0.593	0.472	0.471	0.520	**0.155**	0.918
Spiraea prunifolia		(0.068)		**0.095**		0.102

Table 8. Species that do not naturalize in at least one study area and produce a false positive. The values shown are the predicted probabilities of naturalization for each model that includes the species. Values in (parentheses) are for species that have a record of naturalizing in that area, so that entry is not a false positive. A blank for a local model indicates a species that was absent in that model. A large predicted P[nat] is a large error, since a perfect prediction would be P[nat] = 0. Bold text indicates that the species generated one of the five largest errors in that model. Species with large errors in the most models are listed first.

	Data set					
Species	IA	Chicago A	Chicago B	MO	MN	Regional
Salix caprea	0.332	**0.878**	**0.845**	0.210	**0.564**	**0.915**
Prunus cerasifera	0.775		**0.642**	**0.851**	0.219	**0.756**
Cotoneaster divaricatus	**0.838**	0.518		**0.934**	0.168	0.590
Rosa rugosa		**0.671**	(0.873)	**0.844**		0.800
Sambucus nigra subsp. *nigra*			**0.787**	0.404		**0.863**
Alnus glutinosa		(0.615)	(0.844)		**0.641**	0.888
Callicarpa dichotoma			0.183	0.559		**0.704**
Castanea mollissima	0.054	0.009		**0.656**		0.541
Cornus alba				**0.688**	0.389	0.815
Corylus avellana		**0.741**		0.288		0.617
Daphne mezereum		**0.545**				0.238
Elaeagnus angustifolia	(0.906)	(0.932)	**0.696**	(0.869)	(0.388)	0.973
Euonymus bungeanus	**0.830**					0.468
Euonymus europaeus		(0.829)			**0.890**	0.578
Larix decidua	0.066	**0.537**	0.184			0.154
Picea abies	0.111	0.313	0.068	0.002	**0.422**	0.197
Prunus avium	**0.934**		(0.392)			0.719
Prunus maackii	**0.948**				0.186	0.247

Quercus robur	0.247	0.456	**0.652**			0.155
Ribes alpinum	0.698		0.326	0.168	**0.615**	0.465
Spiraea trilobata					0.262	**0.728**
Tamarix ramosissima	**0.865**			(0.877)		0.885

Figure 1. Map of five local study areas in the Upper Midwest, USA. In the text, the darkly-shaded area in IL and WI is designated as “Chicago A” and the lightly-shaded area in IN and MI as “Chicago B.”

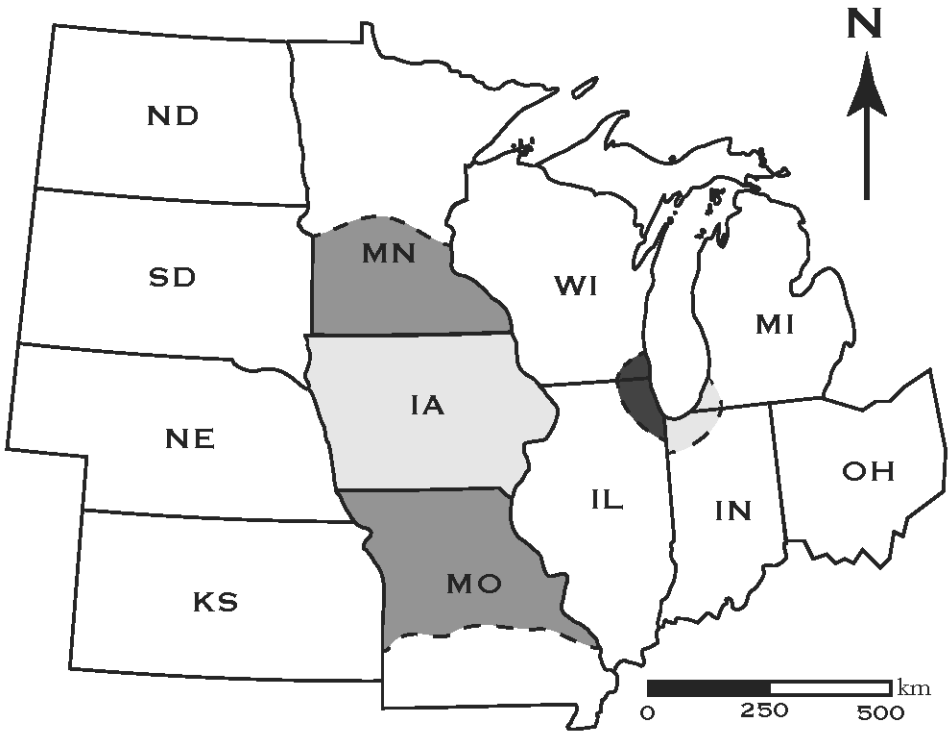


Figure 2. Combined variable importance plot for all five local models and the regional model. Variables are ranked in order of decreasing importance in the regional model. Bars indicate importance in the five local models; horizontal black line is the importance in the regional model. Variables are: G: G-value (continuous), quick: quick maturity, invade: species invades outside North America, fleshy: fruits are fleshy and bird dispersed, fast: species has fast vegetative spread, family: family is invasive in North America, germ: requires germination pretreatment, evergreen, native: species is native to North America.

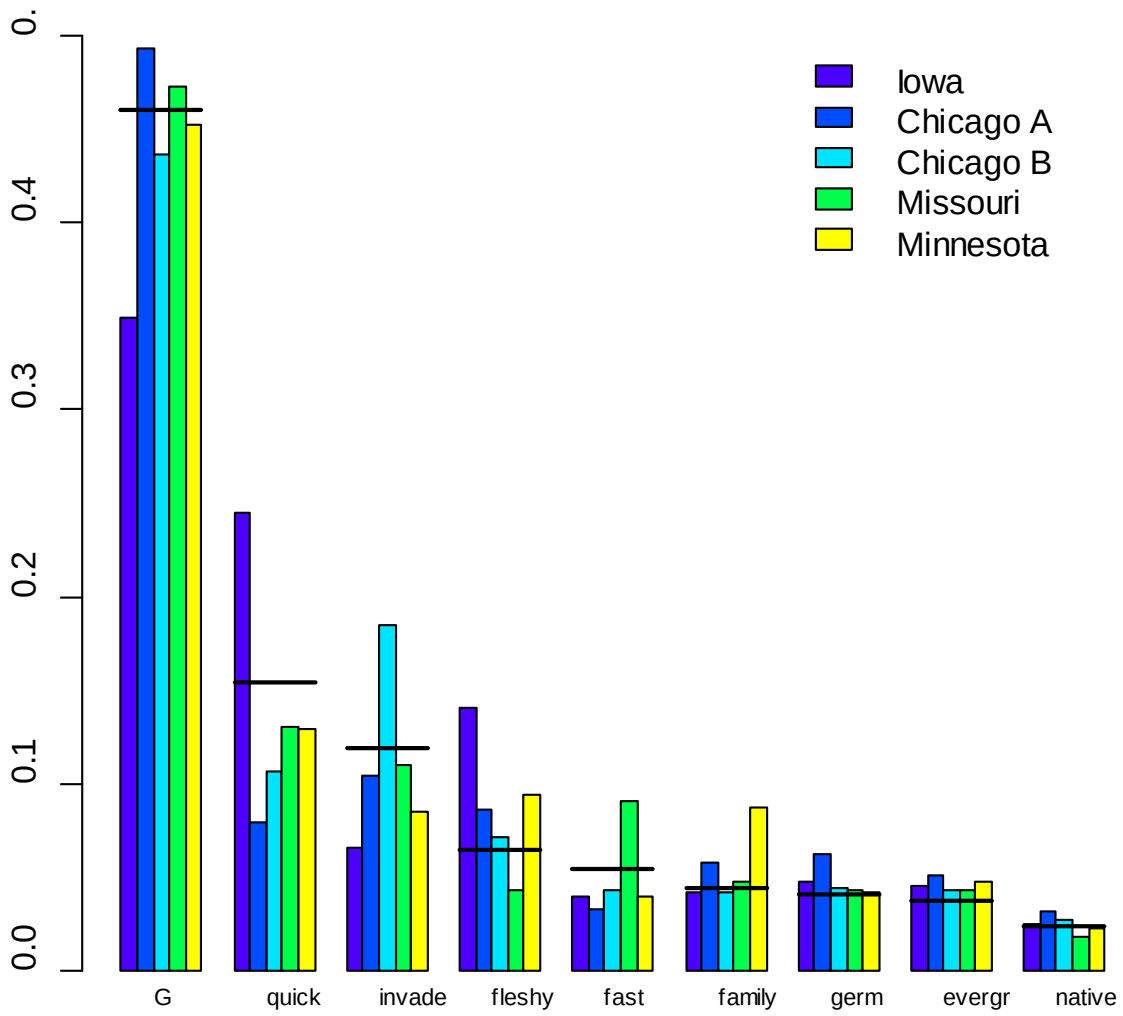
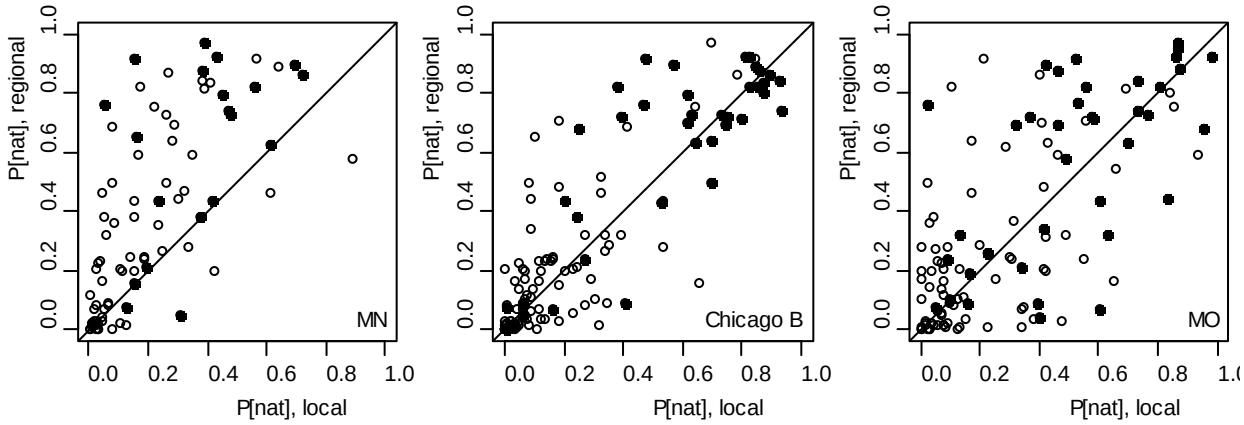


Figure 3. Out-of-sample predicted probability of naturalization, P[nat], from the local random forest model and the regional random forest model for species in Minnnesota (MN), southeast of Chicago, IL (Chicago B), and Missouri (MO). Out-of-sample probabilities are calculated using leave-one-out cross-validation. Filled-in points represent species that naturalize in that area; open points represent species that do not naturalize. Points above the diagonal line are species for which the regional model predicts a greater probability of naturalization; points below the line are species for which the local model predicts a greater probability of naturalization.



Supplemental Appendix:

Table A1. Comparison of predicted probability of naturalization for species that do not naturalize anywhere in the region and "potential naturalizers", species that naturalize somewhere within the region but not in a specific local area. Values are the median predicted probability of naturalization, P[nat], from the regional model for the sets of species included in each study area.

Data set	Do not naturalize	Potential naturalizer
Southern Minnesota	0.20	0.66
Iowa	0.10	0.49
Northern Missouri	0.08	0.33
Chicago A	0.14	0.64
Chicago B	0.11	0.20

Table A2. Comparison of local model predicted probability of naturalization for species that do not naturalizate anywhere in the region and "potential naturalizers", species naturalize somewhere within the region but not in a specific local area. Values are the median predicted probability of naturalization, P[nat], from the local model for the sets of species included in each study area.

Data set	Do not naturalize	Potential naturalizer
Southern Minnesota	0.06	0.28
Iowa	0.07	0.22
Northern Missouri	0.07	0.26
Chicago A	0.10	0.17
Chicago B	0.06	0.36

Figure A1. Out-of-sample predicted probability of naturalization, P[nat], from the local random forest model and the regional random forest model for species in Iowa (IA) and west of Chicago, IL (Chicago A). Out-of-sample probabilities are calculated by using leave-one-out cross-validation. Filled-in points represent species that naturalize in that area; open points represent species that do not naturalize. Points above the diagonal line are species for which the regional model predicts a greater probability of naturalization; points below the line are species for which the local model predicts a greater probability of naturalization.