NOTES AND CORRESPONDENCE

Comments on "Determining the Relative Frequency of Occurrence of Local Cumulonimbus Activity through Discriminant Analysis"

HARRY R. GLAHN

Techniques Development Laboratory, National Weather Service, NOAA, Silver Spring, Md. 20910

1 September 1977

It is good to see studies that produce operationally useable objective techniques such as that by Randerson (1977), and I hope nothing I will say will imply his predictions are not useful. However, I believe some aspects of the study should be highlighted for the unwary reader.

Randerson makes somewhat of a point of choosing predictors by using the 95% level of the F distribution. Many screening regression programs (whether they are "forward selection," "backward elimination," or a combination) use the F test as a stopping procedure just as Randerson has done.[1] However, one should be aware, and state for his readers, that it is just that—a stopping procedure—and not state or imply that any probability level can be attached to it.

There are various reasons why exact—or even approximate—probability levels are inappropriate in this and similar studies. The most important of these is the fact screening is being done.[2] Consider that 100 possible predictors are available. If each of these were tested for its correlation with a predictand, one would expect to find that about five of the 100 were "significant" at the 95% level even if the predictors were drawn from a set of random numbers. Now, when forward screening is done the first predictor picked is the one which has the highest correlation with the predictand, which is also the one that has the highest F ratio associated with it. So, one should be surprised if the "best" predictor did *not* show significance at the 95% level, even if the predictors were random time series.

For $N=500$ independent cases (Randerson used 1079 days, not necessarily independent), 40 uncorrelated available predictors (he used, as nearly as I can tell, 57 correlated predictors), the (tabled) 95% F value is about 3.86 for a one-predictor regression equation. Monte Carlo simulation (Zurndorfer and Glahn, 1977) shows that to test the *best* of the 40 an F value of about 9.48 should be used. If the predictors are highly correlated, the value is lower; in this case a critical value of 7.0 is reasonable (about twice the tabled value).

This brings up another knotty problem—what level of significance should be used? There are two types of error in significance testing. One, called Type I error, is judging a result to be significant when it really isn't (rejection of a true null hypothesis). The other, called Type II error, is not judging a result to be significant when it really is (failure to reject a false null hypothesis). In many applications, one wants the probability of Type I error (α-region) to be low, say 5%; therefore, the 95% F value is used. However, if one insists on a low probability of Type I error, he may make a Type II error quite frequently. I am of the opinion that when one is deciding whether or not to put another predictor into a regression equation that is going to be used for prediction, an α-region closer to 50% than 5% is more appropriate, provided the operational use of this added predictor presents no more problems than the previous one. (By α-region, I mean a region that actually corresponds to an α probability, not a tabled $1-\alpha$ F value.)

An α-region of 50% implies a considerably smaller critical F value than an α-region of 5%. For the Monte Carlo simulation described above, the critical F value is about 5.8 for uncorrelated predictors and 3.2 for highly correlated predictors. The same argument holds for the second, third, etc., predictor selection, although the difference between tabled F values and a value appropriate for screening would undoubtedly not be as large.

Therefore, the error of using tabled F-values when a predictor has been selected as "best," and the "error" of using the tabled F-value corresponding to a small α-region when a much larger region is more

[2] Autocorrelation of variables and nonnormal distributions are sometimes serious problems.

[1] The F value is calculated as the ratio of mean squares not as the ratio of sum of squares as Randerson states.

appropriate, partially compensate each other; the result is a workable stopping procedure for a screening program!

The above discussion concerns forward selection. Similar arguments can be made concerning the backward elimination Randerson used. However, even less is known about testing the significance of predictors "dropped" than of predictors "selected."

To summarize, in order not to mislead readers, one should not imply that such a "stopping procedure" relates to a particular significance level unless such a statement has some backing such as Monte Carlo simulations.

A second point concerns the use of both regression and two-category discriminant analysis. First, Randerson used the total sample of 1079 days and chose by screening the "best" predictors for predicting the number of hours of Cb activity. Then he used those predictors in a two-category discriminant analysis program that would accept only 600 of the original 1079 cases to determine a linear discriminant function [his Eq. (1)]. The relative frequency of a Cb day was then found for various values of the discriminant function (his Fig. 2). The reason for this two-step process was not mentioned.

It seems not to be generally recognized that the single linear discriminant function for two groups is equivalent to the linear regression equation for the same predictors and those same two groups, where group membership is indicated by some arbitrary constant value for one group and some other arbitrary constant value for the other group. Also, if the constant for group 1 (the occurrence of an event) is 1 and the constant for group 2 (the nonoccurrence of the event) is 0, the resulting equation can be used directly for estimating the probability of the event, given specific values of the predictors (using care not to use predictor values outside the data region in the dependent sample). The regression formulation specifies the predictor coefficients and the constant exactly; however, the discriminant formulation does not provide a control for the variance of the discriminant function. For that reason the coefficients in the discriminant function will usually only be *proportional* to the corresponding regression coefficients.

However, either the regression equation or the discriminant equation can be used to develop a correspondence between equation value and event probability as Randerson has done. In fact, the *same* curve should result if all cases are used in the analysis.

I am of the opinion that some meteorologists attribute more "magic" to discriminant analysis than to regression for two groups. That seems to be the case here. Probably a screening regression program was available but not a screening discriminant analysis program. So the former was used to select predictors and those predictors were input to the "magical" discriminant analysis program, even though the latter program could accommodate less than 60% of the data. Actually, the selection of predictors by the regression program was not optimum for the discriminant program because the predictand was hours of Cb days rather than a binary variable denoting group membership.[3] However, if a Cb day is denoted by a 1 (or any constant) and a non-Cb day denoted by 0 (or any other constant), the screening program will pick predictors which are optimal (to the degree that the screening procedure will pick optimal predictors) for either regression or discriminant analysis.

Therefore, use of a discriminant analysis program is not necessary. Screening regression will give the "best" (again, to the degree that screening will pick optimal predictors) equation either for direct use as a prediction equation or (usually better) in the development of a figure such as Randerson's Fig. 2.

Given a developmental data sample, two or more objective techniques will many times yield similar results. Also, a good (poor) predictor for number of Cb hours would likely be a good (poor) predictor for group membership. Therefore, I would not expect the results to be much different from those presented by Randerson if this data sample were processed the way I suggest. The point is, why go to the extra effort of a two-step process when a one-step seems more appropriate?

I am puzzled by another aspect of the study. The list of potential predictors included the K, the Total Totals and the SWEAT indices. Each of these variables is a linear combination of some of the basic predictors—temperature and dew point at 850, 700 and 500 mb. In backward elimination, an inversion (or what is equivalent to an inversion) of the complete predictor cross product matrix is necessary. This matrix will not invert (will be singular) if one predictor is a linear combination of (some of) the others. Here, it seems, the three indices are linear combinations of the other predictors. Why was there not a problem with the computer program? Perhaps roundoff error allowed a solution.

Finally, if Fig. 2 was actually prepared from the biased sample of 600 rather than from the complete sample, I should think the probabilities derived from it would be quite biased toward high values: 470 (61%) of the 770 nonoccurrence cases were evidently ignored;

[3] This is the statement made in the paper—"Using the number of hours of Cb activity as the dependent variable . . ." However, one of the references (Quiring, 1974) states that screening runs were made in which the predictand was binary (thunderstorm or no thunderstorm day) as well as runs with the quasi-continuous number of hours of Cb activity; it is possible Randerson is discussing the former rather than the latter. In fact the Quiring reference strongly suggests this to be the case [in the Appendix, compare the predictors selected in column $Y(1)$ of the first table and those selected in column $XY'(1)$ of the second table with the predictors in Randerson's Eq. (1)].

only 9 (3%) of the 309 occurrence cases were not used. Quiring (1974) presents relative frequency curves prepared from the complete sample that bear this out.

REFERENCES

Quiring, R. F., 1974: Comments on the Randerson *Z*- and *U*-Index thunderstorm prediction scheme. ARLV-351-42, Air Resources Laboratory, Las Vegas, 9 pp.

Randerson, D., 1977: Determining the relative frequency of occurrence of local cumulonimbus activity through discriminant analysis. *Mon. Wea. Rev.*, 105, 709–712.

Zurndorfer, E. A., and H. R. Glahn, 1977: Significance testing of regression equations developed by screening regression. *Preprints 5th Conf. on Probability and Statistics*, Las Vegas, Amer. Meteor. Soc., 95–100.