**Evaluation of Habitat Data as Correlates of Bird Community Metrics and Species Occurrence**

Lloyd W. Morrison

Quantitative Ecologist,

Heartland Inventory and Monitoring Network

National Park Service

Final version

9 April 2007

BACKGROUND

The purpose of these analyses is to evaluate the value of habitat data collected in conjunction with bird community monitoring at several Heartland Inventory and Monitoring Network parks. Different habitat types are recognized at each park, and it was desired to have the data analyzed separately for each park/habitat type (hereafter “PHT”). Annual sample sizes for most PHTs are relatively small. Only at the TAPR and AGFO upland habitats were annual sample sizes > 20. Because a rotating panel is employed at the TAPR upland PHT, the total sample size is larger for this PHT. Data were not pooled across years for analyses involving this PHT, however, as inter-annual variability would likely bias the results. A total of 32 habitat variables were included for evaluation. Data were collected for all variables in each PHT (in most years), even if the habitat of interest was not present (i.e., only 0’s were recorded).

|  |  |
| --- | --- |
| PHT | Annual Sample Size |
| AGFO-riparian | 14 |
| AGFO-upland | 13-40 |
| HEHO-upland | 9 |
| HOCU-riparian | 5 |
| HOCU-upland | 20 |
| HOCU-edge | 2 |
| TAPR-riparian | 16-18 |
| TAPR-upland | 79-242 |

# EVALUATION OF INDIVIDUAL HABITAT VARIABLES

I evaluated all habitat variables made available to me (n=32) for AGFO and TAPR. Separate evaluations were made for riparian corridor and upland habitats. Five years of data were available for both parks and habitat types. In-depth analyses were conducted for the first and last years (2001 and 2006) for each data set. If results differed dramatically between the two years, the intervening years were examined.

Analyses consisted of examining the data range and distribution of each habitat variable for each park, habitat type, and year separately. Habitat variables were then classified according to their likely information content in correlational analyses with bird species data.

Definitions of groupings:

**Uninformative:** Data consisted of all or mostly 0’s for the variable of interest. In other words, that particular habitat type wasn’t present at any or at most sites.

**Questionable:** Distribution of data highly skewed, many small or large values, or an otherwise narrow data range; transformations necessary for any analyses requiring a normal distribution of data. May only be informative as predictors of occurrence (e.g., logistic regression) rather than abundance. **Or** this variable may meet the definition of potentially informative in some years and be considered uninformative in others, not a consistently useful variable.

**Potentially Informative:** Good data range and distribution near normal or likely to be transformed to normal or near normal.

It should be noted that a number of variables appeared to be reasonably informative in some years and not in others. This may be partially due to natural inter-annual variability, and in part to the fact that many of the sites sampled each year were different. Whatever the reason, this aspect of the data set needs to be emphasized. Although I have made an effort to summarize the usefulness of each variable generally, prior to conducting in-depth analyses with specific habitat variables it would be wise to examine each year of data.

Note: HVPB = Horizontal Vegetation Profile Board

AGFO--Riparian Corridor

Uninformative

Canopy height

Canopy cover

Basal area

Conifer litter

Rock

Woody debris

Moss & lichen

Shrubs & vines

Tree seedlings

Warm season grass

HVPB 5-m;1.25m

HVPB 5-m;1.50m

HVPB 5-m;1.75m

HVPB 5-m;2.00m

HVPB 15-m;1.25m

HVPB 15-m;1.50m

HVPB 15-m;1.75m

HVPB 15-m;2.00m

Questionable

Deciduous litter

HVPB 5-m;1.00m

HVPB 15-m;1.00m

Potentially Informative

Grass litter

Bare soil

Unvegetated

Cool season grass

Forbs

Total foliar cover

HVPB 5-m;0.50m

HVPB 5-m;0.75m

HVPB 15-m;0.50m

HVPB 15-m;0.75m

H

AGFO--Upland

Uninformative

Canopy height

Canopy cover

Basal area

Deciduous litter

Conifer litter

Woody debris

Tree seedlings

HVPB 5-m;1.00m

HVPB 5-m;1.25m

HVPB 5-m;1.50m

HVPB 5-m;1.75m

HVPB 5-m;2.00m

HVPB 15-m;1.00m

HVPB 15-m;1.25m

HVPB 15-m;1.50m

HVPB 15-m;1.75m

HVPB 15-m;2.00m

Questionable

Rock

Unvegetated

Moss & lichen

Shrubs & vines

Warm season grass

HVPB 5-m;0.75m

HVPB 15-m;0.75m

H

Potentially Informative

Grass litter

Bare soil

Cool season grass

Forbs

Total foliar cover

HVPB 5-m;0.50m

HVPB 15-m;0.50m

TAPR--Riparian Corridor

Uninformative

Conifer litter

Questionable

Rock

Woody debris

Moss & lichen

Shrubs & vines

Tree seedlings

Warm season grass

HVPB 5-m;0.50m

HVPB 5-m;0.75m

HVPB 5-m;1.25m

HVPB 5-m;1.50m

HVPB 5-m;1.75m

HVPB 5-m;2.00m

HVPB 15-m;0.50m

HVPB 15-m;0.75m

HVPB 15-m;1.25m

HVPB 15-m;1.50m

HVPB 15-m;1.75m

HVPB 15-m;2.00m

Potentially Informative

Canopy height

Canopy cover

Basal area

Deciduous litter

Grass litter

Bare soil

Unvegetated

Cool season grass

Forbs

Total foliar cover

HVPB 5-m;1.00m

HVPB 15-m;1.00m

H

TAPR--Upland

Uninformative

Canopy height

Canopy cover

Basal area

Conifer litter

Woody debris

Tree seedlings

HVPB 5-m;1.25m

HVPB 5-m;1.50m

HVPB 5-m;1.75m

HVPB 5-m;2.00m

HVPB 15-m;1.25m

HVPB 15-m;1.50m

HVPB 15-m;1.75m

HVPB 15-m;2.00m

Questionable

Deciduous litter

Grass litter

Rock

Unvegetated

Cool season grass

Moss & lichen

Shrubs & vines

HVPB 5-m;0.75m

HVPB 5-m;1.00m

HVPB 15-m;0.75m

HVPB 15-m;1.00m

H

Potentially Informative

Bare soil

Forbs

Total foliar cover

Warm season grass

HVPB 5-m;0.50m

HVPB 15-m;0.50m

# 

# ANALYSES

The large number of predictor variables and small sample size extremely limits the statistical approaches that can be employed (at least that could include most or all habitat variables). Stepwise multiple linear or logistic regression analyses are frequently used as model-building approaches when the goal is to select a small set of informative predictors from a large number of potentially intercorrelated variables. Factor analyses are often employed prior to such analyses to reduce a large number of predictor variables to a smaller set of derived, uncorrelated variables. A combination of factor analysis followed by stepwise multiple linear regression, and stepwise multiple logistic regression was applied to the bird data to evaluate habitat correlates. Minimum sample sizes necessary for these types of analyses precluded their application to most PHTs.

|  |  |  |
| --- | --- | --- |
| **Method** | **Minimum sample size** | **Reference** |
| Multiple linear regression analyses | 10-20 for each predictor variable | Harrell 2001 |
| Multiple logistic regression analyses | At least 10 EPV, where EPV = the smaller of the counts for the values of the binary variable/the number of predictor variables | Peduzzi et al. 1996 |
| Factor analysis | Depends on various things; 10 per variable to 300 overall | MacCallum et al. 1999 |

In interpretation of the following analyses, it should be kept in mind that all such sequential (i.e., stepwise) methods suffer from compromised Type I error rates, as many variables are considered for inclusion simultaneously. Also, it must be remembered that correlation is not causation. It has been demonstrated using synthetic data that “significant” models may be constructed from data sets in which there are no functional relationships between the response and predictor variables (Mac Nally 2000).

## Determinants of bird community metrics

Multivariate approach

The first approach was to focus on a metric of overall bird communities, and species richness was used as the response variable. It was desirable to evaluate as many predictor variables in the models as possible. A multiple linear regression analysis could not be employed on the raw data because of sample size limitations. A factor analysis was run for the TAPR upland PHT in 2006. The sample size was too small for this analysis according to information published in the literature (although within the range of analyses in other published studies).

The statistical algorithm used to extract factors was a principal components analysis. This method forms linear combinations from the observed variables, resulting in a smaller number of derived variables. The first principle component is the combination that accounts for the largest amount of variance in the sample. Only the components with eigenvalues greater than one were retained. This resulted in five principal components that explained a cumulative variance of 77.5%. An orthogonal rotation was performed to increase ease of interpretation.

The five principal components were then used as predictor variables in a multiple linear regression analysis with bird species number as the response variable. The sample size was entirely within the specified limits for this type of analysis. A forward stepwise selection process was used, with a default criteria of *p* = 0.05 for variable inclusion and *p* = 0.10 for variable removal. None of the principal components was selected as a significant predictor.

Univariate approach

The next step was to evaluate all pairwise correlations of habitat and bird community variables. A Spearman correlation coefficient was used as most habitat variables did not appear to come from a normal distribution. This was done for the AGFO upland and riparian and TAPR upland and riparian PHTs in 2006. Surprisingly, this analysis revealed very few “significant” correlations among species richness and any of the habitat variables:

Significant habitat correlates of bird species richness:

TAPR upland 2006: (n=1) moss and lichen, r = 0.227

TAPR riparian 2006: (n=1) cool season grass, r = 0.487

AGFO upland 2006: (n=1) forbs, r = -0.485

AGFO riparian 2006: (n=8) total foliar cover, r=0.538; 7 HVPB variables, all negative correlations, highest r = -0.649

Similar results were obtained with diversity as the response variable (diversity and species richness were highly positively correlated; r always > 0.95).

Because so many multiple comparisons were made simultaneously, the probability of a type I error is no longer 0.05. Some sort of correction for multiple comparisons should be made or, alternatively, this analysis could be viewed as “descriptive” rather than inferential, and potentially biologically important variables identified based on the size of the correlation coefficients. Comparisons of *p*-values among PHTs should not be done, as the PHTs have different sample sizes, and *p*-values are dependent upon sample size.

Conclusions from these analyses:

(1) The sample sizes will be too small for multiple linear regression analyses or factor analyses (except possibly for upland sites at TAPR when only a small number of predictor variables are included).

(2) Very few “significant” correlations of bird community metrics with habitat variables were observed by pairwise comparisons, and few of them were very strong. Correcting for multiple comparisons will likely result in no “significant” habitat correlates for most PHTs.

(3) Simple pairwise comparisons may be the best way of evaluating habitat correlates in terms of community level metrics, at least in a descriptive manner.

Determinants of species occurrence

I evaluated species occurrence patterns at AGFO and TAPR with a stepwise multiple logistic regression. A forward stepwise selection process was employed, using the score statistic for evaluating variables for entry into the model, and the conditional statistic for evaluating variables for removal.

A list of birds from each PHT was specified as follows (D. Peitz, email correspondence):

AGFO: GRSP, LASP, MODO, RWBL, and WEME.

TAPR: BARS, BHCO, BRTH, CONI, DICK, EAKI, GRSP, KILL, LASP, RWBL, UPSA, and WEME.

A list of species was also provided for other PHTs, although small sample sizes precluded multiple logistic regression analyses.

It was indicated that the “potentially informative” habitat variables determined earlier should be included in these analyses, as well as the “questionable” variables, to see how the resulting models might be influenced (D. Peitz, email correspondence). Thus two sets of models were constructed for each PHT, one including only the “potentially informative” habitat variables, and one including the “potentially informative” habitat variables and the “questionable” variables. Analyses were run for upland sites at TAPR and AGFO in 2006. The habitat variables [HVPB 5-m;0.50m] and [HVPB 15-m;0.50m] were among the “potentially informative” variables for both PHTs, and the remaining HVPB variables designated as “questionable” were not included in the analyses, as the HVPB variables were in general highly intercorrelated.

It is difficult to obtain an informative model for species present at only a few sites or species present at almost all sites. Thus the following species were removed from consideration because they were too rare: BARS, BRTH, EAKI, and LASP at TAPR; or too common: WEME at TAPR and AGFO.

Total sample sizes were clearly too small for such a large number of predictor variables, and the resulting models are not expected to be stable. The output, however, allows a univariate examination of each of the independent variables, which is informative and recommended as the first step in any logistic model building process (Hosmer and Lemeshow 2000). The coefficients for each predictor variable considered alone will not vary with the total number of predictor variables considered.

Summary measures for evaluating the goodness of fit for a logistic multiple regression include the Cox Snell R2 and Nagelkerke R2. The interpretation of these measures is not the same as the coefficient of determination in a linear regression however, and in general they are difficult to interpret (Norusis 2003). The percent of correct classification is another summary measure that is often used, but it is actually a poor indicator of model fit. It is possible to add highly significant variables to the model, yet have the correct classification rate decline, as was observed in some cases with this data set.

Results:

The interpretation of the logistic regression models is that the probability of a bird species being present at a site is 1/(1+e-Z), where Z is equal to the equations given for each species.

At AGFO, including only “potentially informative” habitat variables (n=7), the results were:

RWBL: -1.960 + 0.145 forbs

No significant predictors were found for LASP, MODO, or GRSP.

At AGFO, including both “potentially informative” and “questionable” habitat variables (n=12; additional HPVB variables were not included since HVPB 5-m; 0.50m and HVPB 15-m; 0.50m were included and were highly correlated with the other HPVB variables), the results were:

GRSP: -0.662 + 0.183 warm season grass

RWBL: 11.147 + 0.183 forbs – 9.957 H

No significant predictors were found for LASP or MODO.

At TAPR in 2006, including only “potentially informative” habitat variables (n=6), the results were:

DICK: -2.821 + 0.047 HVPB 15-m; 0.50m

GRSP: -0.235 + 0.043 warm season grass

KILL: 0.865 – 0.051 HVPB 5-m; 0.50m

No significant predictors were found for BARS, BHCO, CONI, RWBL, or UPSA.

At TAPR, including both “potentially informative” and “questionable” habitat variables (n=14; additional HPVB variables were not included since HVPB 5-m; 0.50m and HVPB 15-m; 0.50m were included and were highly correlated with the other HPVB variables), the results were:

DICK: -12.374 + 0.053 HVPB 15-m; 0.50m + 0.108 unvegetated + 0.096 grass litter - 0.117 forbs

GRSP: -0.363 – 0.129 cool season grass + 0.037 total foliar cover

KILL: 0.865 – 0.051 HVPB 5-m;0.50m

RWBL: -1.182 + 0.046 cool season grass

UPSA: 0.637 - 0.062 cool season grass

No significant predictors were found for BARS, BHCO, or CONI.

Note that for both AGFO and TAPR, the selection of predictor variables in the models was dependent upon the set of variables entered into the model building process.

This analysis was also conducted for TAPR upland sites in 2001, as an assessment of model validation. Including only “potentially informative” habitat variables (n=6), the results were:

DICK: -11.385 + 0.098 HVPB 15-m; 0.50m + 0.035 Bare Soil + 0.017 HVPB 5-m;0.50m

GRSP: 0.072 + 0.032 warm season grass

RWBL: -4.248 – 0.133 warm season grass + 0.066 total foliar cover

UPSA: -3.00 + 0.038 bare soil

No significant predictors were found for BARS, BHCO, CONI, or KILL.

Comparing these results to the same PHT in 2006, with only the “potentially informative” habitat variables (n=6) used in both analyses, it is obvious that different models were the result. Some species had significant models in one year but not in the other. This indicates these models in general are not stable, have very low predictive value, and are not applicable to areas of the park that were not included in the respective sample.

Conclusions from these analyses:

(1) The sample sizes will be too small for multiple logistic regression analyses (except possibly for upland sites at TAPR when only a small number of predictor variables are included).

(2) The models presented here are not the only, nor necessarily the “best” models for a given species. Many different options for model building are available, and the options selected, along with the habitat variables included, can potentially produce many different models. Any given model building algorithm cannot guarantee a “best” model in a statistical sense, and the choice of a model will ultimately depend upon factors such as objectives of the study, ease of variable acquisition, and interpretation (Norusis 2003).

(3) The resulting models are likely to vary among years (i.e., panel of sample sites), indicating low predictive value. This is a result of a small sample size and may also result from interannual variability, heterogeneity across different sampling sites, or problems with the sampling design.

(4) The most informative use of multiple logistic regression analyses is likely to be a focused study on one or a few species, rather than trying to construct models for a relatively large number of species in each PHT. The most informative model building will result from careful evaluation of all habit variables based on the natural history of the species and the issues associated with each PHT. If certain habitat variables are known to be important for the species of interest, they should be included in the model regardless of their significance level (Norusis 2003). Competing models should be evaluated (both statistically and practically) and the most informative model or models determined. This process should include evaluation of model diagnostics (not done here) and model validation.

# SUMMARY CONCLUSIONS AND RECOMMENDATIONS

There are other, more sophisticated methods that could be applied to this data, although the large number of potential predictor variables coupled with the relatively small sample sizes will be a problem with any multivariate analyses. With any model building approach, the results of the model will be dependent upon the subset of predictor variables chosen for evaluation.

Overall conclusions

Assuming the five-minute surveys accurately depict the bird species composition of a given site, the general lack of basic explanatory power for the habitat variables may be due to two potential sources:

(1) Because parks have already been divided into different habitat types, there may be too much homogeneity within each habitat type. The data for a number of habitat variables exhibit a narrow range of values, supporting this hypothesis.

(2) The territory or home range of the bird species may be larger than the area sampled for habitat, or have a low degree of overlap with the area sampled for habitat (i.e., habitat data are sampled out to 50m from the plot center, but birds are observed out to 100 m). If the scale of the birds’ territory or home range is not similar to the scale over which habitat is measured, or if the two do not overlap to a large degree, the habitat data may not be representative of the bird species in question. This is likely to be a larger problem for some species than others.

Recommendations

I have the following overall recommendations:

(1) Reduce the collection of habitat data and reallocate the time to collecting more bird data, by sampling more sites when possible. The habitat data, as currently collected, do not appear to be very useful as correlates of overall bird community metrics or consistent predictors of species occurrence. Collecting habitat data from multiple subplots at a site probably does not yield more accurate information than collecting habitat data from a single plot.

(2) Instead of collecting a wide variety of generalized habitat data, focus on a suite of habitat variables likely to be important to each of a small group of target bird species. This will require knowledge of the natural history of the species involved, as well as the management issues and goals of each park or PHT with reference to the bird species. Such an approach seems necessary for informative model building.

(3) The most informative use of the data as currently collected may be the estimation of habitat parameters on a PHT scale, and evaluation of change in these parameters over time. It may be possible to correlate overall long-term changes in the abundance of target bird species (within a PHT) with changes in certain habitat variables.

# REFERENCES

Harrell, F.E. (2001) *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York and Berlin, Springer.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied logistic regression*. New York, John Wiley and Sons.

Mac Nally, R. (2000) Regression and model-building in conservation biology, biogeography and ecology: The distinction between--and reconciliation of--'predictive' and 'explanatory' models. *Biodiversity and Conservation* **9**: 655-671.

MacCallum, R.C., Widaman, K.F., Zhang, S. and Hong, S. (1999) Sample size in factor analysis. *Psychological Methods* **4**: 84-99.

Norusis, M.J. (2003) *SPSS 12.0 Statistical Procedures Companion*. Upper Saddle River, New Jersey, Prentice Hall.

Peduzzi, P.N., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A. (1996) A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* **99**: 1373-1379.