Weed Science, 53:236–241. 2005

Symposium

Predicting the occurrence of nonindigenous species using environmental and remotely sensed data

Lisa J. Rew
Corresponding author. Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717;
lrew@montana.edu

Bruce D. Maxwell
Department of Land Resources and Environmental Sciences, Montana State University,
Bozeman, MT 59717

Richard Aspinall
Geographic Information and Analysis Center, Montana State University, Bozeman, MT 59717

To manage or control nonindigenous species (NIS), we need to know where they are located in the landscape. However, many natural areas are large, making it unfeasible to inventory the entire area and necessitating surveys to be performed on smaller areas. Provided appropriate survey methods are used, probability of occurrence predictions and maps can be generated for the species and area of interest. The probability maps can then be used to direct further sampling for new populations or patches and to select populations to monitor for the degree of invasiveness and effect of management. NIS occurrence (presence or absence) data were collected during 2001 to 2003 using transects stratified by proximity to rights-of-way in the northern range of Yellowstone National Park. In this study, we evaluate the use of environmental and remotely sensed (LANDSAT Enhanced Thematic Mapper +) data, separately and combined, for developing probability maps of three target NIS occurrence. Canada thistle, dalmation toadflax, and timothy were chosen for this study because of their different dispersal mechanisms and frequencies, 5, 3, and 23%, respectively, in the surveyed area. Data were analyzed using generalized linear regression with logit link, and the best models were selected using Akaike's Information Criterion. Probability of occurrence maps were generated for each target species, and the accuracies of the predictions were assessed with validation data excluded from the model fitting. Frequencies of occurrence of the validation data were calculated and compared with predicted probabilities. Agreement between the observed and predicted probabilities was reasonably accurate and consistent for timothy and dalmation toadflax but less so for Canada thistle.

Nomenclature: Canada thistle, *Cirsium arvense* L. CIRAR; dalmation toadflax, *Linaria dalmatica* (L.) P. Mill. LINDA; timothy, *Phleum pratense* L. PHLPR.

Key words: Generalized linear model, invasive species, logistic regression, nonnative species, predictive mapping, survey, stratified sampling.

Considerable resources are directed toward the management of nonindigenous species (NIS), and obtaining information on their location is important. However, if NIS are relatively infrequent and spread over large areas, financial and logistical constraints will make it impossible to locate and manage all populations. Thus, small portions of the total management area are generally sampled (surveyed). If such data are collected using an unbiased survey design, in which data on NIS occurrence and possibly associated variables are recorded, the data can be used to produce probability maps of species occurrence for the areas which were not surveyed (Franklin 1995; Guisan and Zimmerman 2000; Shafii et al. 2003). Such probability of occurrence maps would help land managers to decide where to send crews to search for additional NIS populations. The survey data also could be used to select populations or patches to monitor from the range of environments within which the target species exists. The relative invasiveness and the potential effects of populations in the different environments can then be evaluated. These monitoring results would serve to prioritize management of populations in the environments that pose the greatest threat to the ecosystem.

Many countries have designated specific areas to be maintained as "wilderness" or "natural areas" for recreational or wildlife benefit, or both. Exactly how management of these wildlands is defined obviously varies, but in many cases it is linked to maintaining flora and fauna at a level observed before settlement by Europeans or at least the early 1900s. For example, the National Park Service has a mandate to maintain natural areas under their jurisdiction as unaltered by human activities as possible (National Park Service 1996). Thus, considerable effort is extended to the management of NIS, particularly plant species.

Disturbance is often suggested as a key factor enhancing the probability of nonindigenous plant establishment in plant communities (Grime 1979). Natural disturbance has a variety of biotic and geomorphic causes including soil disturbance by fauna, weather-related events such as drought, floods, wind, fire, and geological events such as landslides. In most areas of the United States, the natural fire regime has been altered; so, fire should be considered a quasihuman disturbance. Human disturbance also includes construction and use of roads and trails, buildings, utility corridors, and campgrounds. Anthropogenic disturbances such as roads or trails (Gelbard and Belnap 2003; Parendes and Jones 2000; Trombulak and Frissell 2000; Tyser and Worley 1992; Watkins et al. 2003), cultivation, grazing, trampling, and domestic ungulates (Mack and Thompson 1982; Tyser and

TABLE 1. Coefficient values for the best fit combined variable model for Canada thistle, dalmation toadflax, and timothy for Data Subset 1 (n = 42,317).

Coefficients	Canada thistle	Dalmation toadflax	Timothy
Intercept	−7.92551	3.12522	2.00790
Proximity to road (m)	0.00019	−0.00017	0.00016
Proximity to trail (m)	0.00010	−0.00065	0.00025
Elevation (m)	0.00033	−0.00249	−0.00308
Cosine of aspect (°)	−0.18760	−0.39421	−0.22335
Sine of aspect (°)	0.66578	−0.30232	0.17488
Presence of wildfire (binary)	0.67134	−1.32851	0.65564
Slope (°)	0.02320	0.03210	0.00336
Solar insolation (Wh m^{-2})	0.00022	−0.00007	—
LANDSAT ETM[a] Band 1	0.03656	−0.04306	0.02932
LANDSAT ETM Band 2	0.05022	0.05954	0.04364
LANDSAT ETM Band 3	−0.07421	—	−0.08256
LANDSAT ETM Band 4	−0.01688	—	0.02936
LANDSAT ETM Band 5	0.00764	−0.03073	0.00831
LANDSAT ETM Band 7	−0.01699	0.01859	−0.01336
Isocluster class	0.01768	0.01423	0.00388

[a] Abbreviation: LANDSAT ETM+ LANDSAT Enhanced Thematic Mapper+.

Key 1988; Young et al. 1972) are often considered to have more effect on the occurrence of NIS than natural disturbances.

If the occurrence of a target species is known to be correlated with a particular variable, one could stratify the sampling scheme on that variable and improve the probability of finding the target (Hirzel and Guisan 2002). In this study of NIS in the northern range of Yellowstone National Park, we accepted the assumption that human disturbance in the form of rights-of-way (ROW) increases the chance of finding NIS and stratified our sampling using this variable, but sampled away from this disturbance to generate an unbiased data set. The aim of this study was to generate predictive maps of target NIS occurrence using generalized linear models for the entire area of interest—the northern range of Yellowstone National Park. To generate a predictive map requires that the independent variable data are available for the entire area of interest. To achieve this, we used environmental data obtained from digital elevation maps and reflectance data from LANDSAT Enhanced Thematic Mapper (ETM)+ imagery. The influence of environmental and reflectance data on the occurrence of target NIS was assessed, and the benefit of using the environmental and reflectance data, independently or combined, to improve model fit was evaluated. The accuracy of the resultant probability of occurrence predictions and maps was evaluated for three target species.

Materials and Methods

Yellowstone National Park covers an area of 899,121 ha predominantly in Wyoming, United States. A total of 187 nonindigenous plant species have been recorded within the Park, which comprises 15% of the total plant species (Whipple 2001). This study concentrates on the area within the northern elk winter range of the Park (152,785 ha). Sixty-two NIS were targeted by this study, but we are only reporting on three of those species here.

A stratified sampling approach was used to collect field data. Transects were stratified on ROW, which include roads and trails in this instance. Field sampling was performed from early June to late August in 2001 to 2003. During the 3 yr, a total of 305 transects were completed, most of which were 2,000 m in length, although some were shorter if the terrain proved impassable. All transects were 10 m wide. The total area surveyed was 53 ha, representing 0.035% of the study area.

Transect start locations were randomly allocated on ROW in a geographical information system (GIS), before commencing field work. In 2001, the start position of each transect was randomly located on a ROW but ran 2,000 m perpendicular to ROW from that point. This approach needed to be partially modified for subsequent years to provide a more similar number of data points at all distances from ROW. In 2002 and 2003, the start locations of transects were still randomly generated but fit the following set

TABLE 2. Best model fits for Canada thistle, dalmation toadflax, and timothy using seven remotely sensed (LANDSAT ETM+)[a] and eight environmental data variables, combined and independently, for Data Subset 1 (n = 42,317). Akaike's Information Criterion values of the best fit models are provided with number of variables retained in the best model in parentheses.

Target species	All variables (15)	LANDSAT ETM+ variables (7)	Environmental variables (8)
Canada thistle	14,657.42 (15)	16,066.92 (7)	14,768.24 (7)
Dalmation toadflax	9,513.46 (13)	11,293.14 (6)	9,789.77 (7)
Timothy	38,388.81 (14)	40,702.72 (7)	42,956.91 (7)

[a] Abbreviation: LANDSAT ETM+ LANDSAT Enhanced Thematic Mapper+.

TABLE 3. Best model fits for Canada thistle, dalmation toadflax, and timothy using all 15 independent variable data (reflectance and environmental data variables) for Data Subsets 1 to 3 (n = 42,317). Akaike's Information Criterion values of the best fit models are provided with number of variables retained in the best model in parentheses.

Target species	Subset 1	Subset 2	Subset 3
Canada thistle	14,657.42 (15)	14,763.96 (15)	14,914.79 (15)
Dalmation toadflax	9,513.46 (13)	9,575.50 (13)	9,575.65 (15)
Timothy	38,388.81 (14)	38,528.05 (13)	38,799.90 (14)

of confines: starting on a road and finishing 2,000 m from all roads but at all times traversing more than 2,000 m from any known trail; starting on a trail and finishing 2,000 m from all trails but at all times traversing more than 2,000 m from any known road; and starting on a road or trail and finishing 2,000 m from all ROW.

Transects were walked and location and other data recorded with a Global Positioning System (GPS) by two-person teams. Trimble Pro XR and GeoExplorer3® GPS receivers[1] were used, and the data were differentially postprocessed to improve positional accuracy (mean horizontal precision was 1.5 m). The coordinate system and projection used was Universal Transverse Mercator (UTM) Zone 12N, WGS 1984 Datum. Along each transect when a target NIS was intersected, the length of the patch was recorded in the GPS data dictionary. Additional location data were also recorded along each transect. All these data were used to generate continuous NIS data using extensions we created in Arcview[2] (Version 3.2) and an Excel[3] macro. The continuous data were generated at 10- by 10-m resolution.

Environmental and remote sensing data were used as independent variables. To generate predictive NIS maps of the entire area of interest, we need to have variable information of the entire area. Therefore, we used the environmental data from digital elevation maps (10-m resolution) and remote sensing data (30-m resolution). The environmental data including aspect, elevation, slope, and solar insolation were calculated from 10-m resolution digital elevation map; distance from roads and trails were calculated from data layers within the GIS database. Solar insolation was calculated for the summer months using only Swift's method (Swift 1976). LANDSAT ETM+ remote sensing data, acquired July 13, 1999, were included as individual spectral bands and as an unsupervised classification layer. The unsupervised classification layer was generated using ISOCLUSTER in ERDAS Imagine,[4] and 128 classes were identified. These classes were used by Legleiter et al. (2003) to develop a land-cover map of the Yellowstone watershed with accuracies of between 63 and 100% for individual land-cover classes. The 128 individual ISOCLUSTER classes were used in this analysis. The 30-m resolution Bands 1 to 5 and 7 of the LANDSAT ETM+ data were pan-sharpened to 15-m resolution with the panchromatic data from LANDSAT ETM+ Band 8 and resampled to 10-m resolution using nearest neighbor resampling so that the resolution of the LANDSAT ETM+ data matched the resolution of the digital elevation model available for the study area.

All these data layers were queried at 10-m intervals along the continuous sampling transects, and the sample values stored in the transect attribute database in Arcview. Thus, the final data set contained presence and absence points for 28 NIS, eight environmental variables (aspect was transformed into cosine and sine of aspect), six LANDSAT ETM+ bands, and one unsupervised classification layer, at 52,896 locations. Twenty percent of the data (n = 10,579) were randomly selected from the main data set and set aside to validate the accuracy of the probability models and maps. This random selection was performed thrice to produce Data Subsets 1 to 3, which contained the majority of the data (n = 42,317).

The three subsets of data were analyzed with generalized linear regression models, with binomial distribution and logit link in S-PLUS 2000.[5] Generalized logistic models (GLM) were used because the dependent variable—the NIS species data—is binary (presence and absence) data. The best model was determined with backward stepwise procedure using Akaike's Information Criterion (AIC), where the change in AIC value between models is used to define the "best" model, with the lowest AIC value representing the best model fit (Akaike 1977; Burnham and Anderson 1998). In our analysis, we determined three best models for each of the data subsets, using the AIC value for model selection. The three best models were selected using the reflectance data and environmental data variables, separately and combined. This was to determine if only one type of data were available—i.e., environmental or reflectance, which would make a better model, and how do those models compare with models from the combined data. All the analyses were performed in S-PLUS 2000.

Probability of occurrence predictions and maps of the target species were generated using coefficient values from the GLM applied to continuous spatial variables in rasterized format, using an extension we wrote in Arcview. The extension generated the logit of the GLM by summing the product of each variable in the model and its coefficient value, plus the beta-intercept value. The value of each cell in the output raster, ranging from zero to one, represents the probability that the target species could be present within the area defined by that cell. In this study, the raster cell size was 10 by 10 m. The validation data points, which were not used in the GLM, were overlaid on the appropriate probability maps in the GIS. At each validation data point, the predicted probability value was recorded and collated into 10 probability classes. The frequencies of occurrence were then calculated for the associated validation data, for each target species.

Three target species were chosen for the analysis with GLM and development of probability of occurrence maps. These were: Canada thistle, a wind-dispersed species with rhizomatous growth; dalmation toadflax, a non–wind dispersed rhizomatous species; and timothy, a non–wind dispersed nonrhizomatous species.

Results and Discussion

The frequency of Canada thistle was 5%, dalmation toadflax 3%, and timothy 23% within the area surveyed. Al-

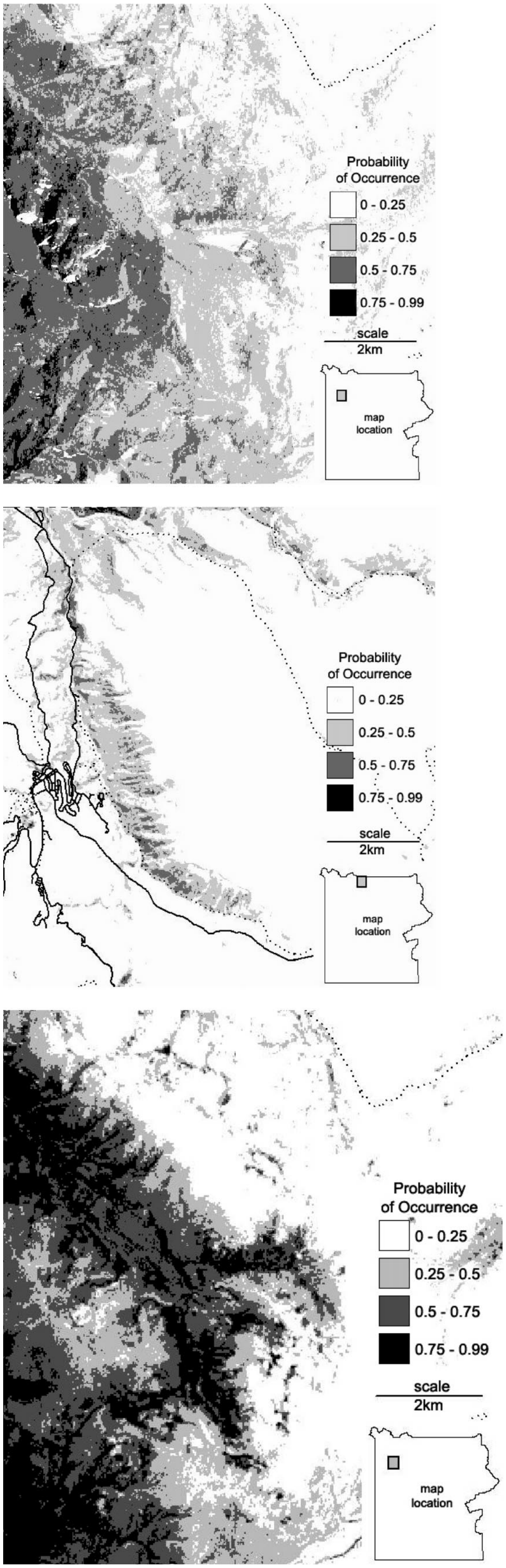

FIGURE 1. Predicted probability of occurrence maps for (a) Canada thistle, (b) dalmation toadflax, and (c) timothy for selected areas of the northern range of Yellowstone National Park. Solid lines represent roads; dashed lines represent trails.

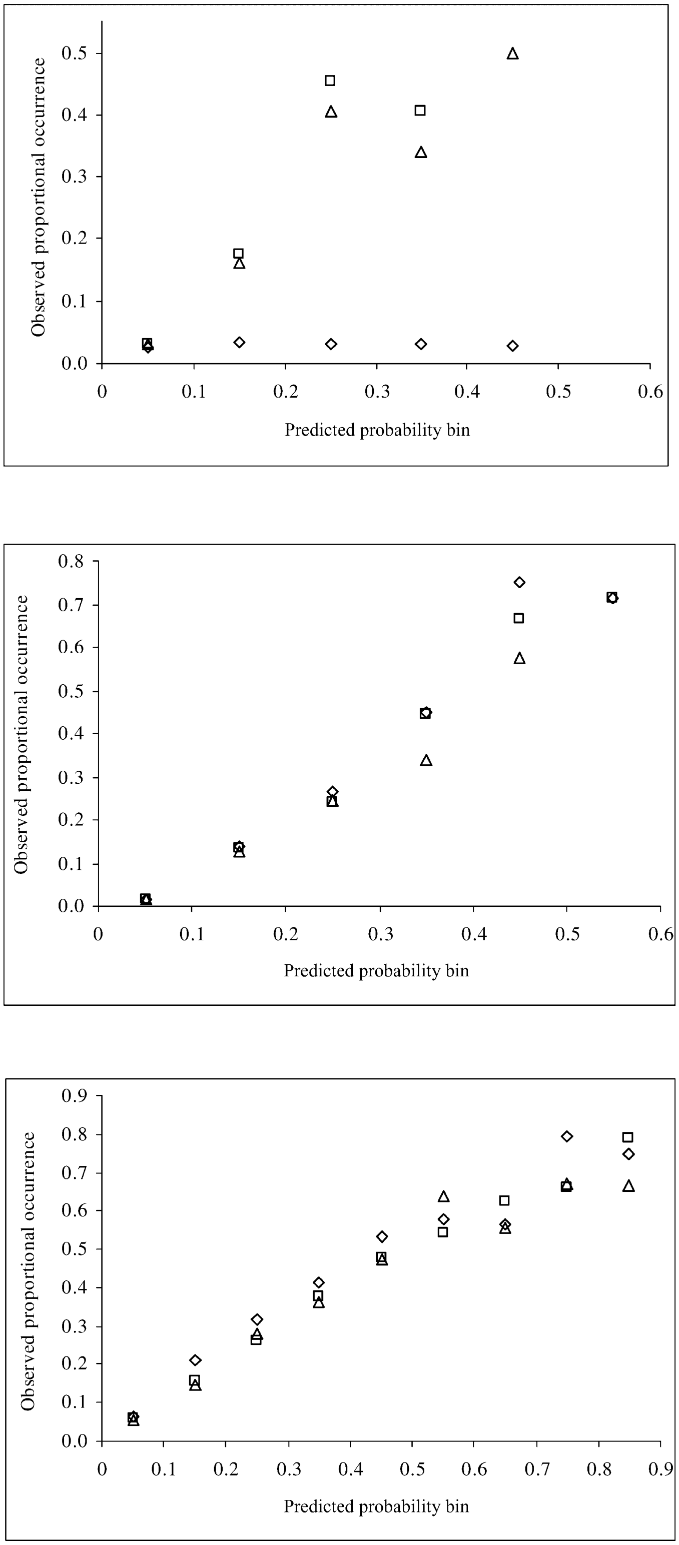

FIGURE 2. Observed frequency of occurrence of the validation data plotted against the predicted probability values, collated into 10 classes for (a) Canada thistle, (b) dalmation toadflax, and (c) timothy. ◇, □, and △ represent validation Data Sets 1, 2, and 3, respectively.

though these values are of interest, they provide no information to improve our understanding of where the species occurred on the landscape. Analyzing the binary NIS data using GLM provides some indication of the environmental variables that are associated with the occurrence of target NIS.

The occurrence of Canada thistle, dalmation toadflax, and timothy was correlated with most of the environmental variables and the reflectance measurements of the remote sensing bands, but the importance of the independent variables differed for the three target species (Table 1). This demonstrates that the occurrence of the target species is driven by numerous environmental parameters, but none of the target species had very specific associations with any one of the variables measured.

If only one type of predictor variable, either the environmental or remotely sensed data, were fit with a GLM, the environmental data produced a better model for Canada thistle and dalmation toadflax occurrence than the LANDSAT ETM+ data, whereas the converse was true for timothy (Table 2). Selecting the best model fit from all the available variables always provided a better model than environmental or reflectance variables separately (Table 2—only results from Subset 1 shown). However, the number of variables retained in the best model differed according to the subset of data used, and this was reflected in the different AIC values (Table 3).

Predictive maps of Canada thistle, dalmation toadflax, and timothy were generated from the best models for each data set, which happened to be Subset 1. Examples of approximately 10- by 10-km areas are provided for display purposes (Figures 1a–c); these smaller areas provide better observation of the probability maps than those of the entire area. The validation data sets were then used to evaluate the agreement between the predicted probabilities and the observed frequencies of occurrence. For example, if 200 validation points were recorded in probability class 0.1 to 0.2, we would expect on average 30 presences and 170 absences; in the probability class 0.6 to 0.7 we would expect 130 presences and 70 absences, etc. Agreement between the validation data and the probability predictions was better at the lower than at the higher occurrence probabilities for each of the target species (Figure 2) because too few of the validation data were located within the higher probability classes. And, because the model is predicting locations where the target species is more or less likely to establish and survive, although it may not have arrived there yet. Agreement between the observed and predicted data was good for timothy, particularly in the first six classes, with more variability in the agreement for the next three probability classes (0.6 to 0.7, 0.7 to 0.8, and 0.8 to 0.9); insufficient validation data were recorded in the 0.9 to 1 classes for comparison. Observed vs. predicted agreement of the dalmation toadflax data was good; there was more variation between the validation data sets for the 0.3 to 0.4 and 0.4 to 0.5 classes (Figure 2). Insufficient occurrence data were located in the higher probability classes (more than 0.6). Variation between the validation data sets was greatest for Canada thistle, with Validation Sets 2 and 3 providing similar results to each other but different to Validation Set 1 (Figure 2). Canada thistle model performance was poor for the lower probability classes, and insufficient validation data were available for probabilities greater than 0.5 (Figures 2a–c). This was expected on the basis of the high-residual sum of squares for Canada thistle compared with the other two species.

Conclusions

Many management areas are too large to sample entirely, so developing predictive maps of species occurrence provides information on conditions that are conducive for that species. However, in order for such predictions to be accurate, it is important that the survey methods used to collect the data on which the probability maps are based are unbiased and sample the environmental conditions present in the study area (Hirzel and Guisan 2002). Sampling may be randomly stratified on variables or gradients that are believed to be associated with a species distribution (Hirzel and Guisan 2002). However, stratifying on a number of variables becomes more complex as the number of target species increases because each species may have a different response to individual and multiple variables (Maggini et al. 2002). Therefore, Hirzel and Guisan (2002) suggest that unless correlations between the target species and variables are well known, sampling equally, not proportionally, within the multiple variables would probably be most effective. Because relationships between NIS occurrence and environmental variables are poorly understood, and the knowledge we do have suggests that species respond differently, we stratified on the one variable which is known to be important, proximity to ROW, and extended transects 2,000 m from ROW to provide the best possibility of sampling all environments equally. Analysis of the data using GLM with logit link provided information on target species correlations with environmental and reflectance data variables. Output from these models produced good predictions, particularly for dalmation toadflax and timothy, and we believe that the approach shows potential and will be validated further. Accurate probability maps of species occurrence could be used by land managers to prioritize where to spend the limited resources available for managing NIS in wildland and rangeland areas.

Sources of Materials

[1] Trimble Pro XR and GeoExplorer3® GPS receivers, Trimble Navigation Limited, 749 North Mary Avenue, Sunnyvale, CA 94085.

[2] Arcview (Version 3.2) ESRI Inc., 380 New York Street, Redlands, CA 92373-8100.

[3] Excel (Microsoft Excel 2002®), Microsoft Corporation, 1 Microsoft Way, Redmond, WA 98052-6399.

[4] ERDAS Imagine, Leica Geosystems GIS & Mapping, LLC, Worldwide Headquarters, 2801 Buford Highway, N.E., Atlanta, GA 30329-2137.

[5] S-PLUS 2000: Mathsoft Inc., 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109.

Acknowledgments

This project was partially funded by the Yellowstone Inventory and Monitoring program. We thank Frank Dougher for writing Arcview extensions and collating the field data; the field crew including Judit Barroso, Jerad Corbin, Matthew Hulbert, Jeremy Gay, Mara Johnson, Rebecca Kennedy, Erik Lehnoff, Ben Levy, Amanda Morrison, Nathaniel Ohler, Charles Repath, John Rose, and Kit Sawyer; and Ann Rodman and the GIS lab in Yellowstone National Park for sharing their GIS data layers.

Literature Cited

Akaike, H. 1977. Likelihood of a model and information criteria. J. Econom. 16:3–14.

Burnham, K. P. and D. R. Anderson. 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. New York: Springer-Verlag. 353 p.

Franklin, J. 1995. Predictive vegetation mapping: geographical modelling of biospatial patterns in relation to environmental gradients. Prog. Phys. Geogr. 19:474–499.

Gelbard, J. L. and J. Belnap. 2003. Roads as conduits for exotic plant invasions in a semiarid landscape. Conserv. Biol. 17:420–432.

Grime, J. P. 1979. Plant Strategies and Vegetation Processes. Chichester, UK: J. Wiley. 222 p.

Guisan, A. and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135:147–186.

Hirzel, A. and A. Guisan. 2002. Which is the optimal sampling strategy for habitat suitability modeling? Ecol. Model. 157:331–341.

Hitchcock, C. L. and A. Cronquist. 2001. Flora of the Pacific Northwest. 12th ed. Seattle, WA: University of Washington Press. 730 p.

Legleiter, C. J., R. L. Lawrence, M. A. Fonstad, W. A. Marcus, and R. J. Aspinall. 2003. Fluvial response a decade after wildfire in the northern Yellowstone ecosystem: a spatially explicit analysis. Geomorphology 54:119–136.

Mack, R. N. and J. N. Thompson. 1982. Evolution in steppe with few, large, hooved mammals. Am. Nat. 119:757–773.

Maggini, R., A. Guisan, and D. Cherix. 2002. A stratified approach for modelling the distribution of a threatened ant species in the Swiss National Park. Biodivers. Conserv. 11:2117–2141.

National Park Service. 1996. Preserving our Natural Heritage—A Strategic Plan for Managing Invasive Non-Indigenous Plants on National Park System Lands. www.nature.nps.gov/biology/invasivespecies/strat_pl.htm.

Parendes, L. A. and J. A. Jones. 2000. Role of light availability and dispersal in exotic plant invasion along roads and streams in the H. J. Andrews Experimental Forest, Oregon. Conserv. Biol. 14:64–75.

Shafii, B., W. J. Price, T. S. Prather, L. W. Lass, and D. C. Thill. 2003. Predicting the likelihood of yellow starthistle (*Centaurea solstitialis*) occurrence using landscape characteristics. Weed Sci. 52:748–751.

Swift, L. W. 1976. Algorithm for solar radiation on mountain slopes. Water Res. 12:108–112.

Trombulak, S. C. and C. A. Frissell. 2000. Review of ecological effects of roads on terrestrial and aquatic communities. Conserv. Biol. 14:18–30.

Tyser, R. W. and C. H. Key. 1988. Spotted knapweed in natural area fescue grasslands: an ecological assessment. Northwest Sci. 62:151–160.

Tyser, R. W. and C. A. Worley. 1992. Alien flora in grasslands adjacent to road and trail corridors in Glacier National Park, Montana (U.S.A.). Conserv. Biol. 6:253–262.

Watkins, R. Z., J. Chen, J. Pickens, and K. D. Brosofske. 2003. Effects of forest roads on understory plants in a managed hardwood landscape. Conserv. Biol. 17:411–419.

Whipple, J. J. 2001. Annotated checklist of exotic vascular plants in Yellowstone National Park. West. N. Am. Nat. 61:336–346.

Young, J. A., R. A. Evan, and J. Major. 1972. Alien plants in the Great Basin. J. Range Manag. 25:194–201.

Received May 20, 2004, and approved November 16, 2004.