Paper 199-27

Identifying Plant Species: A Botanical Analysis Using PROC DISCRIM

Robert G. Downer, Louisiana State University, Baton Rouge, LA
Philip E. Hyatt, U.S. Forest Service, Pineville, LA

ABSTRACT

The genus Carex (Cyperaceae) causes headaches for professional botanists. The visual similarity of these plants often results in a dependence on reproductive characteristics for identification. The species Carex retroflexa Willd. and Carex texensis (Torr.) L. H. Bailey are two species which have often been viewed as one in the past However, in this very general paper, the distinctiveness of the two species is demonstrated through the use of PROC DISCRIM and other SAS® procedures.

INTRODUCTION

The genus Carex (Cyperaceae) is a mystery to professional botanists. There are 484 species and 634 taxa (including varieties and subspecies) in North America, north of Mexico (Kartesz 1999). In the eastern United States, the number of taxa by state generally increases as one proceeds from the south to the north (Michigan has 195 taxa) or up the Appalachian mountain chain (Virginia has 188 taxa). Many species have similar vegetative characteristics and hence identification is a problem.

C. texensis and C. retroflexa are species (within the Carex genus) that have previously been difficult to differentiate. Geographically C. retroflexa is more abundant in more dry locations (such as the hills of the Ozarks) but the general habitats for the two species do overlap. Both can be found along streams and disturbed settings such as lawns, roadsides and parks in all areas of the Unites States.

The purpose of this research was to investigate whether certain plant variables could help differentiate between these two species. Exploratory analysis revealed some distinct differences and PROC DISCRIM confirmed the findings.

DATA

Field experience suggested that a key for identification may be possible (Hyatt 1998). Plants were borrowed from several collections around the country. From sixty of these plants, data on six several characteristics were taken: perigynium length, spongy layer length, width of the widest leaf, length of the longest bract and height of the tallest culm. Each flower contains periygynia, a packet that enclose the seeds of the plant and a perigynium has a soft spongy layer at the bottom. The flowering stems are called culms and bracts are leaves at the base of the spikelet of flowers. (See drawings in Mackenzie (1940)). The average and median for each plant were calculated for each characteristic as input to the analysis but results did not differ and hence the mean was used for simplicity.

Discriminant Analysis

In discriminant analysis (Johnson and Wichern, 1998), a criterion is developed to classify an observation into one of the possible groups or populations. The two assumed groups in our case are the classifications of C. retroflexa and C. texensis. The sample means, variances and correlations of the input variables are used to define a measure of squared statistical distance Dj^2 for each observation. For classification group j, the distance

$$Dj^2 = (\mathbf{x} - m_j)'S^{-1}(\mathbf{x} - m_j),$$

is defined using: **x**, the set of input variables (perigynia width, spongy layer length etc.), m_j is the set of sample means for these variables for group j and S^{-1} is the inverse of the matrix containing their variances and correlations. It is usually assumed to be common for each group. The input variables are each assumed to be approximately normal but the technique is generally quite robust to this assumption.

Prior to observing the plant characteristic variables for the two species we initially assume the probability of classification into either species is 0.5 and then the probability of the observation being from group one as opposed to group two after observing **x** becomes $\exp(-0.5D^2_1(\mathbf{x}))/(\exp(-0.5D^2_1(\mathbf{x}) + \exp(-0.5D^2_2(\mathbf{x})))$. The probability of falling into group two is 1 minus this probability. The higher of the two estimated probabilities defines the classification for an observation. In cross-validation prediction of plant i, the sample means and their sample correlations in the discriminant function are computed using all plants other than observation i, and hence the value of the discriminant criterion changes for each plant.

PROC DISCRIM

As well as classifying observations into groups, the PROC DISCRIM function in SAS will output means, standard deviations and correlations of the input variables.

The estimated discriminant function (from the calibration set) can be used to classify a new set of observations.

Some Relevant Syntax:

```
PROC DISCRIM DATA = calibration CROSSVALIDATE
OUTCROSS = resultset ;
CLASS group;
BY ….;
PRIORS….;
VAR ; ....
```

The input data set is typically a calibration set to establish the discriminant function. The OUTCROSS option gives the resulting classification results for the entire data set. Prediction of each observation is done through discriminant modeling without this observation. The group or population variable is indicated in the CLASS statement. All variables in the input data set are assumed to be used as predictors in the analysis unless otherwise specified in the VAR statement. By default, the prior probability of classification into a particular group (without input into the estimated discriminant function) are 0.5 but this specification can be changed through the PRIORS statement

DATA ANALYSIS

PROC UNIVARIATE was used to analyze the observed distributions of each of the variables. The minimum, maximum, twenty-fifth percentile (Q1), median (Med) and seventy-fifth percentile (Q3) are given for each of the variables in Tables 1a and 1b below. This exploratory analysis was quite revealing and immediately suggested that an identification key based on these characteristics could be developed.

Table 1a Summary of C. retroflexa characteristics

Characteristic	*C. retroflexa*				
	Min	Q1	Med	Q3	Max
Perigynium width	1.34	1.65	1.71	1.83	2.19
Spongy Layer Length	1.17	1.46	1.59	1.71	1.94
Perigynium Length	2.65	3.06	3.17	3.29	3.56
Longest Bract Length	0	0.85	1.75	3.41	7.80
Width Widest Leaf	1.30	1.76	2.03	2.30	3.40
Length Long Culm	18.20	35.38	42.30	52.08	81.30

Table 1b Summary of C. texensis characteristics

Characteristic	*C. texensis*				
	Min	Q1	Med	Q3	Max
Perigynium width	0.98	1.09	1.15	1.20	1.25
Spongy Layer Length	0.91	0.99	1.04	1.09	1.25
Perigynium Length	2.80	3.04	3.10	3.20	3.36
Longest Bract Length	0	1.25	1.93	3.13	7.20
Width Widest Leaf	1.10	1.35	1.63	1.79	2.50
Length Long Culm	18.10	28.52	33.80	40.43	54.70

From the summary information of Tables1a and 1b, one can see that the two species differed most for the measurements of the perigynium width and the length of the spongy layer (while considerable species overlap occurs in the distribution of each of the other variables) .

The entire C. texensis perigynium width distribution is less than the minimum of the corresponding C. retroflexa distribution. The third quartile of the C. texensis spongy layer length distribution is less than the minimum of the corresponding C. retroflexa distribution. In fact, only three out of thirty C. texensis specimens had an average spongy layer length which exceeded the minimum sample average (1.17) of the C. retroflexa.

Hence, exploratory analysis of these characteristics suggested:

If the average perigynium width is less than 1.3 mm, one should classifiy as C. texensis .

If the average spongy layer length is less than 1.1 mm then one should classify as C. texensis

PROC DISCRIM Results

A discriminant analysis with all variables as predictors was run as follows:

```
Title2 ‘Prediction with all variables’;
proc discrim data = avgdat  crossvalida;te
outcross = avgcvout;
class specid;
run;
```

Output of the cross-validation prediction for this analysis is given below:

```
               Prediction with all variables
               Cross-validation Summary
               Number of Observations and
               Percent Classified into specid


specid         retr          texe         Total


 retr              30             0            30
              100.00          0.00        100.00
 texe               1            29            30
                3.33         96.67        100.00
 Total             31            29            60
               51.67         48.33        100.00

               Priors           0.5           0.5


               Error Count Estimates for specid

               retr         texe         Total
Rate         0.0000       0.0333       0.0167
```

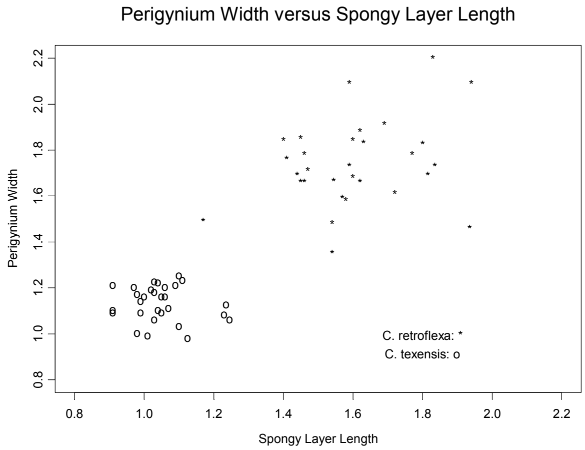
The rows of the outputted table represent the true species id and the columns represent the predicted classification. The total in the (1,1) or (2,2) positions of the table are correct classifications. Hence, 30 out of 30 retroflexa plants were correctly classified using all predictor variables and 29 out of 30 texensis were classified correctly.

.As suggested by the observed variable percentiles, the average perigynimum width by itself (rather than all the predictor variables) was equally as successful in a linear discriminant function. However the single classification error occurred with the other species (retorflexa) as displayed below.

```
Title2 ‘Prediction with average pwidth only ’;
proc discrim data = avgdat  crossvalidate
outcross = avgcvout2;
class specid;
var avgpwid;
run;
```

```
               Prediction with average pwidth only
               Number of Observations,
               Percent Classified into specid

      specid              retr            texe           Total

    retr                  29               1               30
                          96.67            3.33          100.00
    texe                   0              30               30
                           0.00          100.00          100.00

    Total                 29              31               60
                          48.33           51.67          100.00

      Priors              0.5             0.5


               Error Count Estimates for specid
                          retr           texe        Total
Rate                    0.0333         0.0000        0.0167
```

The exploratory analysis also revealed a distinct separation for the observed distributions of the average spongy layer length. A linear discriminant function with this variable as the only variable for prediction was almost as successful as the average perigynium width but one classification error occurred for each species. The distribution separation and classification success of these two variables suggests plotting these variables against each other as an additional visual aid to species identification. The single classification error of a true C. retroflexa by the discriminant function with only average perigynium width as a predictor is the point in between the two clusters at (1.17,1.48)



Logistic regression is an alternative technique for this binary outcome (the two species). A linear regression model is fit with the log-odds of one species over another as the response. PROC LOGISTIC was used to fit this model. In a model with all available variables included as predictors, only perigynium width was significant (p = .0107) and the same cross-validation prediction results were obtained as in the PROC DISCRIM analysis with that single predictor.

SUMMARY

A simple key to identification has been suggested through this analysis. Exploratory work using PROC UNIVARIATE revealed distinct separation in the observed distribution of the average perigynium width and average spongy layer length. Perigynium width and spongy layer length are generally much less for C. texensis and this leads to the following distinct recommendations:

If the average perigynium width is less than 1.3 mm, one should classifiy as C. texensis .

If the average spongy layer length is less than 1.1 mm then one should classify as C. texensis

The classification technique of discriminant analysis verified the original exploratory findings and PROC DISCRIM was successful in correctly classifying 29 out of 30 C. retroflexa and 30 out of 30 C. texensis using only the average perigynium width. Use of all available variables in the discriminant function was equally as successful. A similar analysis using logistic regression through PROC LOGISTIC gave identical results.

REFERENCES

Hyatt, P.E. (1998), “Arkansas Carex (Cyperaceae): a briefly annotated list,” *SIDA,* 18, 535-554.

Johnson, R. A. and Wichern, D.W. (1998), *Applied Multivariate Statistical Analysis, Fourth Edition*. Prentice Hall, Upper Saddle River, NJ.

Kartesz, J.T. (1999), “A synomized checklist and atlas with biological attributes for the vascular flora of the United States, Canada, and Greenland,” In: Kartesz, J.T. and Meacham, C.A. *Synthesis of the North American Flora, Version 1.0.* North Carolina Botanical Garden: Chapel Hill, NC.

Mackenzie, K. K. (1940*),* *North American Cariceae*. New York Botanical Garden, New York, N. Y., 2 volumes.

SAS Institute, Inc. (1999), *SAS/STAT User’s Guide Version 8*, Cary, NC: SAS Institute, Inc.

ACKNOWLEDGMENTS

The authors would like to sincerely thank Dr. Anton A. Reznicek at the University of Michigan for his work with many of the specimens.

CONTACT INFORMATION

Contact the authors at:

Robert G. Downer
Dept. Experimental Statistics
161 Agricultural Administration Building
Baton Rouge, LA 70803-5606
(225) 578-8373
rdowner@lsu.edu

Philip E. Hyatt
U.S. Forest Service
2500 Shreveport Hwy
Pineville, LA 71360
(318) 473-7262
phyatt@fs.fed.us