Assessing Forecast Skill through Cross Validation

J. B. ELSNER

Department of Meteorology, The Florida State University, Tallahassee, Florida

C. P. SCHMERTMANN

Department of Economics, The Florida State University, Tallahassee, Florida

(Manuscript received 14 March 1994, in final form 30 June 1994)

ABSTRACT

This study explains the method of cross validation for assessing forecast skill of empirical prediction models. Cross validation provides a relatively accurate measure of an empirical procedure's ability to produce a useful prediction rule from a historical dataset. The method works by omitting observations and then measuring "hindcast" errors from attempts to predict these missing observations from the remaining data. The idea is to remove the information about the omitted observations that would be unavailable in real forecast situations and determine how well the chosen procedure selects prediction rules when such information is deleted. The authors examine the methodology of cross validation and its potential pitfalls in practical applications through a set of examples. The concepts behind cross validation are quite general and need to be considered whenever empirical forecast methods, regardless of their sophistication, are employed.

1. Introduction

The past decade has seen renewed interest in statistical weather and climate predictions. From forecasting thunderstorm probabilities using expert systems to outlooks of El Niño using canonical correlations, there is a resurgent interest in using historical data to predict the future. This movement away from complete reliance on dynamical prediction models can be explained in part by the availability of better data, the relative ease at which empirical models can be implemented, and the emergence of new techniques such as expert systems, neural networks, and genetic algorithms. Perhaps more importantly, however, in many situations forecast skill from relatively simple statistical models can compete with forecast skill from dynamical models of significantly greater complexity. For example, forecasts of El Niño from multivariate linear correlation analyses (Barnston and Ropelewski 1992) can compete with forecasts from coupled atmosphere–ocean dynamical models (Cane et al. 1986).

When developing statistical models, an accurate estimate of skill can be achieved through a procedure called "cross validation." The method of cross validation, including important issues related to its successful implementation for regression models, is outlined in Michaelsen (1987). The effectiveness and validity of cross validation for estimating forecast skill (Barnston and van den Dool 1993) and the potential for misapplication have motivated the present article. The authors emphasize some of the potential pitfalls when estimating forecast skill and, in doing so, hope to clarify the procedure of cross validation. Section 2 describes the purpose and nature of cross validation and some of the problems with implementation. Section 3 follows with several specific examples, and section 4 contains a summary and closing remarks.

2. The procedure of cross validation

In any dataset of observations there is both useful information (signal) about the underlying physics of the process being studied and extraneous information (noise) related to coincidences in the sampling process, measurement errors, and possible inclusion of irrelevant variables. The primary challenge to the researcher following a statistical approach to prediction is to devise an empirical method that strengthens signals in the available data and dampens noise. A desirable method for selecting a prediction rule from a historical dataset is one that accurately captures the fundamental relationships between variables, without relying too heavily on past coincidences that are unlikely to be repeated in the future.

Given the widespread availability of fast-computing machines, it is now relatively easy to build complex rules that "predict" the past accurately. The central question, however, is whether such rules reflect the underlying physics or are simply exploiting past coinci-

Corresponding author address: J. B. Elsner, Department of Meteorology—3034, The Florida State University, Tallahassee, FL 32306-3034.

dences. Rules that overemphasize coincidences are likely to perform poorly in actual forecast situations.

Cross validation attempts, with a limited data sample, to simulate actual forecast situations and thus provide an honest measure of an empirical procedure's ability to produce a skillful prediction rule. Predictive skill is the skill expected when the prediction rule, chosen by the procedure, is used in practice to forecast the future. Cross validation works by developing a separate prediction rule for each observation in the dataset based only on the remaining observations. The other observations represent a fictitious reordering of "history" from which to predict the omitted observation, and the resulting "predictions" are termed "hindcasts." A successful cross validation will remove the noise specific to each observation and assess how well the chosen procedure selects prediction rules when this coincidental information is deleted.

Some formal notation will facilitate discussion. Let x (independent variables) denote a vector of predictors and let y (dependent variable) denote an outcome. A prediction rule f is a deterministic mapping that produces a predicted outcome $\hat{y} = f(x)$ for any vector of predictors x. In this general formulation, a prediction rule f might be a linear regression model with specific coefficients (e.g., $\hat{y} = 3 - 4x_1 + x_2$) but it might also be something more exotic, such as a neural network with a specific set of weights or an expert system with a specific set of decision rules.

Let F denote the set of all prediction rules under consideration by the researcher. For example, F could be the set of all linear combinations of x or the set of all neural nets with a given architecture. The problem facing the researcher is to use the available data on x and y to choose a single prediction rule $f \in F$. Let A denote a deterministic algorithm that makes this choice; that is, an algorithm A is the predefined procedure that takes as input a dataset containing multiple observations on x and y values and produces as output a single "best" prediction rule f out of the set F. Standard ordinary least-squares (OLS) regression is such an algorithm; it selects $f(x) = x^T[(X^TX)^{-1}X^Ty]$ from a set F of all prediction rules having the form $f(x) = x^T[\beta]$, where the superscripts T and -1 denote the matrix operations of transpose and inverse, respectively, and where β is a vector of coefficients. However, the algorithm A could be considerably more complicated. For example, the algorithm might first consist of choosing the number of independent variables used in the prediction rule followed by an OLS estimation of the coefficients. The essential point of this article is that cross validation is performed on an algorithm (A) rather than on a particular prediction rule (f).

There are three important considerations in performing cross validation for the purpose of accurately estimating forecast skill of a chosen prediction rule. The first concerns the condition that hindcasts must be performed using out-of-sample data. In-sample hindcasts reflect only the degree to which the prediction rule chosen by A fits the data. Second, prediction rules used in hindcasts should not be chosen based on decisions that require information from the entire data sample. If this condition is violated, then the algorithm has not been truly cross validated. Third, the subsample from which a hindcast is generated must be independent of the omitted observations. If neighboring events in time are nonnegligibly correlated, then successively removing a single observation is not appropriate since "future" information will be used in the algorithm's choice of a prediction rule.

The three considerations are all variations of the general requirement that the forecast target be independent of the development sample, where the development sample is defined as the portion of the data from which the rule is derived, and the forecast target is the portion of the data used for predictions. More specifically, the forecast target must not be allowed, in any way, to influence the development of the prediction rule f. The separate effects of these three considerations are demonstrated with the following examples.

3. Examples

We will illustrate the important considerations involved in estimating forecast skill with a cross-validation procedure on the problem of predicting 48-h hurricane intensity changes. The problem is particularly relevant since it is one for which statistical rules can sometimes outperform dynamical models. A dataset consisting of variables known to have an effect on hurricane intensity was obtained from M. DeMaria of the Hurricane Research Division of the National Hurricane Center in Coral Gables, Florida. Independent variables are listed in Table 1. Values of the dependent

TABLE 1. A description of the 11 independent variables (predictors) used in developing prediction rules for forecasting 48-h hurricane intensity changes.

Variable no.	Variable
1	day of year
2	longitude of storm
3	previous 12-h intensity change (kt)
4	maximum possible intensity determined from SST–current intensity
5	magnitude of 200–850-mb vertical shear of horizontal wind
6	storm-size parameter
7	200-mb eddy relative angular momentum flux convergence
8	200-mb planetary angular momentum flux convergence
9	distance to nearest land
10	time tendency of vertical shear
11	square of variable 4

variable (48-h intensity change) are extracted from 43 separate hurricanes and given in Table 2. Standardized values of the corresponding predictor variables are available via anonymous ftp on metlab1.met.fsu.edu in file/pub/elsner/hurint.dat3. It is assumed that data from each storm represent independent information.

a. In sample versus out of sample

The problem is to make an accurate forecast of 48-h hurricane intensity change from the values of these 11 predictors. Assuming that an OLS linear regression will work well in this situation, the question is how well. Let F be the set of all 11 variable linear prediction rules ($x^T\beta$) and let A be OLS regression. There are two ways to estimate forecast skill through hindcasts, in sample and out of sample. An in-sample hindcast results from using all 43 storms to estimate a single set of coefficients (a particular $f \in F$) and then using that particular rule to hindcast each storm's intensity change. The alternative is to use cross validation to obtain an out-of-sample estimate. Cross validation is performed on A by successively excluding single observations and repeating the OLS procedure on the remaining 42 observations 43 times. Each time the algorithm is run a different f will be chosen to hindcast the excluded observation. As expected, the cross-validated procedure yields a larger error estimate (Table

TABLE 2. A list of the storms, dates, times, and magnitudes of 48-h hurricane intensity changes used as the dependent variable (predictand). The intensity change is determined by the change in maximum sustained winds in knots over the 48-h period.

Storm	Year	Month	Day	Hour	48-h intensity Δ
Debby	82	9	15	0	30.0
Josephine	84	10	10	0	30.0
Gloria	85	9	25	0	−30.0
Arlene	87	8	10	12	25.0
Bret	87	8	18	0	15.0
Cindy	87	9	6	0	15.0
Dennis	87	9	10	0	15.0
Emily	87	9	20	0	35.0
Floyd	87	10	10	0	30.0
Gilbert	88	9	11	12	30.0
Joan	88	10	11	12	0.0
Barry	89	7	10	12	15.0
Dean	89	7	31	12	40.0
Erin	89	8	18	0	10.0
Felix	89	8	27	0	10.0
Gabrielle	89	8	31	0	40.0
Hugo	89	9	11	12	30.0
Iris	89	9	17	0	20.0
Jerry	89	10	12	12	30.0
Karen	89	11	28	12	25.0
Arthur	90	7	23	0	15.0
Bertha	90	7	28	0	25.0
Cesar	90	8	2	0	10.0
Edouard	90	8	7	0	10.0
Fran	90	8	12	0	5.0
Gustav	90	8	24	12	35.0
Hortense	90	8	25	12	15.0
Isidore	90	9	4	12	30.0
Josephine	90	9	21	12	0.0
Klaus	90	10	3	12	40.0
Lili	90	10	11	0	0.0
Nana	90	10	16	12	45.0
Ana	91	7	3	0	20.0
Bob	91	8	16	0	45.0
Claudette	91	9	4	12	50.0
Danny	91	9	7	0	15.0
Erika	91	9	9	0	25.0
Andrew	92	8	17	0	15.0
Bonnie	92	9	18	0	60.0
Charley	92	9	22	0	55.0
Danielle	92	9	22	12	10.0
Earl	92	9	27	0	5.0
Frances	92	10	23	12	5.0

3, case 1). The cross-validation exercise yields a mean absolute error (MAE) that is 37% larger than the MAE when the algorithm is not cross validated. More significantly, when individual hindcasts are compared, hindcasts made in sample are always better than their cross-validated counterparts. However, since cross validation mimics actual forecast situations, resulting error estimates are more accurate in terms of what can be expected when the procedure is used to choose a rule for forecasting the future.

b. *False versus true cross validation*

Cross validation operates on algorithms (A) rather than on prediction rules (f). This is a subtle but very important point. For example, suppose $I_k(x)$ is a matrix whose columns are the eigenvectors corresponding to the k most significant eigenvalues of X^TX. Suppose further that the choice of k is part of the algorithm A. Imagine an A that works as follows: 1) select a trial value of k, and thus a trial $I_k(x)$; 2) estimate out-of-sample errors for an OLS regression of y on $I_k(x)$ by cross validation; 3) select the value of k (k_{min}) that performs the best in step (2); and 4) select as the prediction rule f the OLS regression of y on the k_{min} surrogate variables.

Is the estimated error for k_{min} from step 2 an honest assessment of the forecast skill of the rule produced when A is applied to the full dataset? No. Step 2 cross validates for a specific k but does not cross validate the procedure for choosing k. The usual cross-validation method still needs to be applied, even though A itself contains an "inner" cross validation in step 2. To truly cross validate A, we must successively omit single observations, apply A (including the choice of k) to the remaining data to generate prediction rules, and use the prediction rules to hindcast the omitted data.

An example using the hurricane intensity-change data illustrates how artificial skill will be introduced if k is chosen with the help of the entire data sample, even if the error for OLS regression of y on the selected indices is estimated using cross validation. In this example, data reduction is based on an empirical orthogonal function (EOF) analysis of the 11 independent variables. The method requires a decision about the number of significant eigenvalues to retain. As stated above, if this decision is made by selecting the number of eigenvalues that minimizes the cross-validated hindcast error, then it is part of the algorithm for choosing a prediction rule and must also be cross validated.

Examining the errors for each successively added EOF yields a best decision at nine EOFs, with an inner cross-validated MAE of 12.95 (Table 3, case 2). This is not, however, an accurate estimate of forecast skill of the resulting prediction rule. A true cross validation of the algorithm is performed by deleting each storm in turn and selecting a prediction rule by applying the algorithm that includes selecting k to minimize the cross-validated MAE over the remaining $N - 1$ observations. A hindcast is then made with the selected rule on the original deleted storm. Performing true cross validation on this dataset produces an MAE of 15.28, which is 18% greater than the MAE of the false cross-validation exercise. In this dataset k_{min} takes on the values 2, 4, 5, and 9 depending on which data point is removed. True cross validation produces a more conservative and more accurate estimate of forecast skill since true cross validation does not use own-storm information at any step in the algorithm that produces the hindcasts.

The message of the above example is particularly relevant for more sophisticated algorithms associated with neural networks or genetic algorithms, for example. Suppose a decision as to how many hidden layers to include in the network architecture is made by considering out-of-sample performance of the network. Then, as the previous example demonstrated, this decision is part of the algorithm and must also be cross validated. Failure to perform such an "outer" cross validation will lead to unrealistically high estimates of forecast skill for the neural network.

In comparing the properly cross-validated hindcast skill of the more complex EOF-based algorithm used in this subsection with cross-validated hindcast skill of the simpler multivariate linear regression algorithm

TABLE 3. Comparison of hindcast skills for the three cases examined in this study. Case 1 is a comparison between no cross validation and cross validation. Case 2 is an example where cross validation is used improperly, and case 3 is an example where cross validation is performed without regard to serial correlation.

				Cumulative distribution of errors		
		MAE	Percent of hindcasts where $error_a < error_b$	25%ile	50%ile	75%ile
Case 1	a. In sample	9.76	100%	−9.28	−0.90	5.97
	b. Out of sample	13.39	—	−13.49	−0.59	10.40
Case 2	a. False X validation	12.95	95%	−10.91	−1.71	9.58
	b. True X validation	15.28	—	−12.66	−3.75	11.58
Case 3	a. Serial correlation	12.65	93%	−13.47	−7.13	5.90
	b. No serial correlation	13.47	—	−14.13	−7.79	6.77

used in the previous subsection, we note that the simpler algorithm performs better. Just as simple statistical models can equal or outperform complex dynamical models, so too can they sometimes outperform more complex statistical models. This is particularly true when sample sizes are modest, as they often are in climate forecasting problems. A proper cross validation helps you see this point very clearly.

c. *Serial correlation versus no serial correlation*

Another consideration when employing cross validation is the presence of serial correlation. In cases where serial correlations exist, the simple cross-validation procedure of successively omitting single observations from the data will introduce bias into the estimation of forecast skill. Bias enters in part because nearby observations (both past and future) contain noise related to noise in the omitted observation. It also enters because the predictor values of nearby future observations are likely to be close to those of the omitted observation. These nearby future points are thus likely to be especially informative about the omitted predictant, but they would be unavailable in a real forecast situation. In this case, if only single observations are omitted in the cross-validation exercise, hindcast prediction rules may incorporate information that would not be available in an actual forecast situation.

As an illustration, consider an extended set of the same hurricane intensity data. Instead of one observation per storm (as was used in the above two examples) the number of observations now ranges from 1 (Hurricanes Gloria and Floyd) to 27 for Hurricane Josephine in 1990. In this case the total number of observations is 376. However, since 48-h intensity changes for individual storms often represent overlapping periods (generally observations are only 12 h apart), the data are highly correlated. As a result it is inappropriate to cross validate an algorithm by removing only 1 observation at a time. A more accurate estimate of forecast skill is obtained when all data from an entire storm are withheld when performing hindcasts. Table 3 (case 3) contains a comparison of the two procedures using a standard OLS regression algorithm; in one case, cross validation is performed by removing 1 observation at a time, and in the other case, cross validation is performed by removing *storms* one at a time. Serial correlation decreases the errors made with drop-one hindcasts, but these hindcasts incorporate information that would not be available to a forecaster.

One could argue that the comparison in this case is not fair, since if only 1 observation is removed, then 375 remain for use in the algorithm, whereas if an entire storm is removed, then (in general) the number of observations available for choosing a prediction rule is smaller. Differences in cross-validation errors in Table 3 (case 3) could therefore merely represent differences in sample size. To test this possibility the cross validation by observation exercise is repeated, but in addition to removing the observation used in the hindcast error calculation, $N - 1$ randomly selected observations are ignored, where N is the number of observations of the storm from which the original removed observation is taken. In this way the number of observations used in the algorithm is the same as in the case of storm-by-storm cross validation. This procedure results in the same MAE (to the second decimal place) as the original procedure when the process is averaged over 50 trials. Clearly the difference in hindcast skill seen in Table 3 (case 3) is due to serial correlation and not sample size.

4. Summary and remarks

When choosing a statistical prediction rule from a historical dataset, it is important to have an accurate estimate of how well the rule will perform in actual forecasts. This can be done by employing cross validation on the algorithm that is used to choose a prediction rule. The result is an estimate of out-of-sample error that will be more representative of forecast errors than will in-sample error estimates. Straightforward cross validation involves the successive removal of single observations from the dependent sample, with the choice of prediction rules made from the remaining data. These prediction rules are subsequently used to hindcast, in turn, the data points removed. Some algorithm tests may require multiple nested cross validations as illustrated in section 3b and in Livezey et al. (1990), for example.

There are several important considerations when employing this type of error estimation that we hope to have clarified in this article. Cross validation is performed on a rule-choosing algorithm, not on a specific prediction rule. If information from the omitted dependent data is used at any point in the procedure that selects the prediction rule used in hindcasting, then the algorithm is not truly cross validated and forecast skill will be overestimated.

If serial correlation is present in the dependent dataset, then the successive removal of single observations must be replaced by the removal of blocks of observations—the sizes of which are estimated by physical reasoning or by a decorrelation time. In this case a hindcast is made using only data from outside of the removed block to generate the prediction rule. Failure to account for autocorrelation in a cross-validation exercise will bias the estimate of forecast skill upward because in real-time forecasting only the effect of serial correlation from the preceding observation is available, as opposed to simple cross validation in which both past and future observations offer their skill-enhancing serial correlation effects.

When cross validation is applied with care and caution it will, in general, produce error estimates that

accurately represent the level of skill expected on future data. The general principles apply regardless of the complexity of the algorithm that chooses a prediction rule. For more sophisticated algorithms the potential for misapplication increases. Indeed, with the suite of techniques such as nonparametric regressions, neural networks, genetic algorithms, etc., that are now available for forecasting from data, it is more important than ever that researchers clearly understand how to estimate the forecast skill of empirical models accurately.

Acknowledgments. The authors are grateful to M. DeMaria of the Hurricane Research Division for the data used in this study. Special thanks for careful reviews of earlier drafts are extended to S. Applequist, R. Correa-Torres, C. Dartland, J. Lamm, S. Murphy, and R. Treadon. The careful reviews of anonymous referees enhanced the overall readability of this report. The first author received partial support for this work from NOAA through the Cooperative Institute for Tropical Meteorology (CITM).

REFERENCES

Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate,* **5,** 1316–1345.

——, and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977.

Cane, M. A., S. E. Zebiak, and S. C. Dolan, 1986: Experimental forecasts of El Niño. *Nature,* **321,** 827–832.

Livezey, R. E., A. G. Barnston, and B. K. Neumeister, 1990: Mixed analog/persistence prediction of United States seasonal mean temperatures. *Int. J. Climatol.,* **10,** 329–340.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.