HOW ARE CROP YIELDS DISTRIBUTED?

Authors: Bailey Norwood, Postdoctoral Student, Department of Agricultural and Resource Economics, North Carolina State University.

Matthew Roberts, Assistant Professor, Department of Agricultural, Environmental, and Development Economics, The Ohio State University.

Jayson Lusk, Assistant Professor, Department of Agricultural Economics, Mississippi State University.

Contact: Bailey Norwood
Box 8109
NCSU-ARE
Raleigh, NC 27695-8109
Phone: 919-515-8946
Fax: 919-515-1824
Email: fbnorwoo@unity.ncsu.edu

Paper presented at the American Agricultural Association Meetings.

July 28-31, 2002

Long Beach, California

HOW ARE CROP YIELDS DISTRIBUTED?

Bailey Norwood

Matthew Roberts

Jayson Lusk

Keywords: yield distribution, model selection, risk, uncertainty, crop insurance

Abstract

Six popular crop yield distributions are compared to determine which best describes yield fluctuations out-of-sample. For 183 crop and county combinations, each distribution is estimated and ranked according to its log-likelihood function observed at out-of-sample observations. A semiparametric model dominates the contest for all crops and most counties, likely due to its flexibility and treatment of heteroskedasticity. Most other models ranked lower because their variance equation performed poorly out-of-sample.

HOW ARE CROP YIELDS DISTRIBUTED?

Characterizing the behavior of crop yields is an important component of agricultural economic analysis. Accurate knowledge of crop yield behavior is critical in devising farm management tools, farm policy, and crop insurance. For example, low yields may occur during periods of low commodity prices, causing financial distress on the farm sector. Knowledge of the likelihood of such events is necessary for government programs to respond with appropriate policy. However, crop yields can be extremely variable from year to year, perhaps much more than the output of non-agricultural firms. Although understanding the stochastic nature of crop yields is critically important in agricultural economic analysis, characterizing yield distributions can be quite difficult.

How are crop yields distributed? There is no one answer for every crop region, and although we will never know a crop's true distribution, there are ways of measuring model accuracy. Many innovative yield distributions have been offered over the last three decades, six of which are discussed here. This study seeks to understand which of this six models best describe yield fluctuations in out-of-sample settings.

The literature is filled with various candidate yield distributions, few of which can be excluded on theoretical grounds. How should one discern which is best? Previous studies rely almost exclusively on hypothesis tests for model discrimination using in-sample fit. Though useful in many regards, as with any other approach it has its disadvantages. Usually, if a set models are not significantly different the most parsimonious is chosen, but sometimes appeals to parsimony cannot be made. For instance, if a three parameter beta distribution is not significantly different from a three parameter normal distribution, which one should be chosen?

While acknowledging the potential usefulness of hypothesis tests in many settings, this paper considers an alternative approach to model selection extendable to more settings, and is more consistent with the purpose of crop yield distribution estimates. This alternative approach assigns an unambiguous ranking to almost any number and type of yield models, where models are ranked based on their out-of-sample performance. For the most part, yield distributions estimated from historical data are used for extrapolating into the future. For instance, crop insurance premiums are set by estimating yield distributions from historical yields and assuming that distribution will carry over to next year's yield. The analysis of farm policies often make similar assumptions. Thus, it seems natural to rank yield distributions by their out-of-sample performance.

The second and third sections discuss six popular crop yield models and an appropriate methodology for discriminating among them. This model selection criterion is referred to as the Out-of-Sample-Log-Likelihood Function (OSLLF) approach. The fourth section pits each candidate yield model in three contests. These contents are designed to illustrate which model is best across different crops and regions, along with an indication of how much confidence one can place in that chosen model being "best." Overall, this paper contributes to our knowledge of yield distributions firstly by describing a method of model selection particularly suited to yields. Second, estimating yield distributions and performing one's own validation tests can be time consuming.

Much time can be saved in future work by using the results of this study's model rankings.

CANDIDATE YIELD DISTRIBUTIONS

Long ago it was discovered that yields can exhibit various sorts of behavior. Mean yields may be increasing in yield, or it may not. The same can be said for yield variance.[1] Furthermore, yields may be negatively skewed, positively skewed, or symmetric (Day) and even exhibit bi-polarity (Goodwin and Ker). In response, agricultural economists have developed a myriad of rich models allowing any of these descriptions, but implying none of them. Articles from the American Journal of Agricultural Economics from 1990-2000 were reviewed to reveal six popular yield models, each to be analyzed here.

Gallagher and Nelson and Preckel utilize the concept of a maximum attainable yield. Gallagher describes this maximum yield as a time trend, while Nelson and Preckel model it as a constant. Gallagher's model states deviations of yield from its maximum value (which is time dependent) as a gamma distribution, and is thus referred to as the *GAMMA* model. Heteroskedasticity is accounted for in his estimation routine.[2] Nelson and Preckel model deviations from its maximum value (which is not time dependent) as a two parameter beta distribution, and is referred to as the *BETA* model. In its original form, the *BETA* model conditioned the two beta parameters on agricultural inputs, such as nitrogen use and soil characteristics. To facilitate comparison with the other five models, these parameters are instead conditioned upon a time trend.[3] This specification naturally allows heteroskedasticity and a time-dependent mean yield.

Moss and Shonkwiler describe yields as a time trend, but allows the parameters of this trend to be random according to a Kalman Filter. Instead of estimating deviations from this trend as a normal distribution, they estimate a function of those deviations as normal. This function is called an inverse hyperbolic sine transformation of normality and requires two additional parameters. Depending on the parameter values, yield may be positively or negatively skewed or symmetric and may exhibit kurtosis. This model is referred to as *STOCHIHS* where the *STOCH* portion designates yields to be a function of stochastic parameters and the *IHS* portion indicates the use of an inverse hyperbolic sine function.

The *STOCHIHS* model does not allow heteroskedasticity because the Kalman Filter becomes intractable. Ramirez modifies this model to allow heteroskedasticity by replacing the stochastic trend with a fixed-parameter trend, in addition to several other creative reparameterizations. This new model allows a positive covariance between different yields[4] and so becomes a multivariate distribution. This model's name is *MULTIHS*, as it is multivariate and still utilizes an inverse hyperbolic sine function for non-normality.

A simpler but more flexible model is that offered by Goodwin and Ker. This model, denoted *SEMIPAR* for semiparametric, portrays percent deviations of yield from its mean with a nonparametric kernel smoother.   Mean yields are estimated from an ARIMA model[5] making the mean yield component parametric and the remaining portion nonparametric. The kernel smoother is applied to percent deviations of yield from its mean, rather than raw deviations, to account for heteroskedasticity.[6]

A recent article by Just and Weninger suggests that previous findings of skewed yield distributions may be the result of inappropriate detrending and failure to account for heteroskedasticity properly. When using flexible polynomial trends for mean yield and yield variance the authors find that normality is difficult to reject. This last model is referred to as the *NORMAL* model.[7]

There exists plenty of evidence for considering each six models as a candidate for use in yield distribution estimates. The first five models; the *GAMMA, BETA, STOCHIHS, MULTIHS,* and *SEMIPAR* are flexible enough for yield to exhibit a wide array of behavior. The *NORMAL* model is more restrictive, but as illustrated by Just and Weninger, is often difficult to reject. This study is not concerned with which models are significantly different, but rather which one is best for a particular crop and region. Besides, hypothesis tests would be difficult to apply in this situation. No two models are nested implying non-nested tests would have to be employed. But non-nested tests are famous for ambiguous conclusions as they can reject all models and often fail to reject more than one model. Plus, some of these models contain the same number of parameters, making appeals to parsimony infeasible.

The next section describes a different way of looking at model selection. Statistically, it is based on the Kullback-Leibler Information Criterion, but more importantly ranks models by their out-of-sample performance, as yield distributions are ultimately used for making probability statements about the future.

THE OUT-OF-SAMPLE-LOG-LIKELIHOOD FUNCTION (OSLLF) APPROACH TO MODEL SELECTION

Ranking yield models by their out-of-sample performance is worthwhile for two reasons. First, sample sizes are typically low, making it easy to over-fit models. Second, as mentioned previously, many yield models are used for making probability statements about future yields, so it seems natural to rank these models by their ability to describe yields post-sample. By far, the most common method of ranking models according to out-of-sample performance is prediction error. Each yield model has an implied expected value. Prediction error measures the distance between this expected value and actual yields for a series of out-of-sample yields. Examples are out-of-sample-root-mean-squared error; average-absolute-out-of-sample error; and the Ashley, Granger, Schmalensee test (Brandt and Bessler; Kastens and Brester; Norwood and Schroeder).

Be it useful in many instances, in regards to selecting yield distributions, considering only forecast errors leaves much to be desired. Prediction error alone does not account for how well a model captures variance, skewness, kurtosis, and probabilities in general. If the purpose of yield distributions is to generate probability statements, we must consider more than just prediction error. Put differently, we are not just interested forecasted yields relative to observed yields, but forecasted probability statements relative to observed yields. The entire distribution should be considered in the model ranking.

An alternative is to rank models by their out-of-sample-log-likelihood function (OSLLF) values, as likelihood functions also measure fit but take into account the entire distribution. Let $L_j(Y_t|\theta)$ be the log-likelihood function value from Model j and parameter vector θ. A log-likelihood function is then the sum of $L_j(Y_t|\theta)$ over a set of Y_t's. Judging models by their log-likelihood function values has many nice features to be discussed shortly, but first, a discussion of how $L_j(Y_t|\theta)$ can be considered "out-of-sample" must be given.

For $L_j(Y_t|\theta)$ to be "out-of-sample" the parameter vector θ cannot be estimated with information on yields at time t (Y_t).[8] Suppose there are T yield observations; Y_1, Y_2, …, Y_T. Cross-validation entails calculating $L_j(Y_t|\theta)$ where θ is estimated using every observations except Y_t, and is appropriately denoted θ_{-t}. An OSLLF value from cross-validation can then be denoted $\sum_t L_j(Y_t|\theta_{-t})$. Grouped-cross-validation is similar, except one or more yields are excluded in addition to Y_t. For instance, the OSLLF value for $Y_{t=15}$ may be calculated using observations $Y_{t=1}$, …$Y_{t=15-3}$ and $Y_{t=15+2}$, …, Y_T. An OSLLF value using grouped cross-validation is denoted $\sum_t L_j(Y_t|\theta_{-((t-i)\rightarrow (t+j))})$. The grouped method will tend to choose more parsimonious models (Shao). The decision of whether to use cross- or grouped-cross-validation, and if the grouped method is pursued which groups to leave out, depends on the true data generating process, which is never known.

Consider the following arguments for using the OSLLF approach for selecting yield distributions. First, we know that models should not be ranked according to their in-sample-log-likelihood function values, as one will tend to pick incorrect models over correct models simply because they have more parameters.[9] The Akaike Information Criterion (AIC) avoids this by penalizing a model's likelihood function value for each parameter it employs.[10] This penalty works: In large samples and under certain conditions (Sawa), it will pick the distribution closest to the true distribution (Akaike).

Stone has shown that when cross-validation is used, the OSLLF value is asymptotically equivalent to the AIC value. This means that the OSLLF approach avoids overfitting and no penalty parameter is needed. However, under certain conditions,[11] both the AIC and OSLLF calculated using cross-validation will pick models with too many parameters, although both are better than using in-sample-log-likelihood functions. The correction for the AIC in these cases is to increase the parameter penalty (Sawa). The correction for the OSLLF is to use grouped-cross-validation (Shao) with an increasing number of observations "left out at a time." Unfortunately, these certain conditions requiring larger penalty parameters or more observations to be left out are never based on observable quantities.

The OSLLF shares the asymptotic properties of the AIC and similar criteria (Stone; Shao). One of these properties is that it chooses the model with the highest information content, as measured by the Kullback-Leibler Information Criterion.[12] Asymptotic properties are nice, but yield samples are typically small, so it is natural to explore the small sample properties of the OSLLF. Previous studies providing a simulation-based comparison between the OSLLF and other model selection criteria in small samples was performed. The simulations were designed to mimic crop yield distributions, and results suggested that the OSLLF picks the true yield distribution with a higher frequency than other methods considered (Norwood, Lusk, and Roberts; Norwood, Ferrier, and Lusk).[13]

For these reasons, the OSLLF approach is a desirable method of ranking models based on out-of-sample fit and will be used to discriminate among the six models mentioned previously. Three contests are conducted to determine which model ranks best across various settings. Before these contests take place, some method must be used to assess how much confidence can be placed in the rankings. This study wishes to ask: For any model ranked highest for a particular crop and region, in repeated samples, how often would this model continually be selected? A nonparametric technique is used which proceeds as followed. Let $\text{OSLLF}_{t,j}$ be the OSLLF value for a yield observation in time t using Model j, OSLLF_j be a vector of those values for Model j, and t = 1, …, T. Suppose Model j =1 is ranked highest because $\text{OSLLF}_1 > \forall\ \text{OSLLF}_{j>1}$. If the vector OSLLF = [OSLLF_1 OSLLF_2 … OSLLF_6] with T rows where each row corresponds to a particular t, a new simulated matrix *SIM*OSLLF is created by randomly selecting rows of OSLLF with replacement. The rows of OSLLF are randomly picked, rather than individual values of $\text{OSLLF}_{t,i}$, because correlations between $\text{OSLLF}_{i,t}$ and $\text{OSLLF}_{j,t}$ $\forall$ i,j seem likely.

A variable called *IFCHOOSE* is created which equals one if the highest ranked model from the simulated OSLLF values is the highest ranked model from the original estimation (if the highest ranked model is still MODEL 1) and zero otherwise. This exercise is repeated 1,000 times. If the value of (*IFCHOOSE*)/(1,000) equals one, then we can say that in repeated samples we would expect to choose Model 1 100% of the time and would therefore place great confidence in that model. Conversely, if

(*IFCHOOSE*)/(1,000) equals 50%, we would say this model is not “truly dominant” over the set of remaining models.

This is analogous to a test for the alternative hypothesis that Model 1 has a higher Kullback-Leibler Information Criterion value than all others (with the null being its value is equal to the largest of all the others) and 1-(*IFCHOOSE*)/(1,000) is the p-value. Or, it can be interpreted as a test for the null hypothesis that Model 1 will be chosen 50% of the time, against the alternative hypothesis that it will be picked more than 50% of the time, with the test statistic being $2T^{1/2}[(IFCHOOSE)/(1{,}000) -1/2]$ (Mendenhall, Wackerly, and Schaeffer). This test does not depend on any of the models being the true, or any other assumption about the true data generating process.

HOW ARE CROP YIELDS DISTRIBUTED? RESULTS OF THREE MODEL RANKING CONTESTS

This section pits the six models in multiple contests to determine which one tends to be ranked highest across various regions and crops. In each contest, the models' out-of-sample-log-likelihood function (OSLLF) is calculated using grouped-cross-validation. The sample size for each contest is divisible by five. First, the models are estimated using observations six and larger, and then used to calculate OSLLF values for the first five observations. Then, observations one through five and eleven and larger are used for the estimation and calculation of OSLLF values six through ten. This continues until the OSLLF values for the last five observations are calculated from parameters that were estimated using all previous observations. This amounts using grouped-cross-validation "leaving five out at a time." This choice is based on recommended procedures for grouped-cross-validation.[14]

The first contest utilizes data for Cornbelt corn, soybeans, and wheat yields for 1950-1989 available in the Appendix of Ramirez. This data was used by Ramirez because the three yields are likely correlated and, since his yield distribution is multivariate, allows a determination of whether taking correlations among crops into account will improve model performance. Ramirez finds that while wheat yields appear independent, corn and soybeans yields are correlated.

This study will ask a similar question using a different methodology. All six models are estimated with this data. The five univariate models; *GAMMA*, *BETA*, *STOCHIHS*, *SEMIPAR*, and *NORMAL* are used to obtain OSLLF values for each crop. The OSLLF values are then summed across the three crops to produce a multivariate OSLLF value. Then, the *MULTIHS* model estimates corn, soybeans, and wheat yields jointly to obtain another multivariate OSLLF value.[15] If *MULTIHS* is ranked higher than the other six, this is evidence that accounting for correlation across crops may improve the accuracy of forecasted probability statements.

Table 1 shows that the semiparametric model, *SEMIPAR*, is ranked highest. Outliers had a huge influence on this ranking though. Both the *GAMMA* and *BETA* models have the lowest possible OSLLF value of negative infinity due to the manner in which they represent the maximum attainable yield. Both models place values on the highest value yield can take, which according to both models, the probability of a yield above this value is zero. However, in out-of-sample forecasts there were yields that exceeded this ceiling, making the log-likelihood function negative infinity. Excluding these models based on one outlier may seem harsh, but remember, both models said the probability of yield exceeding a particular level was zero, and yield did exceed that level. It is doubtful researchers will want to use a model so confident, and so wrong.

The *NORMAL* model performed poorly due to its variance equation. A linear trend was chosen to represent heteroskedasticity, however, when forecasting yields in years in 1950-1955 it predicted an extremely low variance. Then, when matched with a rather high prediction error the result was a yield observation the *NORMAL* model said had an extremely low probability of occurring.[16] Notice this would not be reflected in the model ranking if only prediction errors were considered. Again, the model received a low ranking because it was very confident and very wrong.

The low ranking of the *MULTIHS* model was also the result of a few observations (three out of 40 total). Just like the *NORMAL* model, when forecasting yields in 1950-1955 it predicted a low variance, suggesting the yields that did occur had a very low probability. If the model rankings were repeated using only OSLLF values from 1956-1989 the *MULTIHS* would be ranked highest.[17] Further evidence for this can be seen by noticing the median OSLLF value for *MULTIHS* is considerable lower than the others. Table 1 provides a test indicating this lower median is significantly smaller.

If one wishes to rank models excluding outliers, an alternative is to choose the model with the lowest median OSLLF value. Then, to determine the confidence of picking this model in repeated samples the test for significant medians (shown in Table 1) can be employed. If the inclusion of outliers is desirable, as would likely be the case when applied to crop insurance premiums, the OSLLF values summed across all observations should be used. Finally, when applying the bootstrap procedure described in the previous section, the bootstraps suggest that in repeated samples one would choose the *SEMIPAR* model 73% of the time. Being significantly greater than 50% (the test statistic is $2*(1000^{1/2})*(0.73-0.50) = 15$), this implies the *SEMIPAR* model is truly dominant.

The previous contest concerned multivariate yield models; the next two contests concern univariate yield models. For a single crop and region, the *MULTIHS* model is still applicable by setting the correlations across crops and regions to zero.[18] Using county data for corn, soybeans, and wheat, the thirty counties with the largest harvested acreage were selected. For all 90 counties (largest thirty counties for three crops) all yield models were estimated and ranked according to the OSLLF approach.

The results, shown in Table 2, indicate that the semiparametric model proposed by Goodwin and Ker, *SEMIPAR*, is ranked highest more than half the time for each crop. Models *GAMMA* and *BETA* are second best depending on the crop, while the other models are rarely chosen. The next contest chooses thirty counties at random, excluding those counties in the previous contest, for each crop. Results are in Table 2 and are almost identical to the previous one, except that *STOCHIHS* is picked more frequently and the *GAMMA* model less frequently.

Lastly, a single measure of model performance is provided across all 180 counties of the previous two contests (30 counties per crop per contest = 180 counties). For each six models, a variable is created which equals one if the model is ranked highest and zero otherwise. Then, if its value equals one, the variable is multiplied by the variable *IFCHOOSE*. Recall this variable is a measure of confidence that the highest ranked model for that county would continue to be ranked highest in repeated samples. Finally, this variable is summed across all counties for each model and divided by 180/100,[19] then compared across models. Results are reported in Table 3. This is an index of model performance and is not a statistic to be used for hypothesis tests. The best model for issuing probability statements post-sample is *SEMIPAR* which detrends yields with a flexible polynomial, creates a vector of values equaling the percent deviation of yield from its forecast, and applies a kernel smoother to those values.

SUMMARY AND IMPLICATIONS

Many creative candidate yield distributions have been offered by agricultural economists for use in farm risk management, policy analysis, crop insurance, and similar research functions. Discerning which distribution is best for a particular crop or region can be difficult and time consuming though. This study seeks to provide insight into which distributions perform best by offering a desirable method of model selection and applying it to a variety of crops and regions.

The model selection method has two strong advantages compared to conventional methods. First, it ranks models based on their performance out-of-sample, as most distribution estimates' ultimate purpose is extrapolating into the future. Second, it takes into account the entire model specification, and therefore reflects the relative ability to capture mean yield, yield variance, skewness, kurtosis, and other moments of interest. The approach simply requires picking the model with the highest log-likelihood function value when observed at out-of-sample observations. Since all yields models must be stated as probability density functions, this method can rank any number and type of yield models in a manner that exhibits desirable statistical properties.

Six recent and popular models offered by Gallagher; Nelson and Preckel; Moss and Shonkwiler; Goodwin and Ker; Just and Weninger; and Ramirez were ranked according to this method in numerous settings. The first contest asked which model best describes soybean, wheat, and corn yields in the Cornbelt under a multivariate setting. The second and third contests asked which captures yield fluctuations for these same crops in a univariate setting across 180 counties. The second contest used the thirty counties with the highest harvested acreage for each crop, and the third chose thirty counties for each crop at random.

The model developed by Goodwin and Ker dominated all three contests, as it is the highest ranked in over half of the 183 counties analyzed. The model is semiparametric; yields are detrended with a flexible polynomial, and a kernel smoother is applied to the percent deviation in yields from this trend. There are several reasons why this model frequently dominates. First, perhaps its nonparametric nature is best suited for yields because it is the most flexible and makes little a priori distributional assumptions.

Second, it accounts for heteroskedasticity differently. Other candidates specify yield variance (or standard deviation of yield) according to a polynomial trend, and as mentioned in the previous sections, these models often received low rankings because their variance consistently under predicted the true yield variability. Goodwin and Ker's model does not forecast yield variance, but simply assumes the percent deviation of yield from its trend is homoskedastic. Perhaps this latter approach is superior?

This does not imply that the semiparametric model should always be used, nor does it imply that other models are not better suited for a researcher's needs. For instance, if a research agenda can be simplified using the mean-variance approximation of expected utility, the normal model offered by Just and Weninger may be preferred. Alternatively, if the objective is to estimate the maximum value yield can take, Gallagher or Nelson and Preckel's model may be the top choice.

What this study does do is provide practical guidance for selecting yield distributions. In cases where time does not allow researchers to perform their own model validation tests, they can rely on these findings that, most of the time, the semiparametric model offered by Goodwin and Ker issues the most realistic yield probability statements.

FOOTNOTES

1) For instance, of one regresses US corn yields from 1960-2000 against an intercept and time trend, the trend parameter is significantly positive. Though a plot of yields against time seems to suggest an increasing variance, the White test for heteroskedasticity does not support that claim. Conversely, Miller, Kahl, and Rathwell find that South Carolina and Georgia peach yield distributions exhibit a constant mean and variance over this same time period.

2) Gallagher constructs an index for yield variance which is dependent upon a time trend. In the maximum likelihood estimation, each observation is weighted by its predicted standard deviation, very similar to weighted least squares.

3) One might be more familiar with the two parameter (α,β) beta model where the independent variable lies on the (0,1) interval. In the Nelson and Preckel article, α and β are conditional on data in X, and the dependent variable is yield divided by its maximum value. Nelson and Preckel specify $\alpha = aX^b$, and a similar specification for β, where X denotes an agricultural input. This study replaces X with a time trend, but this particular form made convergence in non-linear estimation extremely difficult. Thus, we replaced it with the form $\alpha = a + bt$ where t is a time trend, with an identical form for β. This implies a five parameter beta distribution; one for maximum attainable yield, two for α and two for β. With this form yield may exhibit a time-varying mean and variance.

4) This correlation may be between different crops in the same region, the same crop in different regions, or different crops and regions.

5) One of the authors suggested using either a quadratic or linear trend instead of an ARIMA model. In response a quadratic trend is used unless the quadratic term is insignificant, in which case a linear trend is employed.

6) If Y_t is yield and $E(Y_t)$ is its expected value, the kernel smoother is applied to $(Y_t-E(Y_t))/E(Y_t)$ rather than $(Y_t-E(Y_t))$, because the former is considered a time invariant distribution while the latter is not.

7) A cubic polynomial is used for expected yield, which can be reduced to a quadratic or linear trend if supported by hypothesis tests. Heteroskedasticity is accounted for by modeling the absolute value of ordinary least square residuals as a quadratic trend. The variance equation may be reduced to a linear trend or a constant if hypothesis tests suggest doing so.

8) Authors often include a data matrix in the likelihood notation, such as $L_j(Y_t|\theta,X_t)$. Sometimes this is important because Y_t must be predicted without knowledge of X_t. However, this is not the case here and so this distinction is not made. This study only considers estimating yields from either a time index or yields from other time periods, so the “data” is always known.

9) Likelihood functions are useful model development tools because they indicate probabilities. In some circumstances, a likelihood function can be interpreted as the probability of observing the data assuming the model (and its parameters) are true. However, in-sample likelihood functions are subject to the inclusion of additional, irrelevant parameters which will always increase the likelihood function value. For instance, using the classical linear regression, one could include a unique dummy variable

for each observation and proceed to maximize a normal homoskedastic likelihood function. This method will always result in zero errors and a variance of zero. In this case, the likelihood function will always equal one because the model is a tautology. Plus, it will equal one no matter whether unique dummy variables are replaced with other unique explanatory variables. The likelihood function will fail to provide information because its value is not a statistic--it is determined by the researcher and will equal one with a probability of 1. Note that if this method were pursued and then one calculated an out-of-sample-log-likelihood function (OSLLF), the OSLLF values will not equal one, will differ across models, and will likely be very low. This is because the out-of-sample-likelihood function is a statistic, and the in-sample function is not.

10) The AIC subtracts K from a model's log-likelihood function value, where K is the number of parameters. Other penalties exist, such as the Final Prediction, Schwartz, and Shibata Criterion. These other criteria are derived assuming all candidate models are nested, and the researcher is determining which explanatory variables to include. They are almost always derived under normality and using asymptotic statistics. Since normality is only one possibility for crop yields, these other penalties may not be as desirable as the AIC.

11) That is, under certain formulations for the true and candidate models.

12) If f(Y) is a candidate model and g(Y) is the true model, where both functions are probability distribution functions, the Kullback-Leibler Information Criterion is

$\int \ln\left(\frac{g(Y)}{f(Y)}\right) g(Y) dY$. The smaller the number, the more information. This information measure is non-negative, based on the assumption that g(Y) is always larger than f(Y), unless the candidate model is true (g(Y)=f(Y)), in which case is zero. This is equivalent to saying the model with the highest expected value of f(Y) is the best.

13) Other criteria considered were the Akaike Information Criteria; out-of-sample-root-mean-squared error; and the Chi-Square, Kolomogorov-Smirnov, and Anderson-Darling statistics. The three latter statistics were applied using both in-sample and out-of-sample observations. The OSLLF picked the true model 25% of the time, which was significantly higher than the other eight criteria and a random pick. The true yield distribution could randomly take one of twelve forms in any simulation. Further details are available from the authors by request.

14) If there are thirty (40) observations, this implies 6 (40/5=8) subgroups of out-of-sample observations. Zhang suggests not using less than five subgroups and Shao suggests dropping more than one observation at a time.

15) The form of *MULTIHS* used follows from the restricted full information estimate shown in Table 7 of Ramirez. This form assumes corn and soybean yields are correlated but wheat yields are not. Following Ramirez, wheat yields are assumed normally distributed while the other two are non-normal.

16) Recall that heteroskedasticity is not a maintained hypothesis of this model, but is determined by the data. See Footnote 7.

17) Using years 1956-1989, the OSLLF values for each model are (1) *GAMMA* = -311 (2) *BETA* = -infinity (3) *STOCHIHS* = -313 (4) *SEMIPAR* = -312 (5) *NORMAL* = -333 (6) *MULTIHS* = -232.

18) In this contest, the *MULTIHS* model was always estimated under the maintained assumption of non-normality.

19) This ensures the index lies in the (0,100) interval.

REFERENCES

Akaike, H. "Information Theory and an Extension of the Maximum Likelihood Principle." *Proceedings of the 2nd International Symposium on Information Theory*. Edited by N. Petrov and F. Csadki. Budapest. Akademiaai Kiado, 1972. Pages 267-281.

Brandt, Jon A. and David A. Bessler. "Price Forecasting and Evaluation: An Application inAgriculture." *Journal of Forecasting*. 2(1983):237-248.

Day, Richard. "Probability Distributions of Field Crop Yields." *Journal of Farm Economics*. 47(1965):713-41.

Gallagher, Paul. "U.S. Soybean Yields: Estimation and Forecasting With Nonsymmetric Disturbances." *American Journal of Agricultural Economics*. 69(November 1987):796-803.

Goodwin, Barry and Alan P. Ker. "Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group-Risk Crop Insurance Contracts." *American Journal of Agricultural Economics*. 80(February 1998):139-153.

Just, Richard E. and Quinn Weninger. "Are Crop Yields Normally Distributed?" *American Journal of Agricultural Economics*. 81(May 1999):287-304.

Kastens, T.L., and G.W. Brester. "Model Selection and Forecasting Ability of Theory Constrained Food and Demand Systems." *American Journal of Agricultural Economics*. 78(2)(March 1996):67-80.

Mendenhall, William. Dennis D. Wackerly, and Richard L. Scheaffer. Mathematical Statistics With Applications. Fourth Edition. PWS-Kent Publishing Company. 1990.

Miller, Stephen E., Kandice H. Kahl, and P. James Rathwell. “Revenue Insurance for Georgia and South Carolina Peaches.” *Journal of Agricultural and Applied Economics.* 32(2000):123-132.

Moss, Charles B. and J.S. Shonkwiler. “Estimating Yield Distributions With a Stochastic Trend and Nonnormal Errors.” *American Journal of Agricultural Economics*. 75(November 1993):1056-1062.

Nelson, Carl H. and Paul V. Preckel. “The Conditional Beta Distribution As a Stochastic Production Function.” *American Journal of Agricultural Economics.* 71(May 1989):370-377.

Norwood, Bailey, Jayson Lusk, and Matthew Roberts. “A Comparison of Crop Yield Distribution Selection Criteria.” Presented at the 2002 Agricultural Economics Southern Meetings in Orlando, Florida. February 2-6, 2002.

Norwood, Bailey, Peyton Ferrier, and Jayson Lusk. “Model Selection Using Likelihood Functions and Out-of-Sample Performance.” Proceedings of the NCR-134 Conference of Applied Commodity Price Analysis, Forecasting, and Market Risk Management, 2001.

Norwood, Bailey and Ted C. Schroeder. “Usefulness of Placement Weight Data in Forecasting Fed Cattle Prices.” *Journal of Agriculture and Applied Economics*. 32(April 2000):63-72.

Ramirez, Octavio A. “Estimation and Use of a Multivariate Parametric Model for Simulating Heteroskedastic, Correlated, Nonnormal Random Variables: The Case of Corn Belt Corn, Soybean, and Wheat Yields.” *American Journal of Agricultural Economics*. 79(February 1997):191-205.

Sawa, Takamitsu. “Information Criteria For Discriminating Among Alternative Regression Models.” *Econometrica.* 46(1978).

Shao, Jun. “Linear Model Selection by Cross-Validation.” *Journal of the American Statistical Association.* 88:422(1993):486-494.

Stone, M. “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion.” *Journal of the Royal Statistical Society*. *Series B (Methodological)*. 39:1(1977):44-47.

Zhang, Ping. “On the Distributional Properties of Model Selection Criteria.” *Journal of the American Statistical Association*. 87:419(1991):732-737.

TABLE 1
MODEL RANKING RESULTS FOR CORN, SOYBEAN, AND WHEAT YIELDS IN THE CORNBELT

Let Y_t be yield at time t
$L_i(Y_t)$ be the out-of-sample-log-likelihood function value at Y_t
$S_t^{i,j} = 1$ if $L_i(Y_t) > L_j(Y_t)$ and 0 otherwise
The "i's" represent the model in column and "j's" represent the model in row

	Out-of-Sample-Log-Likelihood Function (OSLLF)	Average OSLLF Value	Median OSLLF Value	Minimum OSLLF Value	Maximum OSLLF Value
SEMIPAR	**-355**	**-8.88**	-8.54	-14.72	-7.05
STOCHIHS	-363	-9.08	-8.67	-14.43	-6.69
MULTIHS	-3,622	-90.56	**-6.40**	-1716	-4.10
NORMAL	-27,065	-676.62	-8.91	-26,405	-5.48
GAMMA	-infinity	-infinity	-8.60	-infinity	-7.12
BETA	-infinity	-infinity	-8.79	-infinity	-7.81

Test For Significant Differences in Median Out-of-Sample-Log-Likelihood Function Values.
Null hypothesis is the median of the model in column is equal to the median of the model in row.

Test Statistic = $\dfrac{\sum_{t=1}^{T} S_t^{i,j} - T/2}{1/2\sqrt{T}}$ *and is approximately standard normal under the null hypothesis.*[a]

	SEMIPAR	*STOCHIHS*	*MULTIHS*[a]	*NORMAL*	*GAMMA*	*BETA*
SEMIPAR		0.32	-5.38	0.95	0.95	2.85
STOCHIHS			-5.38	0.95	1.26	1.26
MULTIHS[b]				5.38	5.96	5.38
NORMAL					-0.32	1.26
GAMMA						1.26

Note: These are multivariate models, meaning the OSLLF values correspond to the probability of the corn, soybean, and wheat yields being realized simultaneously.
a) See Mendenhall, Wackerly, and Schaeffer (Page 677).
b) The test statistics for the median of *MULTIHS* versus four other models are identical because, for these models, the OSLLF value is smaller at every observation except the years 1950-1952.

TABLE 2
MODEL RANKINGS FOR CORN, SOYBEANS, AND WHEAT FOR VARIOUS CROPS AND REGIONS

Model	Corn	Soybeans	Wheat
Across Thirty Counties With Largest Harvested Acreage of Each Crop			
	Percent of Times Model Is Ranked Highest		
SEMIPAR	**53%**	**53%**	**60%**
STOCHIHS	3%	13%	7%
MULTIHS	3%	0%	0%
NORMAL	3%	3%	0%
GAMMA	20%	23%	13%
BETA	17%	7%	20%
Across Thirty Counties Chosen At Random For Each Crop			
	Percent of Times Model Is Ranked Highest		
SEMIPAR	**53%**	**63%**	**63%**
STOCHIHS	10%	17%	10%
MULTIHS	3%	0%	3%
NORMAL	3%	0%	3%
GAMMA	7%	13%	13%
BETA	23%	7%	7%

Note: Numbers may not add to 100% due to rounding.

TABLE 3
MODEL PERFORMANCE INDEX VALUES

Model	Index Value[a]
SEMIPAR	36
STOCHIHS	6
MULTIHS	1
NORMAL	1
GAMMA	9
BETA	9

a) The model performance index value can take values from 0-100 for each model. It represents the percent of times one would expect to pick each model across all 180 counties from Contests #1 and #2, multiplied by the average percent of times the model, if ranked highest, would continually to be ranked highest in repeated samples. Its values are only relative to other model values, and should not sum to one across all models.