On the Notion of "Best Predictors:" An Application to Tornado Prediction

Caren Marzban[1,2,3] *, E. DeWayne Mitchell[1,2], and Gregory J. Stumpf[1,2]

[1] National Severe Storms Laboratory, Norman, OK 73069
[2] Cooperative Institute for Mesoscale and Meteorological Studies,
[3] Department of Physics, University of Oklahoma, Norman, OK 73019

*Corresponding author's e-mail: marzban@nssl.noaa.gov

Abstract

It is argued that the strength of a predictor is an ill-defined concept. At best, it is contingent on many assumptions, and, at worst, it is an ambiguous quantity. It is shown that many of the contingencies are met (or avoided) only in a bivariate sense, i.e., one independent variable (and one dependent variable) at a time. Several such methods are offered after which data produced by the National Severe Storms Laboratory's Tornado Detection Algorithm are analyzed for the purpose of addressing the question of which storm-scale vortex attributes based on Doppler radar constitute the "best predictors" of tornadoes.

1 Introduction

In statistical model building one is often faced with the task of reducing the number of predictors. The reason for this is, usually, to preclude an "information overload" either for an a posteriori statistical analysis or for the benefit of the prospective user. As an example of the former situation, consider a data set consisting of a number of predictors whose number exceeds the sample size of the data. Such a data set is inadequate for statistical model building because the model [1] is apt to overfit such a data set. Overfitting generally refers to the situation in which a model (e.g., regression, discriminant analysis) performs well on the data set employed for parameter estimation, but performs poorly on an independent data set. In fact, it can occur even when the number of predictors is less than the number of cases; it occurs because the model has more parameters than can be uniquely determined from data. One way for reducing the number of parameters in a model is by reducing the number of predictors without excessive loss of information. This situation is exemplified by numerous algorithms (Stumpf et al. 1998; Mitchell et al. 1998) that offer the user an

[1] Throughout this article, unless otherwise stated, a "model" shall refer to a statistical model.

unwieldy number of variables for predicting weather phenomena such as tornadoes or severe wind. A reduction in the number of variables may aid in better utilizing the algorithms for predictive purposes by avoiding the technical problem of overfitting, and by precluding any information overload.

Such a reduction can occur in at least two ways: One method is to retain only linear (or nonlinear) combinations of the predictors that account for most of the variance in the data. A well-known example is principal component analysis. Such methods make no reference to the dependent variable, and so are not appropriate for selecting the best predictors. A second approach is to take linear (or nonlinear) combinations that actually constitute a set of best predictors of the event at hand. This is equivalent to building a model (regression, or neural network, etc.) for predicting the events. Although the existence of a model simplifies the task of identifying the best predictors, a model is not always readily available. If a model does exist, then it is possible to rank the predictors according to some measure of their predictive strength and retain only the best predictors. Some methods that accomplish this task are stepwise regression and stepwise discriminant analysis. However, stepwise methods usually do not yield an unambiguous ranking of the variables. The reason is as follows: Stepwise methods are based on the improvement of performance upon the inclusion of some variable in the model (i.e., "forward stepwise"), or the loss of performance brought about by the exclusion of some variable from the model (i.e., "backward stepwise"). It is possible that the forward and backward procedures will lead to the same ordering of the variables, but that outcome is neither guaranteed nor likely. Furthermore, the selection criterion and the predictive power are quantities that must be specified. As a result, the list of the best predictors arrived at in this way is not necessarily unique.

There are other methods for ranking variables according to their predictive strength,

but most (if not all) invoke certain assumptions whose violation may be detrimental to the goal of finding the best predictors. The purpose of this article is threefold: first, to review some of the contingencies and difficulties in any attempt at ordering variables according to their predictive strength; second, to identify the conditions under which predictive strengths *can* be assigned; and, finally, to illustrate one example dealing with tornado prediction.

2 Contingencies and Difficulties

In this section, a number of situations are considered that expose some of the contingencies and difficulties associated with the question of best predictors.

Suppose that a parametric model has been found that faithfully relates n predictors, $x_i, (i = 1, n)$ to a single dependent variable, y. And suppose that the model is developed correctly. A well-known situation is when the model is linear in both the parameters and the variables:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + ... \ , \tag{1}$$

where $\alpha_i, (i = 1, n)$ are the parameters of the model. It is often said that the best predictor is the one with the largest (in magnitude) α, or equivalently, that the variables can be ordered according to the magnitude of the α's. However, that conclusion is contingent on at least two assumptions - that the variables all vary over the same range, and that they are uncorrelated. The first assumption may be fulfilled by simply scaling all the predictors so as to vary over the same range. The second contingency, however, is difficult to deal with. It is easy to show that if there exists any collinearity between two (or more) variables, then the corresponding coefficients are ambiguous in that a linear combination of them will produce the same value of y; this can lead to the best estimates for the α's that are excessively large positive (or negative) numbers (Draper and Smith 1981). As such, the α's become meaningless. Furthermore, it can

be shown that the standard error for the α's increase with the amount of collinearity, and as a result, their estimates become less precise (Tacq 1997).

A more general model may be linear in the parameters but with possibly nonlinear terms in the variables:

$$y = \alpha_1 x_1 + \alpha_2 x_1^2 + \alpha_3 x_1^3 + ... + \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_2^3 + ... \ , \tag{2}$$

where α_i, β_i are all parameters. If this is the model that best fits reality, then each variable is no longer associated with a single coefficient, and so it is impossible to assign a single measure of strength to the variables.

The model in (2) is an example of an additive model in that there are no interactions between the variables. If interactions are present, e.g.

$$y = \alpha_1 x_1 + \alpha_2 x_2 + ... + \beta x_1 x_2 + ..., \tag{3}$$

then it is clear that the notion of best predictors is completely meaningless, because the strength of any variable depends on the value of other variables. In other words, a given variable may be a strong predictor when some other variable takes on low values, but not otherwise. As such, it is simply impossible to assign a unique measure of predictive strength to any of the predictors.

It is entirely possible and even likely that the underlying model of a real-world problem is nonlinear in the parameters and includes interactions. An example of such a model is described in Marzban and Stumpf (1996, 1998), and Marzban et al. (1997) wherein a neural network for tornado prediction is outlined. For such nonlinear and interacting problems, the question of best predictors is then entirely unaddressable, at least uniquely, based on the parameters of the model.

It is possible to approach that question from a point of view that does not involve a direct examination of the parameters. One set of such approaches, namely the stepwise set, was mentioned in the Introduction. As mentioned there, although it

is possible to order the predictors according to the gain in performance upon their inclusion in the model, the reverse exercise (i.e., ordering the variables according to the loss in performance upon their systematic exclusion) can yield a different order. Consequently, any stepwise ordering of the variables according to their predictive strength is ambiguous and cannot lead to a unique set of best predictors.

All of the above mentioned methods presume the existence of a statistical model, be it linear regression or nonlinear regression methods such as neural networks. There exist situations, however, where even a statistical model does not exist. Examples include numerous meteorological algorithms that simply produce attributes of radar signatures that are believed to be associated with the phenomenon at hand (e.g., tornado, damaging wind) though without a model to relate the attributes directly to the phenomenon. The utility of such algorithms is not only in providing guidance, but also in providing the user with an arena wherein experimentation along with trial and error can induce a "mental model" that may in turn be employed for predictive purposes. Even when a statistical model exists, one cannot satisfactorily answer the question of best predictors. The issue becomes almost impossible to resolve when a model does not even exist.

It is interesting that the absence of a statistical model suggests an approach wherein the question of best predictors may actually be answered without being affected by the above mentioned difficulties. Regardless of the presense or absence of a statistical model, a reliable method for ordering the variables is a bivariate one (i.e. one predictor at a time). Such a bivariate analysis is model independent in that it does not presume the existence of a multiple regression model, a neural network, etc. As a result, it is unaffected by multicollinearity, interactions, and the other problems that infect multiple (independent) variable models. In this sense, a bivariate approach to ordering is the only meaningful approach, and it offers the additional flexibility of

benefiting the users of an algorithm who have only a "mental model" to guide them in utilizing the various variables for predictive purposes.

3 Bivariate Approaches

Identifying the best predictors in a bivariate analysis is still not free of ambiguities, but this time they are due to ambiguities in the definition of "best." In this section, several bivariate approaches will be proposed to address the question of the best predictors of a dependent variable. In particular, the predictors are assumed to be continuous (e.g., temperature, height), and the dependent variable is assumed to be binary (e.g., tornado or no tornado, rain or no rain, generally referred to as event or nonevent, and labeled as 1 or 0, respectively).

Perhaps the simplest approach for ordering the variables is according to their linear correlation with the dependent variable, specifically, using Pearson's linear correlation coefficient, r. When both the predictor and the dependent variable are continuous, r is a measure of linear correlation between the two. In the current case the dependent variable is binary, but r does still offer a measure of correlation, although a better description may be association (see Appendix). The correlation coefficient between two variables x and y is computed as

$$r_{xy} = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2 \sigma_y^2}}, \tag{4}$$

where the covariance and variance terms are computed as

$$\sigma_{xy}^2 = \overline{xy} - (\overline{x})(\overline{y}), \quad \sigma_x^2 = \overline{x^2} - (\overline{x})^2, \quad \sigma_y^2 = \overline{y^2} - (\overline{y})^2, \tag{5}$$

and the overline signifies the sample average of the corresponding quantity. r varies between -1 and +1. A nonlinear generalization of r will be discussed later in this section.

An alternative approach is suggested by considering the way in which a forecaster employs the variables at his/her disposal. He/she may be interested in issuing forecasts that maximize some measure of performance. Assuming that either higher or lower values of the predictor are associated with events, an important quantity is the value of the variable above (or below) which a "warning" is to be issued. Furthermore, it is important to identify the threshold at which the measure of performance is maximized. This suggests the following method for assigning a predictive strength to the various predictors: For each variable, dichotomize it by introducing a decision threshold, form a 2×2 contingency table for the forecasts and observations, compute some measure of performance based on the contingency table, and then order the variables according to the maximum obtainable value of that measure.

In this approach, it is important to compute a "good" measure of performance, otherwise the inferred predictive strengths may be incorrect or even unreasonable. Many performance measures have been examined in Marzban (1998a) for pathological behavior. It was found that all the examined measures are biased (or "inequitable") in that they induce under- or over-forecasting in a rare-event situation, though to different degrees. [2] A relatively unbiased measure is the Heidke Skill Statistic (HSS), and two relatively biased measures are the Critical Success Index (CSI) and the Likelihood Ratio Chi-square (LRC). They are defined as

$$\text{CSI} = \frac{C_4}{C_2 + C_3 + C_4}, \tag{6}$$

$$\text{HSS} = \frac{Tr(C - E)}{Tr(C^* - E)} = \frac{2(C_1 C_4 - C_2 C_3)}{N_0(C_2 + C_4) + N_1(C_1 + C_3)}, \tag{7}$$

$$\text{LRC} = -\sum_{i=1}^{4} C_i \log \frac{E_i}{C_i}, \tag{8}$$

where Tr is the trace operator (i.e., sum of the diagonal elements), and C and E are the contingency table and its (biased) expected value based on pure chance (i.e.,

[2] A rare-event situation refers to when an event is far more likely than the nonevent.

guessing), respectively. C^* is the contingency table for a set of perfect forecasts. The subscripts refer to the elements of the respective table. In particular, C_1 and C_4 are the number of correctly classified nonevents and events, respectively; C_2 and C_3 are the number of incorrectly classified nonevents and events, respectively. The expected matrix, E, is computed from the marginal probabilities which, in turn, can be estimated from C itself (Marzban 1998a).

Three other facets of forecast quality that are useful to examine are the Probability of Detection (POD), the False Alarm Ratio (FAR), and Bias, defined as (Marzban 1998a)

$$\text{POD} = \frac{C_4}{N_1}, \quad \text{FAR} = \frac{C_2}{C_2 + C_4}, \quad \text{Bias} = \frac{C_2 + C_4}{N_1} \; . \tag{9}$$

POD and FAR range from 0 to 1, and their respective optimal value is 1 and 0. Bias ranges from 0 to ∞, and its optimal value is 1; Bias$>$ 0 (Bias$<$ 0) implies over (under) forecasting.

In spite of its inequitability (Gandin and Murphy 1992; Marzban 1998a; Marzban and Lakshmanan 1999), CSI is a popular measure in meteorology because it can be computed without any knowledge of C_1; in practice, forecasters do not keep track of the number of nonevents when warnings are not issued. Whereas CSI is a measure of accuracy, HSS is a measure of skill, and so it takes into account random forecasts; said differently, if $C = E$, then $HSS = 0$. LRC, too, is a measure of skill, but it has an additional advantage (apart from the fact that it follows a chi-squared distribution (Fienberg 1977)): The appearance of the log function has the effect of magnifying the difference between correct and incorrect forecasts. As such, LRC can better differentiate between the strength of the predictors. As mentioned previously, because each measure captures a different aspect of performance quality, the choice of the best predictors may depend on the choice of the measure.

Finally, one may adopt a probabilistic approach by examining the (posterior)

probability of an event (e.g., tornado), $P_1(x)$, given the value of a predictor, x. This probability can be calculated from the conditional frequency distribution, $N_i(x)$, at a given value of x, where $i = 0, 1$ refers to nonevents and events, respectively. Specifically, it can be shown (Marzban 1998b) that Bayes' theorem implies

$$P_1(x) = \frac{N_1(x)}{N_1(x) + N_0(x)}. \tag{10}$$

A "good" predictor is one whose $P_1(x)$ changes significantly as a function of x. Although one may order the predictors according to the change in $P_1(x)$ over the range of x, it is more instructive to examine the plot of $P_1(x)$ as a function of x *for each variable*, because such plots display a multifaceted view of the importance of a predictor.

The first two methods have a limitation that the last method does not; they are linear. This causes a nonlinear variable (e.g., x2 in Fig. 5) to be assigned an incorrect (and possibly low) predictive strength. The probabilistic method exposes the nonlinearity of such variables and will, therefore, assign a faithful predictive strength. The disadvantage of the probabilistic method is that it does not offer a means of quantitatively ordering the variables according to a single (scalar) measure. In other words, the multifaceted nature of the plot of $P_1(x)$ as a function of x allows only for a coarse classification of the variables into a few classes of predictive strength (e.g., poor, marginal, good) and not a continuous ordering of the variables.

However, it is possible to distill the multi-faceted plot of $P_1(x)$ as a function of x into a single, one-dimensional (scalar) quantity. In fact, this quantity is a nonlinear generalization of the linear correlation coefficient, and is called the correlation ratio, η (Croxton and Crowden 1955). [3] Its exact definition (and its relation to r) is given

[3]The authors are indebted to one of the reviewers for pointing out the existence of this measure.

in the Appendix. For a binary dependent variable, its square can be written as

$$\eta^2 = \frac{N_0 + N_1}{N_0 N_1} \sum_x (N_0(x) + N_1(x))(P_1(x) - p_1)^2 \,, \tag{11}$$

where N_0 and N_1 are the sample sizes for nonevents and events, respectively, and $p_1 = N_1/(N_0+N_1)$ is the a priori (or climatological) probability of the event. Finally, it must be pointed out that some information is lost any time a multi-faceted quantity is reduced to a scalar. Therefore, although it is possible to order the variables according to their η, the plot of $P_1(x)$ as a function of x carries more information regarding the predictive strength of the variables (see the next section).

4 Application to Tornado Prediction

The National Severe Storms Laboratory's Tornado Detection Algorithm (TDA) has recently been added to the WSR-88D system. A descriptive outline of the TDA functionality, performance capability, strengths, and weaknesses can be found in Mitchell et al. (1998). The function of the TDA is to identify regions of strong azimuthal shear in Doppler velocity data that are often, but not always, associated with tornadoes. A strong azimuthal shear implies that a circulation is associated with a vortex. The TDA has replaced the original WSR-88D tornadic vortex signature algorithm, and as such, it is important to offer the prospective users of TDA some guidance so as to allow for a better utilization of the algorithm for predictive purposes. In particular, it is useful to know which of the many attributes of a vortex detected by TDA are most strongly associated with the occurrence of tornadoes. The answer will be considered within the context of the previously mentioned, bivariate methods.

The examined data set consists of 43 cases (or 275 hours) of WSR-88D data containing 207 tornado reports and over 173 severe wind reports from a variety of storm types from across the U.S. This constitutes $N_0 = 7224$ nontornadic circulations

detected by the TDA, and $N_1 = 730$ TDA-detected tornadic circulations [4]. Note that $N_0 >> N_1$.

The predictors computed by TDA are listed in Table 1 (in no particular order); throughout this article, however, they will be referred to by the numerical labels appearing in that table. Most of the variables have a self-explanatory meaning. However, it is worth elaborating on gate-to-gate velocity difference and shear. The former is the difference between two adjacent velocity gates which are adjacent in azimuth and constant in range. In contrast, shear is the velocity difference divided by the distance between the adjacent velocity gates.

5 Results of the Application

It is instructive to identify the variables that are correlated with one another, not only for gaining some substantive understanding of the data, but also as a check of the various methods; for example, if the predictive strengths of two highly collinear predictors are found to be significantly different, then one may suspect an error in the (bivariate) analysis. Pearson's correlation coefficient, r, can again be utilized to identify the mutually correlated predictors. However, the rare-event nature of the data set under study can cause r to be excessively large. For this reason, the correlation coefficients must be computed for the two classes, separately: $r^{(0)}$ for nontornadoes and $r^{(1)}$ for tornadoes. Pairs of variables that are highly correlated for both classes may be considered statistically equivalent (or redundant). The pairs with $r^{(0)} > 0.8$ and $r^{(1)} > 0.8$ are variables 5 and 8, and 6 and 7. The statistical equivalence of these

[4] Tornadic circulations are those which can be associated in space and time with a reported tornado (ground truth). A "time window" is applied such that associated circulations present within 20 minutes before the starting time of the tornado, and 6 minutes after the ending time, are also deemed tornadic.

variables is evident in their scatter plots (Fig. 1). The correlation coefficient between variables 5 and 8 is $r^{(0)} = 0.97$ for the nontornadic circulations and $r^{(1)} = 0.96$ for the tornadic circulations; the correlation coefficients between variables 6 and 7 are $r^{(0)} = 0.88$ and $r^{(1)} = 0.86$. The probability that values as large as these values of r could be obtained by chance was computed to be zero (to 12 decimal places). [5] The standard errors $((1 - r^2)/\sqrt{N})$ for these r's are, 0.001, 0.003, 0.003, and 0.01, respectively. Therefore, to a high level of significance, the corresponding pairs of variables are highly correlated.

The linear correlations between the predictors and the dependent variable (ground truth) are given in Fig. 2. The height of each bar is a measure of the predictive strength of the corresponding variable; a positive (negative) value for r implies that tornadoes are associated with larger (smaller) values of the corresponding variable. The standard error for these values of r is approximately 0.01. It is evident that according to this measure of predictive strength, variables x3, x4, x1, and x9 are the best predictors, respectively, in descending order. Also note that, as expected, the collinear variables have equal predictive strengths (within the standard error).

As described in Section 3, a predictor may be dichotomized by the introduction of a decision threshold, after which some categorical measure of performance may be computed. For example, Fig. 3a shows the dependence of the three measures on the value of the decision threshold placed on variable x2 (i.e., depth). It can be seen that a CSI of 0.15 can be reached if depths larger than approximately $6100m$

[5] The probability that a random sample would produce a value of r as large as the observed value of r is given by (Bevington 1992)

$$\frac{2}{\sqrt{\pi}} \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]} \int_{|r|}^{1} dx (1-x^2)^{(\nu/2-1)},$$

where $\nu = N - 2$ is the number of degrees of freedom for an experimental sample of size N.

are forecast as tornadic. Approximately the same threshold maximizes LRC, while to obtain a maximum HSS, the threshold must be placed at x2=7000m. Fig. 3b shows the corresponding values of POD, FAR, and Bias. For example, it can be seen that a set of forecasts that maximize HSS lead to a POD of 45%, FAR of 82%, and are nearly unbiased (Bias$\sim$ 1). By contrast, forecasts that maximize CSI or LRC are heavily biased (Bias$>>$ 1). Table 2 shows the analogous quantities for all the predictors. Note that the use of CSI or LRC leads to generally higher Bias values than that of HSS.

The maximum score reached by placing a threshold on each of the predictors is displayed in Fig. 4 for the three different measures. The height of a bar is a measure of the predictive strength of the corresponding variable. Recall that since x1, x5, and x8 (and to a statistically insignificant level, x10) are negatively correlated with tornados (Fig. 2) sub-threshold values should be forecast as tornadic.

Evidently, if maximizing CSI is the goal (Fig. 4a), then variables x3, x4, x1 and x9 are the best predictors in descending order. The high (linear) correlation between the variables x5 and x8, and x6 and x7, is manifest in Fig. 4a by their equal predictive strength. Note that employing x10 (i.e., range) appears to yield a nonzero CSI, in spite of the lack of any theoretical or physical reason for range to be a good predictor. This can be traced to the fact that CSI is not a measure of skill in that it does not take into account random forecasting. As advocated previously, the use of CSI may lead to false conclusions regarding the predictive strength of the various predictors.

A manifestation of the aforementioned inequitability of CSI is evident in Fig. 3; note that if one places a decision threshold at zero and proceeds to declare all detected circulations as tornadic, then a nonzero CSI is obtained. This may induce a forecaster to overforecast. In fact, in a rare-event situation (i.e., $N_0 >> N_1$), CSI can reach its maximum at the lowest value of the predictor (Marzban 1998a), causing

severe overforecasting on part of a forecaster who employs CSI to gauge performance. Indeed, CSI would not have been included in this analysis were it not for its popularity (due to its independence of the C_1 element of the contingency table).

Coincidentally, the best predictors according to CSI are the same set of predictors that maximize HSS (Fig. 4b). The noticeable and welcomed difference is that x10 is assigned a much lower predictive strength according to HSS.

If LRC is to be maximized, then the best predictors are x1, x3, x9, and x4, in descending order (Fig. 4c). The ability of LRC to better differentiate between the predictors is apparent in the erratic nature of the vertical bars in Fig. 4c. Also note that x10 emerges with an almost nonexistent predictive strength, and correctly so.

As for the probabilistic approach, the curves for $P_1(x)$ are presented in Fig. 5 for all the predictors. The curve with the error bars is $P_1(x)$ as a function of x, and the curves marked with 0 and 1 are the normalized probability densities ($N_0(x)/N_0$ and $N_1(x)/N_1$). The error bars on the $P_1(x)$ curve reflect the sampling error. The range of the probabilities obtained in these plots is more meaningful if one realizes that the a priori probability of a TDA-detected circulation being tornadic, as estimated by $N_1/(N_1 + N_0)$, is about 0.09. It can be seen that variable x3 is an example of a "good" predictor, while a "poor" predictor is variable x10.

These probability plots are multi-dimensional entities and, as such, do not directly allow for a quantitative ordering of the variables. Therefore, they are coarsely divided into three classes of predictive strength - poor, marginal, and good - corresponding to variables whose posterior probabilities generally vary in the 10%, 20%, and 50% ranges, respectively. The results are tabulated in Table 3.

A finer classification is possible if one allows for some loss of information. As shown in (11), the correlation ratio can be computed from $P_1(x)$. As such, η allows for further distillation of the information contained in $P_1(x)$. The predictive strength

of the variables according to η are given in Fig. 6. This figure is very similar to Fig. 2; in fact, η is almost equal to r for all of the variables. The only exceptions are x1, x2, and x9 which have $\eta > r$; this is consistent with Fig. 5, where it can be seen that only these variables are nonlinear.

As mentioned previously, any distillation of the probability plots leads to loss of information. For example, as seen from Fig. 5, variable x2 has little or no predictive strength for x2$< 5000m$; only for $P_1(x) > 5000m$ does it begin to have any predictive strength. Even a measure like η which is a measure of nonlinear correlation leads to a single number that does not capture such nonlinearity. Said differently, a scalar measure has no diagnostic capability, though it can still determine the predictive strength of the variables.

6 Summary

First, it is argued that the task of assigning predictive strengths to a number of predictors is difficult, at best. Some of the assumptions/contingencies underlying that task are discussed, and it is shown that they are avoided in a bivariate analysis, i.e. one independent variable at a time. Several such methods are offered after which they are illustrated in an application to tornado prediction. It is found that the various tornado predictors in the National Severe Storms Laboratory's Tornado Detection Algorithm portray a wide range of predictive strengths depending on the measure and the method of obtaining the predictive strength. Among the various methods and measures, a consensus does exist, however, regarding the choice of the best predictors.

The analysis suggests that variables x3, x4, x1 and x9, in descending order, have the highest linear correlation and correlation ratio with tornadoes. They also produce the highest performance as gauged by CSI and HSS. Maximizing LRC, on the other

hand, leads to a different order for the same variables, namely x1, x3, x9, and x4. As for the probabilistic method, the outstanding predictors for tornadoes are x2, x3, x4, and x9 (in no particular order). Variables x3 (low-level gate-to-gate velocity difference), x4 (maximum gate-to-gate velocity difference), and x9 (Tornado Strength Index) can be considered to meet the consensus of the best predictors.

7 Acknowledgments

V. Lakshmanan and A. Witt are acknowledged for valuable discussion and a thorough reading of an early version of this manuscript. Support was provided by the Operational Support Facility/NOAA and the Federal Aviation Association.

8 Appendix

In this appendix, the formulae for the linear correlation coefficient, r, and the correlation ratio, η, are given and specialized to the case wherein the dependent variable is binary (0 or 1).

The square of the linear correlation coefficient as written in (4) is in fact equal to the proportion of the total variance that is explained by a least-squares regression line $y(x_i) = ax_i + b$. That quantity is called the coefficient of determination, and can be written as

$$r^2 = \frac{\sum_i (y(x_i) - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}, \tag{12}$$

where y_i is the i^{th} observation of the dependent variable, and $\bar{y}$ is the average of the N observations. When the y_i take 0 or 1 values, then $\bar{y} = N_1/(N_0 + N_1)$, where N_0 and N_1 are the sample sizes for the two classes. Note that this ratio is nothing but the a priori or climatological probability of tornado, p_1. Similarly, the variance of y can be written as

$$\sigma_y^2 = \frac{N_0 N_1}{(N_0 + N_1)^2}, \tag{13}$$

after which the formula for r^2 can be written as

$$r^2 = (N_0 + N_1)^2 \frac{(\bar{x}_1 - \bar{x}_0)^2}{\sigma_x^2}. \tag{14}$$

Therefore, it can be seen that r is proportional to the distance between the means of the independent variables in the two classes. As such, it is better described as a measure of discrimination or association.

A nonlinear generalization of the linear correlation coefficient is the correlation ratio (Croxton and Crowden 1955, Panofsky and Brier 1968), η. As in r, it is defined as the proportion of the total variance that is explained by the fit, but in contrast with r, the fit is not assumed to be linear. However, since the form of the nonlinear curve is not specified, η is instead defined in terms of the average of the dependent variable for specific values of the independent variable. Specifically, its square can be written as

$$\eta^2 = \frac{\sum_x N_x(\overline{y}_x - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} \quad , \tag{15}$$

where $\bar{y}_x$ is the average of dependent variable corresponding to some specified value of the independent variable x, and the $\sum$ is over the full range of x. N_x is the sample size for that value of x, and it should be sufficiently large as to assure the smooth variation of y_x with x. If $y_i = 0, 1$, then

$$\bar{y}_x = \frac{N_1(x)}{N_1(x) + N_0(x)} = P_1(x) \quad , \tag{16}$$

where $N_1(x)$ and $N_0(x)$ are the sample sizes for the two classes but for a specific value of x. Combining the above equations results in

$$\eta^2 = \frac{N}{N_0 N_1} \sum_x (N_0(x) + N_1(x))(P_1(x) - p_1)^2 \quad . \tag{17}$$

Evidently, η^2 is a measure of the amount by which the posterior probability of tornado differs from the a priori probability of tornado, averaged over the full range of x.

9 References

Bevington, P. R., and D. K. Robinson: *Data Reduction and Error Analysis for the Physical Sciences.* McGraw-Hill, 328 pp.

Croxton, F. E., and D. J. Crowden, 1955: *Applied General Statistics.* Prentice Hall, 843 pp.

Draper, N.R., and H. Smith, 1981: *Applied Regression Analysis.* John Wiley & Sons, 709 pp.

Fienberg, S. E., 1977: *The Analysis of Cross-classified Categorical Data.* MIT Press, 190 pp.

Gandin, L. S., and A. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.

Marzban, C., 1998a: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753-763.

———, 1998b: Bayesian probability and scalar performance measures in Gaussian Models. *J. Appl. Meteor.*, **37**, 72-82.

———, and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617-626.

———, and ———, 1998: A neural network for damaging wind prediction. *Wea. Forecasting*, **13**, 151-163.

———, and V. Lakshmanan, 1999: On the uniqueness of Gandin and Murphy's equitable performance measures. *Mon. Wea. Rev.*, **127**, 1134-1136.

———, H. Paik, and G. Stumpf, 1997: Neural networks vs. Gaussian discriminant analysis. *AI Applications*, **11**, 49-58.

Mitchell, E. D., S. V. Vasiloff, A. Witt, M. D. Eilts, G. J. Stumpf, J. T. Johnson, and K. W. Thomas, 1998: The National Severe Storms Laboratory Tornado Detection Algorithm. *Wea. Forecasting*, **13**, 352-366.

Panofsky, H. J., and G. E. Brier, 1968: *Some Applications of Statistics to Meteorology.* The Pennsylvania State University, University Park, 224 pp.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 304-326.

Tacq, J., 1997: *Multivariate Analysis Techniques in Social Science Research.* Sage Publications, 411 pp.

10 Figure Captions

Figure 1. Scatterplots between variables 5 and 8, and 6 and 7. The circles (squares) represent the nontornadic (tornadic) circulations.

Figure 2. The (linear) correlation coefficient, r, between the dependent variable (ground truth) and each of the predictors (see Table 1). Standard error=0.01.

Figure 3. a) Performance measures, CSI (solid curve), HSS (dashed curve), and LRC (dashed-dotted curve), and b) POD, FAR and Bias, as a function of the value of the decision threshold placed on the predictor x2 (i.e. depth). The horizontal (dotted) line has been drawn to point out the threshold at which Bias=1.

Figure 4. The maximum value of three performance measures obtained by dichotomizing the predictors.

Figure 5. The posterior probability of tornado, given the value of the variable (the curve with error bars), and the probability densities for nontornadoes (labeled with 0) and tornadoes (labeled with 1).

Figure 6. The correlation ratio, η, between the dependent variable (ground truth) and each of the predictors. Standard error=0.01.

Variable label	Variable description
x1	Base (m, Above Radar Level (ARL))
x2	Depth (m)
x3	Low-level gate-to-gate velocity difference (ms^{-1})
x4	Maximum gate-to-gate velocity difference (ms^{-1})
x5	Height of maximum gate-to-gate velocity difference (m, ARL)
x6	Low-level shear ($10^{-3}s^{-1}$)
x7	Maximum shear ($10^{-3}s^{-1}$)
x8	Height of maximum shear (m, ARL)
x9	Tornado Strength Index ($ms^{-1} \times 100$)
x10	Range (Km from radar)

Table 1: The list of the variables and their corresponding labels. Consult Mitchell (1998) for a precise definition of the variables.

		Variable								
Measure		x1	x2	x3	x4	x5	x6	x7	x8	x9
CSI	Thresh	1400	6100	28	37	2400	22	26	2400	5000
	POD	51	45	46	42	43	45	50	44	47
	FAR	79	82	73	76	81	80	81	82	79
	Bias	2.5	2.5	1.7	1.8	2.3	2.2	2.7	2.4	2.3
HSS	Thresh	1300	7000	33	45	1500	24	28	1500	5200
	POD	46	28	33	28	24	38	45	26	46
	FAR	79	78	67	69	77	79	80	76	79
	Bias	2.2	1.3	1.0	0.9	1.0	1.8	2.3	1.1	2.1
LRC	Thresh	2400	6000	28	30	2500	16	26	2500	3500
	POD	86	47	46	62	45	64	50	46	77
	FAR	84	82	73	80	82	84	81	81	84
	Bias	5.5	2.6	1.7	3.2	2.5	4.1	2.7	2.5	4.7

Table 2: The decision thresholds yielding the maximum obtainable scores CSI, HSS and LRC, and the corresponding values of POD, FAR and Bias. No thresholds are given for variable x10, since it has no true predictive strength.

Predictive Strength	Variable
Good ($0 < P_1 < 0.5$)	x2,x3,x4,x9
Marginal ($0 < P_1 < 0.2$)	x1,x5,x6,x7,x8
Poor ($0 < P_1 < 0.1$)	x10

Table 3: The predictive strength - good, marginal, poor - of each variable according to the probabilistic approach.

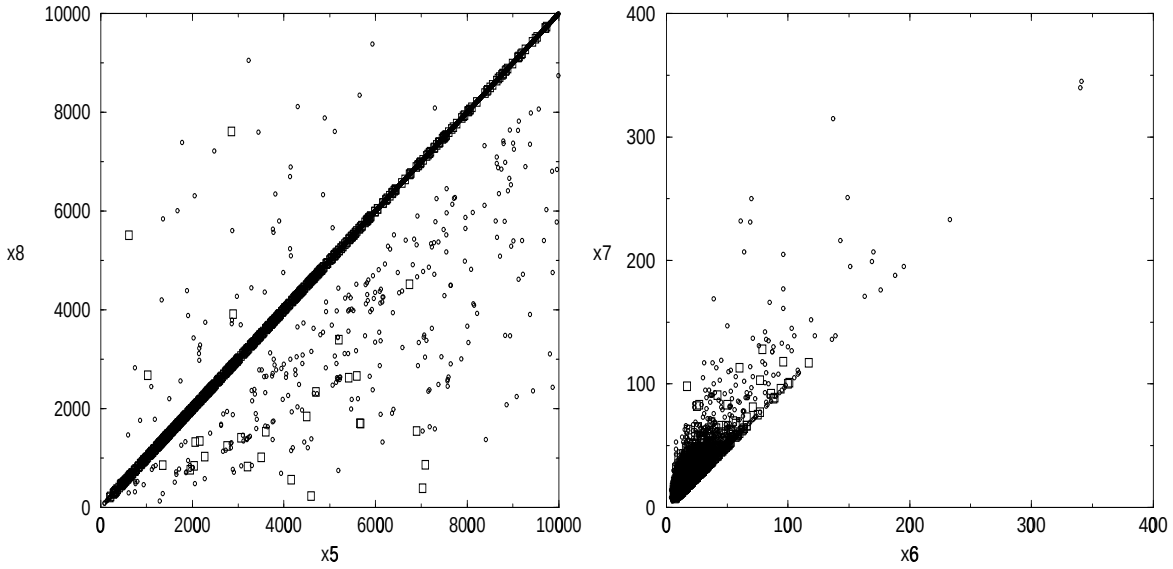

Figure 1: Scatterplots between variables 5 and 8, and 6 and 7. The circles (squares) represent the nontornadic (tornadic) circulations.

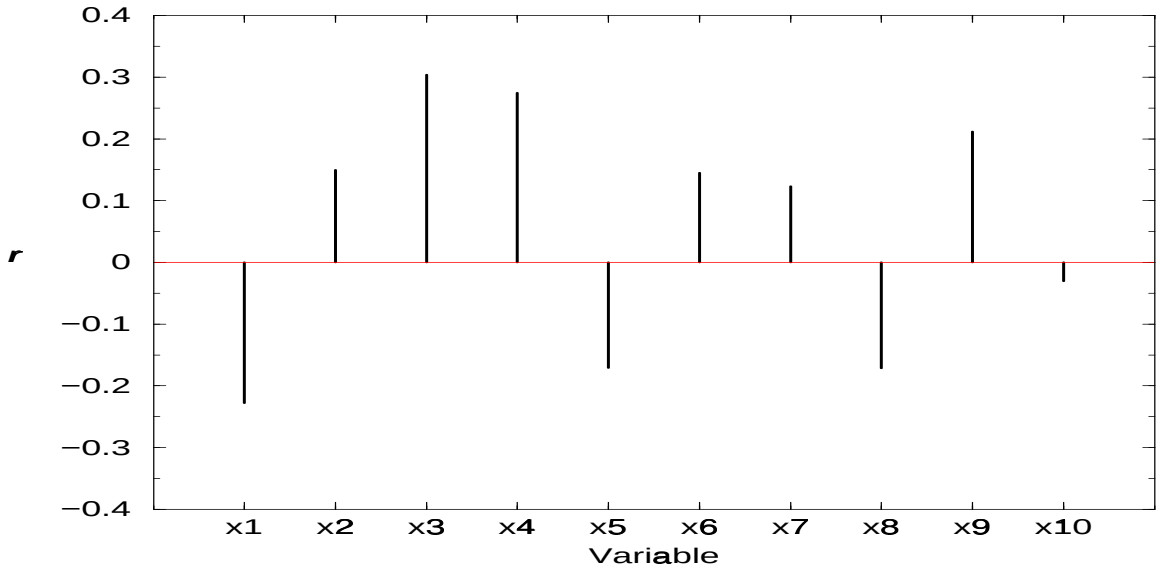

Figure 2: The (linear) correlation coefficient, r, between the dependent variable (ground truth) and each of the predictors (see Table 1). Standard error=0.01.

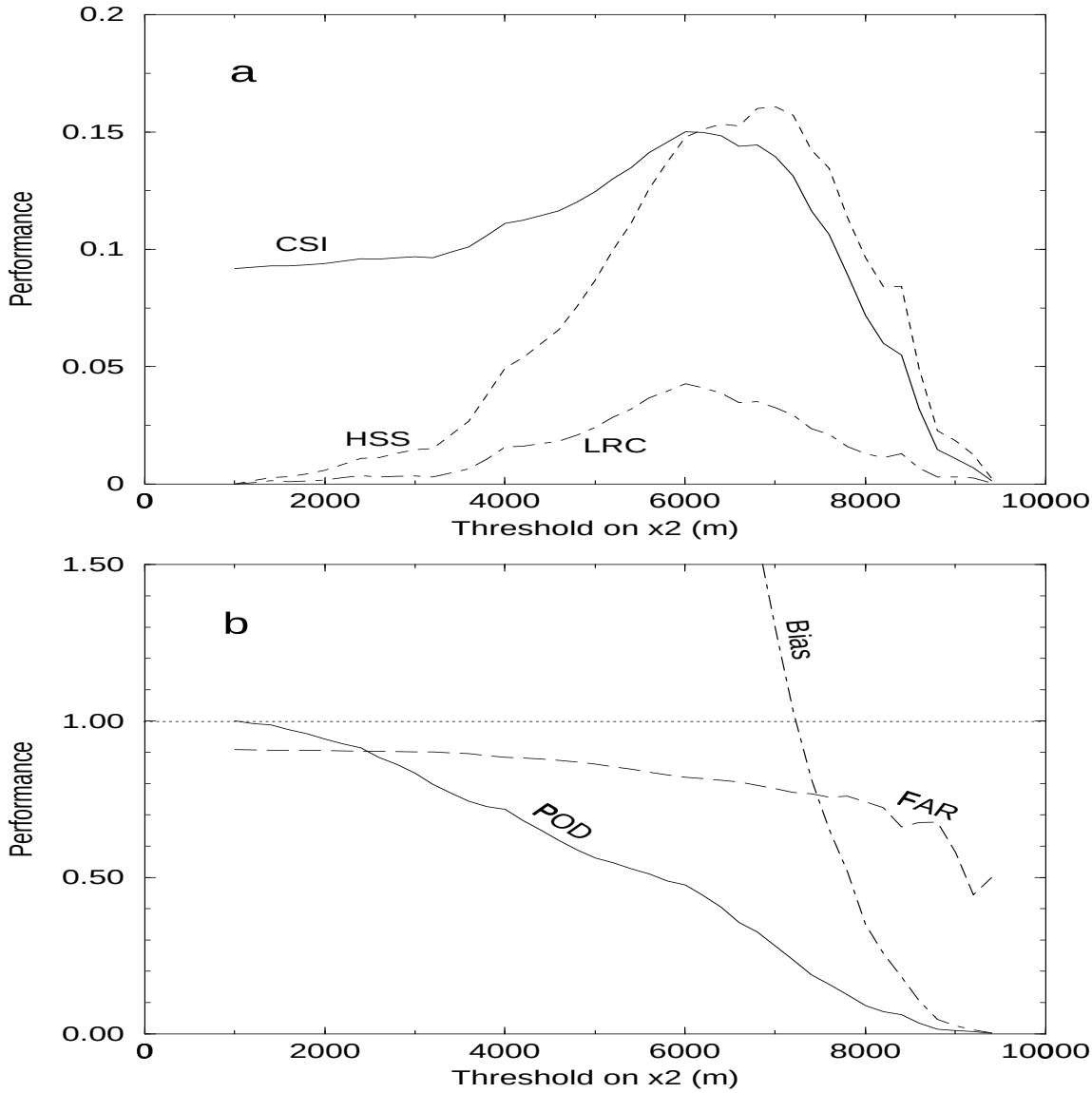

Figure 3: a) Performance measures, CSI (solid curve), HSS (dashed curve), and LRC (dashed-dotted curve), and b) POD, FAR and Bias, as a function of the value of the decision threshold placed on the predictor x2 (i.e. depth). The horizontal (dotted) line has been drawn to point out the threshold at which Bias=1.

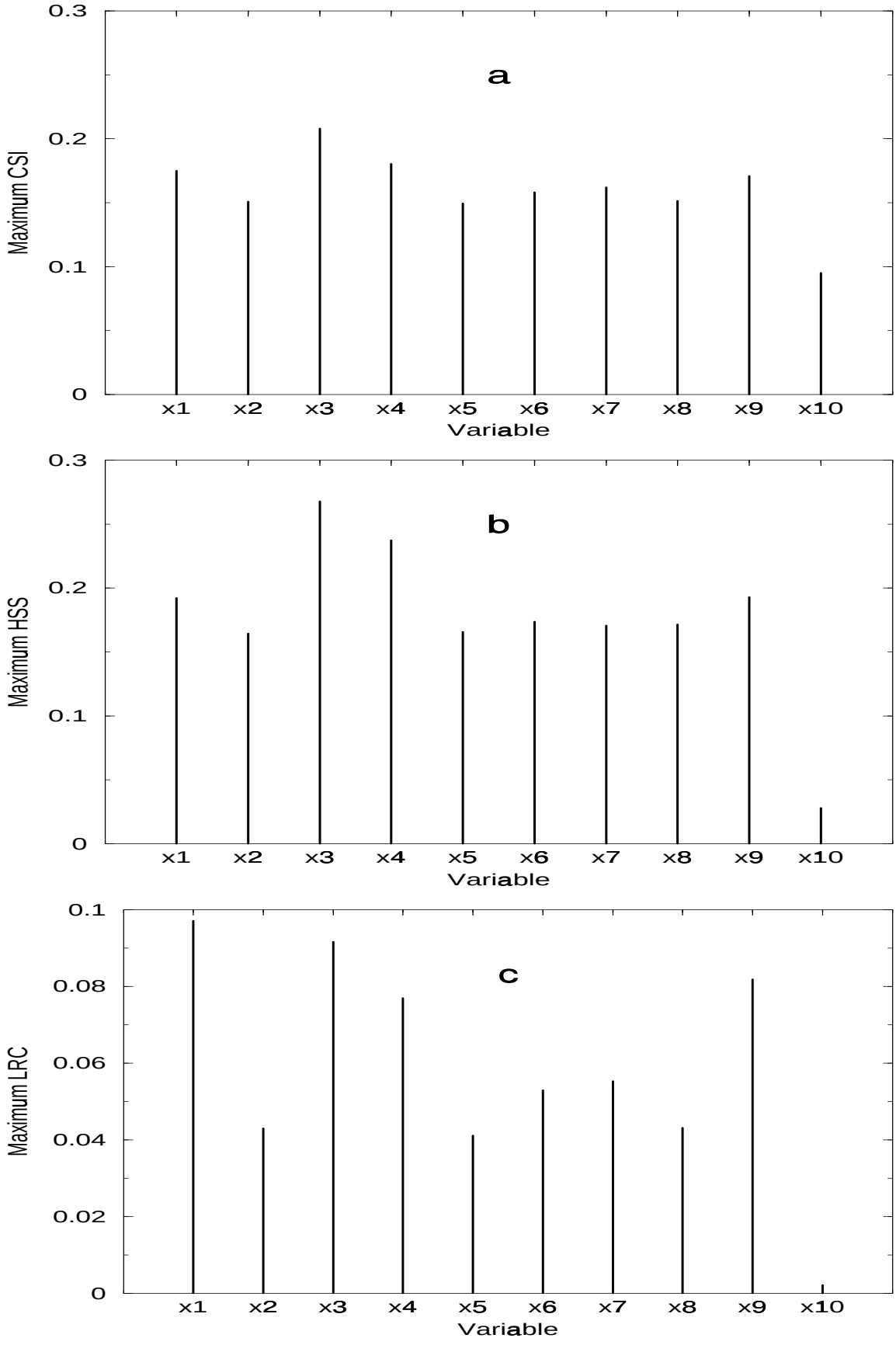

Figure 4: The maximum value of three performance measures obtained by dichotomizing the predictors.

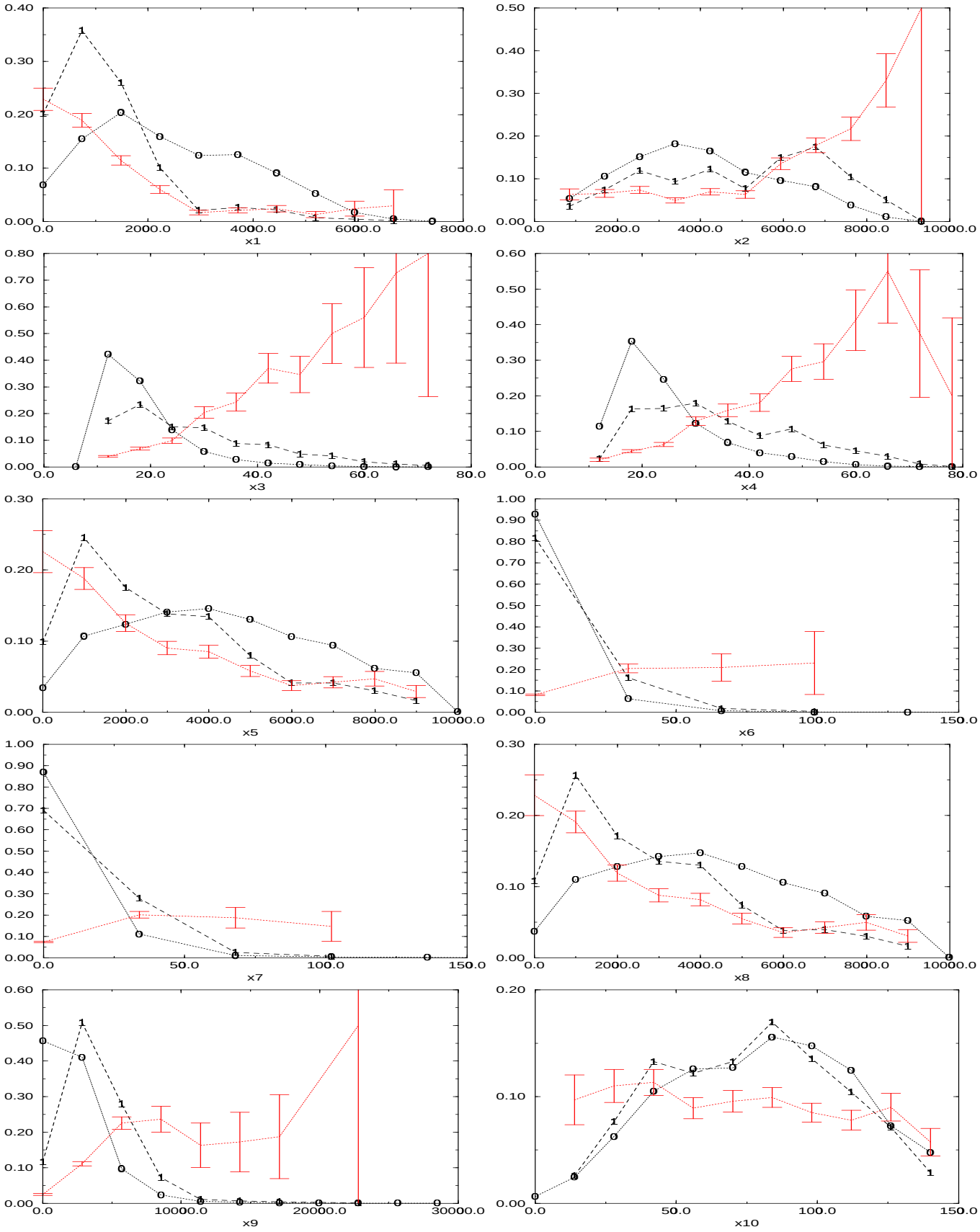

Figure 5: The posterior probability of tornado, given the value of the variable (the curve with error bars), and the probability densities for nontornadoes (labeled with 0) and tornadoes (labeled with 1).

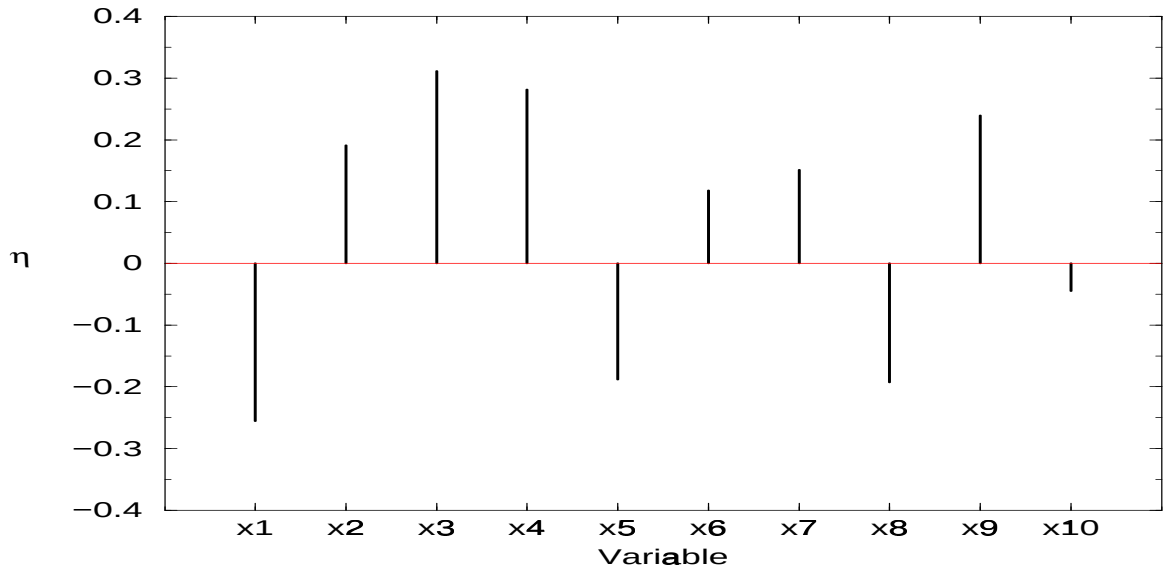

Figure 6: The correlation ratio, η, between the dependent variable (ground truth) and each of the predictors. Standard error=0.01.