ESTABLISHED CRITERIA AND SELECTED METHODS FOR EVALUATING CROP YIELD MODELS IN THE AgRISTARS PROGRAM

Wendell W. Wilson and Jeanne L. Sebaugh
Statistical Reporting Service, USDA

INTRODUCTION

Eight criteria have been established in the AgRISTARS Program to guide the evaluation and comparative analysis of crop yield models. The establishment of these criteria closely corresponds to the spirit of Dr. Hartley's ASA Presidential Address, particularly section three entitled "The Cooperation Between Statistician and Subject Matter Specialist" (1980). In our case the cooperation is between statisticians and two types of subject matter specialists. One type includes commodity specialists or the users of the output of crop yield models as a tool in assembling agricultural intelligence about the production of major crops both in the United States and worldwide. The other type of specialists are those who develop crop yield models or methods of forecasting prospective crop supplies and estimating harvested production (interestingly, most model developers are not statisticians).

Because the users of crop yield models are interested primarily in average yields over rather large areas, the criteria to be discussed might seem rather narrow-minded. This is particularly true with respect to the model developers who often develop models capable of generating farm management and alternative production practice advisories. The criteria reflect the users' needs by concentrating on passive (non-intervention) monitoring of the prospective or actual yield. The information developed will, of course, impact on various market adjustments, including international trade.

AgRISTARS PROGRAM

AgRISTARS is an acronym for Agriculture and Resource Inventory Surveys Through Aerospace Remote Sensing. The six year research program, which began in October 1980, involves five federal agencies. AgRISTARS main thrust is to develop an early warning system able to detect conditions affecting crop production and quality and to provide techniques for more accurate domestic and foreign commodity production forecasts. AgRISTARS research will be concentrated on eight crops grown in the U.S., Canada, USSR, India, Australia, Argentina and Brazil. All analysis techniques will be developed and initially tested in the United States where information to confirm the validity of the techniques is available.

The goal of the program is to determine the usefulness, cost and extent to which aerospace remote sensing and associated technologies can be integrated into existing and future USDA systems to improve the objectivity, reliability, timeliness and adequacy of information required to carry out USDA missions. To achieve this goal the AgRISTARS technical program is structured into eight projects. The Yield Model Development Project (one of the eight) will support the main thrust of the AgRISTARS Program by developing mathematical models using environmental and plant measurement characteristics that represent the yield potential of various crops. Models that have utility for both forecasting and estimation will be developed for the crop/country combinations specified by the USDA. Previous work has demonstrated that climate/crop yield models can successfully provide early, mid and late season (at harvest) yield indications with varying degrees of accuracy. These yield estimates can be combined with independently derived (often based upon satellite remote sensing imagery) crop area estimates to estimate crop production for a selected region.

Specifically, the Yield Model Development Project will support USDA crop production forecasting and estimation efforts by:

1. Testing, evaluating and selecting crop yield models for application testing.
2. Identifying areas of feasible research for improvement of models.
3. Conducting research to modify existing models and to develop new crop yield assessment methods.

AgRISTARS TEST AND EVALUATION CRITERIA

Eight criteria have been established to guide the test and evaluation of crop yield models in the AgRISTARS program: yield indication reliability, objectivity, consistency with scientific knowledge, adequacy, timeliness, minimum cost, simplicity and provision of accurate current indications of modeled yield reliability. The criteria might more appropriately be thought of as desirable characteristics that successful models should possess (from the USDA point of view). Each criterion's importance and the rationale for its use will now be described.

Yield indication reliability is a measure of the degree to which users can rely on crop yield indications from a model as a source for setting official yield forecasts and estimates, and in using them as a basis for policy determinations. Users will often have multiple sources of information. They need to know how much confidence they should have in each source.

Tests of yield indication reliability over a period of years usually involve independent tests to measure such things as: the mean square error; variance; bias; proportion of years beyond a critical error limit; worst and second to worst performance during the testing period; range of accuracy; direction of change from mean yields, previous year yields and from (any) earlier current year forecasts; and the simple correlation coefficient between actual and model predicted yields for a set of independent test years. For the most part this criterion can be evaluated quantitatively.

Objectivity is a more difficult criterion to evaluate quantitatively. In fact, some subjectivity will invariably be involved in the evaluation of yield models for this characteristic. A truly objective crop yield model requires no

subjective judgments which involve adjusting the model form, parameters or input variables. Subjectivity may have been involved in model development, but for the fixed model all parameters and input variables are "measurable," methods of estimation or derivation are fully documented, and the model is exactly repeatable under the same conditions. Even though model users may wish to apply a high order of subjective judgment in using various sources of information to arrive at a yield figure, it is still desirable that to the extent possible all yield information sources be objective. Greater objectivity will allow users to more fully understand each model's characteristics, limitations and capabilities.

Measures of model departures from objectivity can be made by a three step process. First, the model is examined to identify and count the number of subjective judgments required in its use. Then, the degree of subjectivity of each judgment, that is, the repeatability or consistency of each judgment is appraised. The third step involves examining the sensitivity of model yield indications to the subjectively determined values.

Consistency with scientific knowledge can not be examined to its full extent by statisticians alone. The more in depth appraisals of this characteristic will obviously require a high degree of cooperation between statisticians and subject matter specialists. Agreement or consistency of a crop yield model's form and parameter values with experimental data and scientific knowledge is an important criterion in model selection. The sensitivity of model yield to important environmental inputs is an important measure of model capability and acceptance. Understanding when, or under what conditions a model might not be consistent with known physical and biological responses is important.

Consistency with scientific knowledge can be examined in at least three areas. The form of each candidate model, its parameter signs and values can be examined for agreement with available experimental data and, in general, scientific knowledge. Sensitivity analyses can be conducted to evaluate model logic and reasonableness. Models can also be examined for the absence of important environmental inputs or the predominance of a few inputs such that other important variables have a minimal impact.

Adequacy of crop yield models can be assessed in terms of the extent of geographic coverage of a crop, the level of detail provided and in the appropriateness of the model for intended future applications. A model with greater coverage of important producing areas, regions and countries is considered more adequate. Limitations of coverage will often be related to unavailability of or inaccurate measures of input variables in some areas. More detail could provide yield and associated production information for smaller geographic subdivisions, for various production systems and for different crop utilization groupings. In crop production aggregation, the provision of yield indications for the same strata used in estimating crop area is desirable. Appropriateness of a model for intended future applications will be constrained by data availability in some foreign areas and lack of geographic detail for domestic use. Even though the immediate application (perhaps for a domestic test) may be within the coverage of a model, the extendibility of an adapted model to areas of potential future application will be considered.

Timeliness constitutes availability of sufficiently precise or accurate crop yield information at the time when the information is needed and can be used. Timing of a season's first forecast is determined by the coincidence of two factors, (1) a yield indication is needed and (2) a reasonably reliable indication can be provided by the model. Timeliness of subsequent forecasts and estimates will be related to need and when significant updates in earlier forecasts are possible. Evaluation of timeliness often is a matter of answering the question, "Can the model provide a useful forecast or estimate by the required date?" Models that do not meet this criterion for an early season forecast application will not necessarily be excluded from consideration for later forecasts. However, some consideration may be given in model selection to the economy of selecting models with similar input data requirements for all forecast and estimation dates.

Minimum cost is obviously a very desirable characteristic for successful crop yield models. Cost of the operating system associated with a model will be the primary consideration in comparing the cost of candidate models. Cost of operating models will be appraised for various types of activities. Some of these activities are: acquiring, formatting and using historic data bases to estimate model parameters; acquiring and updating current year values in a timely manner for model execution; those associated with the need for frequent model updates, number and kind of variables and general model complexity; and transferring the operating model to a different computer system, if necessary.

Simplicity is a desirable model characteristic. If two models were equal for the other seven criteria, then it is suggested that one would, of course, select the simpler model. Simplicity in crop yield model form and use of input data are often associated with cost. A very important aspect of model similicity is the ability of the user to understand the concept, capabilities and limitations of the model. A thorough understanding allows the user to evaluate the model's indication in the light of other information and make valid judgments. A simpler model would generally have lower user training and experience requirements.

The availability, at the time of model use, of a model generated indication of the reliability of the model's yield point estimate is desirable if it provides any information on the actual reliability of that point estimate. This provision of accurate current indications of modeled yield reliability will be appraised for its availability and utility for each candidate yield model. The degree to which such an indication (when available) corresponds to subsequently determined actual performance will be assessed. The basic task is to ascertain the degree to which the user can depend upon a model's indication of reliability for guidance on the degree of confidence to be placed in that model's yield indication.

In general, these current indications of

modeled yield reliability may reflect where current model input values are with respect to the base period range of input variables, the accuracy with which input variables are measured, temporal and spatial variability of variables, and the underlying population variability. These indicators of reliability might, in some cases, reflect departures from model assumptions in a particular year and therefore be indicative of yield indication reliability. For regression models the model standard error of the predicted yield has often been used to provide the indication of current reliability. The standard error is a function of the residual mean square for the model base period and the distance of the current independent variable values in the prediction year from their average during the base period. Indications of current reliability will be retrospectively compared to measures of the difference between predicted and actual yields and the strength of this relationship will be evaluated.

While the application of the first and last of the criteria (those dealing with reliability) are most amendable to quantitative statistical techniques, objectivity and consistency with scientific knowledge may involve statistical procedures.. The other four criteria--adequacy, timeliness, minimum cost and simplicity--will probably involve little direct use of statistical methods.

PURPOSES WHICH CROP YIELD MODEL TEST AND EVALUATION SERVE

There are basically three objectives for developing test and evaluation criteria, in applying them to evaluate individual crop yield models and using the evaluations as a basis for comparing alternative models. They are:

- o to provide guidance in the selection of promising crop yield models for application testing in the AgRISTARS Program and for use by USDA;
- o to provide a common reference for describing the performance (and likely future performance) of models in terms of their capabilities and limitations;
- o to aid in identification of model deficiencies and areas of feasible research for improving performance of both selected and non-selected models.

This last objective is probably the most important in terms of creating the basis for improved tools for accomplishing USDA's mission of improved agricultural intelligence on worldwide crop production.

SELECTED STATISTICAL PROCEDURES FOR APPLYING TWO CRITERIA

The criterion of yield indication reliability is examined through the use of various indicators and the application of statistical tests. Indicators of yield reliability (described below) require that the parameters of the model be estimated for a set of data and that a yield prediction be made based on that data for a given "test" year. The values required to generate indicators of yield reliability include the predicted yield, $\hat{Y}$, the actual (reported) yield, Y, and the difference between them, $d = \hat{Y}-Y$, for each test year. It is desirable that the data used to estimate the parameters for the model not include data from the test year.

In order to accomplish this, a "bootstrap" technique is used (Wilson, et. al., 1980). Years from an earlier base period are used to fit the model. A predicted yield is generated for the following year. Then, the base period is shifted one year forward and the process is repeated. Continuing in this way, ten predictions of yield are obtained, each independent of the data used to fit the model. The Y, $\hat{Y}$ and d values for the ten-year test period may then be summarized into various indicators of yield reliability.

From the d value, the mean square error (root and relative root mean square error), the variance (standard deviation and relative standard deviation), and the bias (its square and the relative bias) are obtained (see Appendix-Statistical Formulas).

The root mean square error (RMSE) and the standard deviation (SD) indicate the accuracy and precision of the model and are expressed in the original units of measure (quintals/hectare). Accurate prediction capability is indicated by a small RMSE.

A non-zero bias means the model is, on the average, overestimating the yield (positive bias) or underestimating the yield (negative bias). The SD is smaller than the RMSE when there is non-zero bias and indicates what the RMSE would be if there were no bias. If the bias is near zero, the SD and the RMSE will be close in value. We prefer a model whose bias is close to zero.

The relative difference, $rd = (100\ d/Y)$, is an especially useful indicator in years where a low actual yield is not predicted accurately. This is because years with small observed actual yields and large differences often have the largest rd values.

Several indicators are derived using relative differences. In order to calculate the proportion of years beyond a critical error limit, we count the number of years in which the absolute value of the relative difference exceeds a critical limit, say 10 percent. The worst and next to worst performance during the test period are defined as the largest and next to largest absolute value of the relative difference. The range of yield indication accuracy is defined by the largest and smallest absolute values of the relative difference.

Another set of indicators demonstrates the correspondence between actual and predicted yields. It would be desirable for increases in actual yield to be accompanied by increases in predicted yields. It would also be desirable for large (small) actual yields to correspond to large (small) predicted yields.

Two indicators relate the change in direction of actual yields to the corresponding change in predicted yields. One looks at change from the previous year (nine observations) and the other at change from the average of the previous three years (seven observations). A base period of three years is used since a longer base period would further decrease the number of observations, while a shorter period would not be very

different from the comparison to a single previous year.

Finally, the Pearson correlation coefficient, r, between the set of actual and predicted values for the test years is computed. It is desirable that $r(-1 \leq r \leq +1)$ be large and positive. A negative r indicates smaller predicted yields occurring with larger observed yields (and vice versa).

Model performance may be compared using these indicators of yield reliability. However, it is also desirable to run a statistical test comparing the reliability of competing models. A formal statistical test considers the variability of model performance over time and allows the user to specify an upper limit on the probability of incorrectly declaring one model better than another. This probability is known as α, the level of significance, or the Type I error.

However, because of the manner in which models are chosen for testing, it is challenging to construct a meaningful statistical test. Only yield models which have been presented in the literature or developed by known experts are considered. Therefore, a priori, great differences between the reliability of the models are not expected. A powerful statistical procedure is needed which is able to detect small, although important, differences in reliability. Also, the test should be able to function well with relatively small samples of data for each model, say ten years.

The test should also perform well when only two models are being compared. Often only two models of a particular type, for example, two monthly weather data models or two daily weather data models, are competitive and available for testing. When models of different types are to be compared, it is unlikely that all possible model comparisons will be made. It is more likely that the best models of each type will be compared.

It would appear that an F test could be useful in comparing the mean square errors of two models. However, if the mean square errors are based on ten years of test data and $\alpha = .05$, then one model's mean square error must be four times larger than anothers before the models can be declared different. This is an unreasonable requirement since models which are in the evaluation process will almost always be more competitive than this.

A test may be constructed by considering that one model is considered more reliable than another model if its predicted yields, $\hat{Y}$'s, are closer to the actual yields, Y's. No difference in the reliability of two models for a particular year means that the absolute value of the difference between their predicted yields and the actual yield is the same. The reliability of a model for that year is related to the amount of the discrepancy, not its direction. We may define $|d_1| = |\hat{Y}_1 - Y|$, $|d_2| = |\hat{Y}_2 - Y|$, and $D = |d_1| - |d_2|$. Then the models are equally reliable in a year for which D equals zero. If D is not equal to zero, one model is more reliable than the other for that year. In formal terms, we want to test the null hypothesis that there is no difference in the reliability of the models over all years. To do so the values of D from the ten test years may be used to compute a test statistic and a decision made whether or not to reject the null hypothesis. Since the results for the models are paired each year, paired-sample statistical tests are used.

Two types of paired-sample statistical tests are used: a parametric test using the student "t" test statistic and a nonparametric test using the Wilcoxon signed rank test statistic. One reason for applying both tests is that they require different assumptions. The parametric t-test assumes the D values are normally distributed while the nonparametric test does not. The d values may be considered to be approximately normally distributed. The $|d|$ values would then be folded normals rather than normally distributed. Although both models are folded at $|d| = 0$, their means may be different and the distribution of D has a possibility of not being normally distributed. The t-test is robust with respect to the normality assumption; however, this possible violation of the assumption is one reason for also running the non-parametric test.

The other reason for running both tests concerns the conditions under which the null hypothesis is rejected by each test. Using the parametric test, the basis for rejecting the null hypothesis is the average size of the D values as compared to their variability. The hypothesis will be rejected and the model with the smaller $|d|$ values declared more reliable if t is large (either positive or negative). However, it is possible that one model could have a smaller $|d|$ value for each of the test years, in other words, be very consistent in outperforming the other model, and still the null hypothesis may not be rejected by the parametric test unless the average value of D is large enough.

Using the nonparametric test, the null hypothesis will always be rejected if one model has smaller $|d|$ values for each of the test years, regardless of the magnitude of the D values. Therefore, if the models are very competitive in terms of the $|d|$ values each year, but one model consistently, although slightly, outperforms the other model, the nonparametric test will still declare the consistent model to be more reliable.

The hypothesis of equal model performance will only be rejected by the nonparametric test if one model has more years with smaller $|d|$ values than the other model. The model with more smaller $|d|$ values is considered the more reliable model in terms of consistency of performance. However, to reject the null hypothesis and declare one model clearly better than another, consistency of performance is not a sufficent requirement (although it is necessary). Consider the situation in which one model is more consistent than the other but the largest D values occur when the less consistent model performs better. In the few years the less consistent model performs better, it performs much better. A dilemma exists since one model is more consistent than the other but the biggest differences between the models occur when the consistent model performs worse. The null hypothesis will not be rejected and the consistent model will not be declared better if this situation occurs. The null hypothesis will be rejected only if one model is more consistent and the biggest differences between the models occur

when the consistent model performs better.

The other criterion or model characteristic to be discussed here is its ability to provide an accurate, current measure of modeled yield reliability. Although a specific statistic was not discussed in the paper, Crop Yield Model Test and Evaluation Criteria, (Wilson, et al., 1980), it was stated that:

> "This 'reliability of the reliability' characteristic can be evaluated by comparing model generated reliability measures with subsequently determined deviation between modeled and 'true' yield."

For regression models, this suggests the use of a correlation coefficient between two variables generated for each test year. One variable is an indicator of the precision with which a prediction for the next year can be made, based on the model development base period. The other variable (obtained retrospectively) is an indicator of how close the predicted value for the next year actually is to the "true" value. The estimate of the standard error of a predicted value from the base period model, $s_{\hat{Y}}$, is often used for the first value and the absolute value of the difference between the predicted and actual yield in the test year is used as the second variable, $|d|$.

A non-parametric (Spearman) correlation coefficient, r, is employed since the assumption of bivariate normality cannot be made. A positive value of $r(-1 \leq r \leq +1)$ indicates agreement between $s_{\hat{Y}}$ and $|d|$, i.e., a smaller (larger) value of $s_{\hat{Y}}$, is associated with a smaller (larger) value of $|d|$. An r value close to +1 is desirable since it indicates that a small standard error of prediction (and therefore a narrow prediction interval about the yield being predicted) is associated with small discrepancies between predicted and actual yields. If this were the case, one would have confidence in $s_{\hat{Y}}$ as an indicator of the accuracy of $\hat{Y}$.

A model generated reliability measure other than $s_{\hat{Y}}$ could be suggested for use. In particular, non-regression models will need to provide some measure in order to be evaluated based on this criterion.

NUMERICAL EXAMPLES OF STATISTICAL PROCEDURES FOR EVALUATING AND COMPARING CROP YIELD MODELS

Table 1 shows the results from a ten-year bootstrap test for two spring wheat yield models in two North Dakota (ND) Crop Reporting Districts (CRDs). One of the regression models, called Straw Man, is a simple linear regression of yield over time. The other model was developed by the Center for Environmental Assessment Services (CEAS) and estimates yield as a function of trend and monthly weather-related variables.

Table 2 shows the values of the various indicators of yield reliability. In both CRDs and for each indicator, the CEAS model exhibits better performance than the Straw Man model. The results for the parametric and non-parametric statistical tests were similar in that the performance of the models could not be declared different in CRD 10 but could in CRD 20. Neither model demonstrated the ability to provide an accurate, current measure of modeled yield reliability using $s_{\hat{Y}}$. Values of the Spearman correlation coefficient were 0.36 and 0.27 for the Straw Man model and were -0.20 and -0.12 for the CEAS model in CRDs 10 and 20 respectively. Instances of years with smaller prediction intervals about the yield being predicted were all too often associated with larger observed discrepancies between the actual and predicted values.

CONCLUSION

As Dr. Hartley pointed out in the JASA article, previously referred to, there are two types of possible errors in applying statistical techniques. A paragraph from the article is quoted below.

> "Now it is usually accepted without question that the input of the subject matter specialist is vital. On the other hand, we witness with some concern the tendency to question the role of statisticians. This tendency is fanned by the fact that there are indeed many problems that are not amendable to statistical treatment. We statisticians must certainly refrain from forcing statistics into a problem where it is not needed. But the error 'of the second kind' that our colleagues often commit is not to use statistics at all when it is badly needed. How many failures in the public sector would have been avoided by a timely statistical analysis, revealing risks of an unreasonable order of magnitude?"

In our efforts to test and evaluate crop yield models we certainly do run the risk of applying statistics where it is not needed. This is particularly true for some of the criteria. However, we must also be aware of the pitfall of not using statistical techniques at all in cases where they are badly needed.

REFERENCES

Hartley, H. O., 1980. The Cooperation Between Statistician and Subject Matter Specialist. JASA, 75: 1-7.

Wilson, Wendell W., Barnett, Thomas L., LeDuc, Sharon K., Warren, Fred B., 1980. Crop Yield Model Test and Evaluation Criteria. AgRISTARS Yield Model Development Project, Document YMD-1-1-2 (80-2.1).

Table 1

Bootstrap Test Results For Spring Wheat Yields
in North Dakota Comparing Straw Man and CEAS Models

CRD	Test Year	Actual Yield (Q/H)	Predicted Yield (Q/H)		d=Predicted -Actual (Q/H)	
			Straw Man	CEAS	Straw Man	CEAS
10	1970	16.2	18.1	17.0	1.9	0.8
	1971	20.0	18.5	20.6	-1.5	0.6
	1972	19.9	19.4	24.5	-0.5	4.6
	1973	20.1	20.2	16.7	0.1	-3.4
	1974	14.8	21.2	15.8	6.4	1.0
	1975	16.7	20.7	17.5	4.0	0.8
	1976	17.6	20.4	17.6	2.8	0.0
	1977	16.5	20.3	14.5	3.8	-2.0
	1978	21.9	19.7	20.4	-2.2	-1.5
	1979	14.5	20.5	15.7	6.0	1.2
20	1970	14.9	18.3	16.9	3.4	2.0
	1971	20.7	18.6	19.4	-2.1	-1.3
	1972	19.2	19.7	19.2	0.5	0.0
	1973	19.8	20.3	19.2	0.5	-0.6
	1974	12.9	21.1	15.6	8.2	2.7
	1975	16.4	20.5	16.1	4.1	-0.3
	1976	16.4	20.3	18.0	3.9	1.6
	1977	14.8	19.9	16.5	5.1	1.7
	1978	19.7	18.9	18.3	-0.8	-1.4
	1979	16.6	19.2	17.3	2.6	0.7

Table 2
Indicators of Yield Reliability
For Spring Wheat Yields in North Dakota

Indicator of Reliability (Unit)	CRD 10		CRD 20	
	Straw Man	CEAS	Straw Man	CEAS
Bias = B (Q/H)	2.08	0.21	2.54	0.51
Relative Bias = RB (%)	11.7	1.2	14.8	3.0
Mean Square Error = MSE $(Q/H)^2$	12.12	4.30	14.91	2.13
Root Mean Square Error = RMSE (Q/H)	3.55	2.07	3.86	1.46
Relative Root Mean Square Error = RRMSE (%)	19.9	11.6	22.5	8.6
Variance = Var $(Q/H)^2$	8.29	4.26	8.46	1.87
Standard Deviation = SD (Q/H)	2.88	2.06	2.91	1.37
Relative Standard Deviation = RSD (%)	14.5	11.7	14.8	7.8
Percent of years $\|RD\| > 10\%$ (%)	70.	30.	70.	30.
Largest $\|RD\|$ (%)	43.2	23.1	63.6	20.9
Next Largest $\|RD\|$ (%)	41.4	-16.9	34.5	13.4
Smallest $\|RD\|$ (%)	0.5	0.0	2.5	0.0
Range $\|RD\|$ (%)	42.7	23.1	61.0	20.9
Percent of Years Direction of Change from the previous year in the $\hat{Y}$'s agrees with the Y's (%)	33.	78.	33.	78.
Percent of Years Direction of Change from the average of the previous three years in the $\hat{Y}$'s agrees with the Y's (%)	14.	71.	43.	100.
Pearson correlation coefficient between Y and $\hat{Y}$.	-0.39	0.70	-0.40	0.92

APPENDIX - STATISTICAL FORMULAS

Measures of Model Performance

Definition of Terms:

Y_i = Yield as reported by U.S.D.A. for year i ("true" or "actual" yield).

$\hat{Y}_i$ = Yield as predicted by a model for year i.

$d_i = \hat{Y}_i - Y_i$ = difference between predicted and actual yield for year i.

$rd_i = 100\ d_i/Y_i$ = relative difference for year i.

s_{Y_i} = Standard error of regression = (Residual or Error Mean Square from Model Development Base Period)$^{1/2}$ for year i.

$s_{\hat{Y}_i}$ = Standard error of a predicted value for year i = $s_{Y_i}\ (1 + \underline{x}_o'(\underline{X}'\underline{X})^{-1}\underline{x}_o)^{1/2}$, where $\underline{X}$ is the regression design matrix of independent variable values and $\underline{x}_o$ is the vector of independent variable values for the year the prediction is being made.

$i = 1, \ldots, n$ = number of test years and $\Sigma = \sum_{i=1}^{n}$ = summation over the test years.

$\bar{Y} = 1/n\ \Sigma Y_i$ = average actual yield.

Measures:

Bias = $B = 1/n\ \Sigma\ d_i = \bar{d}$.

Relative Bias = $RB = 100\ B/\bar{Y}$.

Mean Square Error = $MSE\ 1/n\ \Sigma\ d_i^2$.

Root Mean Square Error = $RMSE = (MSE)^{1/2}$.

Relative Root Mean Square Error = RRMSE = $100\ RMSE/\bar{Y}$.

Variance = $Var = 1/n\ \Sigma\ (d_i - \bar{d})^2$.

Standard Deviation = $SD = (Var)^{1/2}$.

Relative Standard Deviation = $RSD = 100\ SD/(\bar{Y} + \bar{d})$.

Mean Square Error = Variance + $(Bias)^2$, or

Accuracy = Precision + $(Bias)^2$.

Pearson r between $\hat{Y}_i$ and Y_i:

$$r = \frac{\left[\Sigma\hat{Y}_i Y_i - \frac{(\Sigma\hat{Y}_i)(\Sigma Y_i)}{n}\right]}{\left[\left(\Sigma\hat{Y}_i^2 - \frac{(\Sigma\hat{Y}_i)^2}{n}\right)\left(\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}\right)\right]^{1/2}}$$

Spearman r between $|d_i|$ and $s_{\hat{Y}_i}$:

Let $R(|d_i|)$ = the rank of $|d_i|$, $R(s_{\hat{Y}_i})$ = the rank of $s_{\hat{Y}_i}$, and $f_i = R(|d_i|) - R(s_{\hat{Y}_i})$, $i = 1, \ldots, n$. Then, $r = 1 - \frac{6\Sigma f_i^2}{n^3 - n}$.

Paired-Sample Statistical Tests Comparing the Performance of Two Crop Yield Models

Definition of Terms:

$\hat{Y}_{1_i}$ = Yield as predicted by model 1 for year i.

$\hat{Y}_{2_i}$ = Yield is predicted by model 2 for year i.

$|d_{1_i}| = |\hat{Y}_{1_i} - Y_i|$ = Absolute value of the difference between model 1 predicted and actual yield for year i.

$|d_{2_i}| = |\hat{Y}_{2_i} - Y_i|$ = Absolute value of the difference between model 2 predicted and actual yield for year i.

$D_i = |d_{1_i}| - |d_{2_i}|$.

Rank $(|D_i|)$ = Ranks of the absolute values of D_i assigned in ascending order (smallest value of $|D_i|$ = rank 1, ..., largest value of $|D_i|$ = rank n). If two or more years have the same value for $|D_i|$, assign each year the average of the ranks.

Parametric Test - Student t:

H_0: $\mu_D = 0$

H_a: $\mu_D = 0$

Test Statistic = $t = \frac{\bar{D}}{s_{\bar{D}}}$, where

$\bar{D} = 1/n\ \Sigma D_i$,

$s_{\bar{D}} = (s_D^2/n)^{1/2}$, and

$s_D^2 = \{\ \Sigma D_i^2 - 1/n\ (\Sigma D_i)^2\}\ /(n-1)$.

Reject H_0 if $|t| > t_{\alpha,\ (n-1)}$.

Nonparametric Test - Wilcoxon Signed Rank:

H_0: One model does not perform better than the other model.

H_a: One model performs better than the other model.

Procedure to compute test statistic, T:

1. Compute the D_i.
2. Assign ranks to $|D_i|$.
3. Assign signs to Rank ($|D_i|$) corresponding to the signs of D_i.
4. Let T = the absolute value of the sum of the ranks with the less frequent sign (corresponding to non-zero D_i).

Reject H_0 if $T < T_\alpha$ (1 tailed), n.