Forecasting the Annual Almond Crop Production in California

Project No.: **12-ABCBOD1-Wang**

Project Leader: Dr. Jane-Ling Wang
Department of Statistics
University of California, Davis
One Shields Avenue
Davis, CA 95616
530.752.2361
janelwang@ucdavis.edu

Project Cooperators and Personnel:
Dr. Neil H. Willits, Senior Statistician, Statistical Laboratory, Department of Statistics, University of California, Davis
Cong Xu, Graduate Student Researcher, Department of Statistics, University of California, Davis

Objectives:

The focus of the project was to answer three questions:

1. What are the statistical operating characteristics of the existing methods for forecasting California almond production?
2. What changes can be made to the existing methods in order to improve their accuracy and precision?
3. Can Nonpareil production be forecast with better accuracy and precision?

Interpretive Summary:

While the results are fairly mathematical, the analyses allow us to identify the most important sources of variability and bias in the crop estimates, for either the total almond crop or specifically for the Nonpareil crop. This allows us to identify some areas in which the estimates could be improved, as well as to some likely bottlenecks which will limit the amount of improvement that is possible.

Materials and Methods:

This work is based on statistical analysis of historical sampling data, based on the annual samples collected by the National Agricultural Statistics Service (NASS). The analyses that were run on this data were carried out primarily using the SAS statistical software (SAS, Inc., Cary, NC), version 9.2.

Results and Discussion:

The annual crop estimates produced by the National Agricultural Statistics Service (NASS) represent a combination of three separate pieces of information:

1. An estimate of the number of nuts per tree (*N*), which is based on extensive sampling of individual almond trees, employing a *random path* methodology to count the nuts within a randomly-selected portion of each tree in the sample.
2. An estimate of the number of trees per acre (*T*) or planting density for almond orchards.
3. An estimate of the number of acres planted (*A*) in almonds, or in a given variety of almonds.

The toal crop estimate is based on the product of these three terms, namely

$$Crop = N \cdot T \cdot A = f(N, T, A)$$

which is adjusted for historical discrepancies between this arithmetic product and the actual total for that year's almond crop. Thus imprecision or bias in the estimation of any of these three components can result in error in estimating the total almond crop for a given year.

Before proposing changes to the method of estimation, it's important to recognize the relative contribution of errors in these three components to the accuracy of the overall crop estimate. The *Delta Method* is a general statistical technique that can be used to approximate the variance of a smooth function of one or more component random variables. In general, the formula says that if the function in question is $f(X_1,X_2,X_3)$, then

$$\operatorname{Var}(f) \approx \sum_i \left[\left(\left.\frac{\partial f}{\partial X_i}\right|_{X=\mu}\right]^2 \operatorname{Var}(X_i) + 2\sum_{i<j}\left(\left.\frac{\partial f}{\partial X_i}\right|_{X=\mu}\left(\left.\frac{\partial f}{\partial X_j}\right|_{X=\mu} \operatorname{Cov}(X_i, X_j)\right)$$

For the data on estimated nut counts, trees per acre and acres in crop, each component variable comes from a different source, and so it's reasonable to assume that the three variables are all uncorrelated, eliminating the need for the covariance terms in this equation. Making this simplification and then dividing by the expected crop estimate gives the equation

$$\frac{\operatorname{Var}(f(N,T,A))}{\mu_{NTA}^2} = \frac{\operatorname{Var}(N)}{{\mu_N}^2} + \frac{\operatorname{Var}(T)}{{\mu_T}^2} + \frac{\operatorname{Var}(A)}{{\mu_A}^2}.$$

From this we see that the contribution of the error in each term to the overall error (i.e., variance) of the crop estimate is the square of the proportional error in each term. Thus for example, if one term, say the nut count estimate, had a 5% error of estimation and another, say the tree density estimate, had a 1% error of estimation, then the error in the nut count would account for 25 times as much of the error in the overall crop estimate. This equation helps place the errors of estimation into perspective, since an improvement in the estimation of an already accurate component variable is apt to have a negligible impact on the accuracy of the overall crop estimate.

Of the three component variables, the least reliable are the estimate of the average number of nuts per tree, since it relies heavily on a yearly sampling effort, and the estimate of the acres in crop, since in that case there's <u>*no*</u> data source that provides a comprehensive and current estimate of this quantity. For these reasons, most of the focus of this research will be directed at these two quantities. By comparison, the numbers of trees per acre are to a large extent standardized, being set in accordance with best orchard management practices, and so the variability in the estimate of this component is apt to be of lesser concern.

The most complicated of the three component estimates is the estimate of the average number of nuts per tree. The sampling of trees for this estimate is intricate and the sampling within a tree is based on the selection of a *random path* through the tree, so that a representative portion of the nuts on a tree can be counted without having to count *all* of them. For this to work, however, you need to be able to "scale up" the count based on the random path to represent an estimate of the number of nuts on the entire tree. The way that this is done in practice is to take the actual nut count on each segment of the path and divide it by the probability that the segment in question would have been included in the random path, based on the protocol used in selecting the path. The way in which a random path is selected is by starting at the trunk of the tree, and proceeding to a series of subsequent branch points. At a given branch point, the probability of selecting a given branch is proportional to the cross-sectional area (CSA) of that branch. This process continues through the path until none of the remaining branches are greater in cross sectional area than a fixed cutoff point, at which point *all* of the remaining nuts on that branch are counted.

The process of "scaling up" the estimate from the random path to represent the uncounted portion of the remaining branch will produce an unbiased estimate of the number of nuts only if the multiplier for the nut count on a given branch estimates the ratio between the total nuts on the remaining branch and the nuts that were on the proportion of the branch that was counted, which will happen only if the number of nuts on a branch is proportional to the cross sectional area of that branch, or

$$\#\{\text{nuts}\} \propto CSA,$$

This is an assumption, but it's one whose validity can be examined based on the data that have been collected. If this assumption is valid, then the following relationship should hold:

$$\ln\{\#nuts\} = b_0 + \ln(CSA) + \epsilon,$$

\

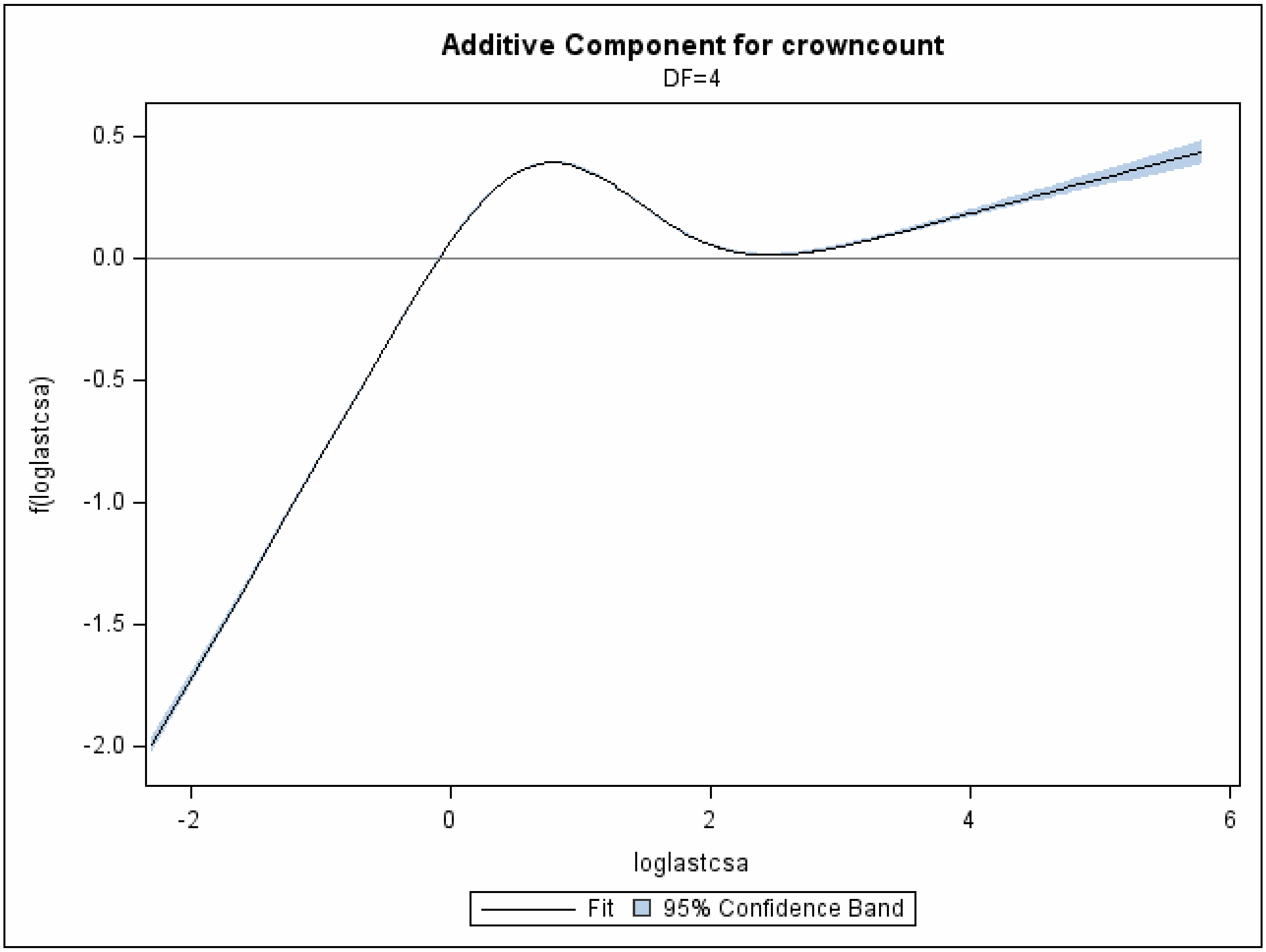
where b_0 is the log of the constant of proportionality.

To assess whether this assumption is consistent with the data, a series of generalized additive models (GAMs) were run on the historical data, in order to look at the relationship between cross sectional area and the observed number of nuts on a branch. A GAM fits a model of the form

$$\ln\{\#nuts\} = f(\ln(CSA)) + \epsilon,$$

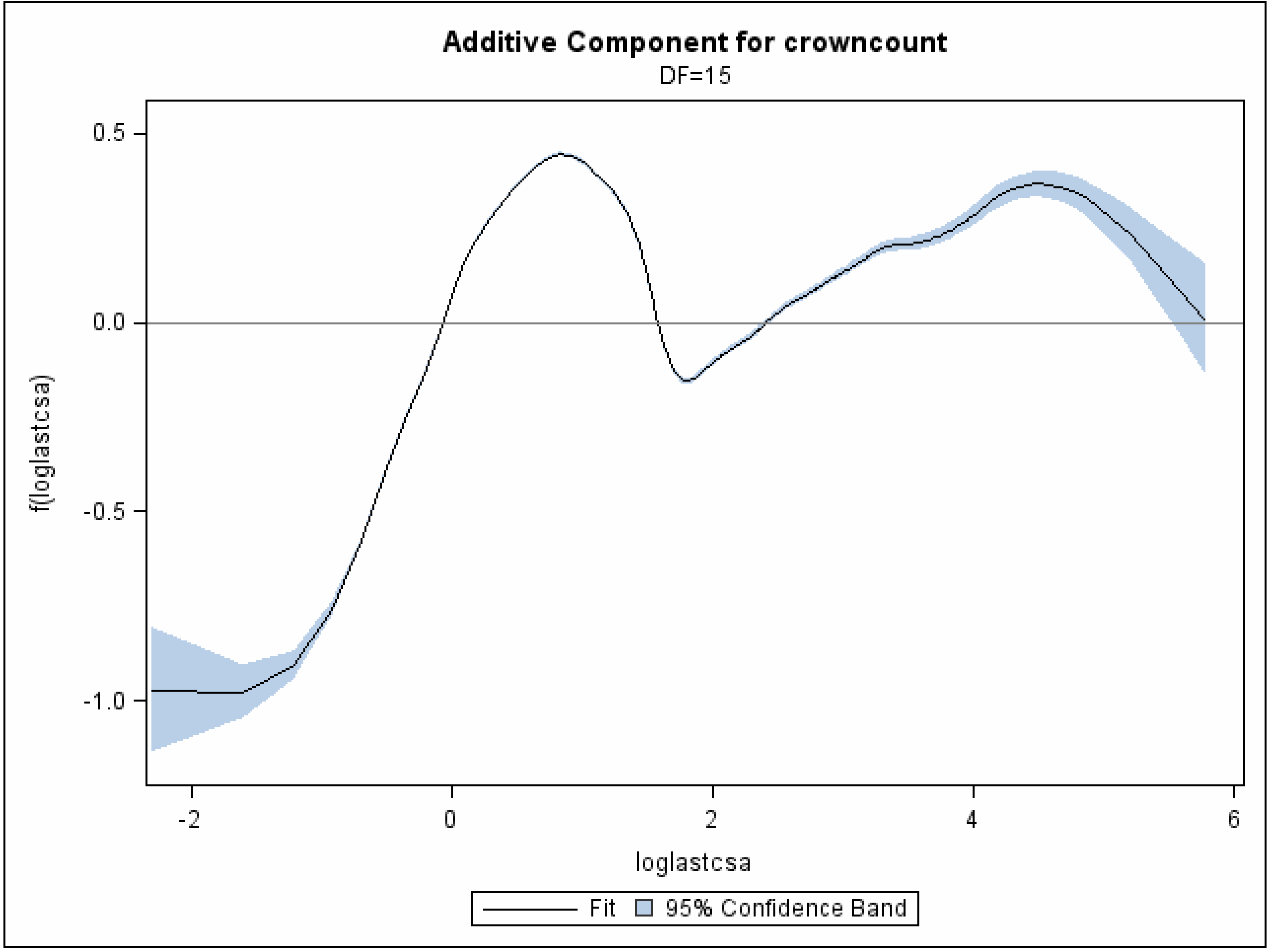
where *f* represents a smooth nonlinear function that's estimated using either spline or local smoothing methods.

It's plausible that the form of this relationship may depend on the position a branch within a tree, and in particular the form may differ between branches that are transitional (that start at one branch point and end at another) and ones that are terminal. For that reason, the initial models of this form were fit to the terminal branches only, since those branches are structurally more similar to each other than to transitional branches. This model was fit first to the random path data from all terminal branches, irrespective of the variety or the age of the tree. The following plot is a *smoothing component plot* that estimates the contribution of ln(CSA) to the response.



If the proportionality assumption was satisfied, then this graph would be roughly linear, with a slope of one. For small cross sectional areas, this is at least approximately the case, since the curve starts around (-2,-2) and passes close to (0,0). However, for larger cross sectional areas, the curve flattens out, and even dips a little. In light of this, the proportionality assumption seems invalid, and so the method used for "scaling up" a nut count for the larger terminal branches needs to be adjusted.

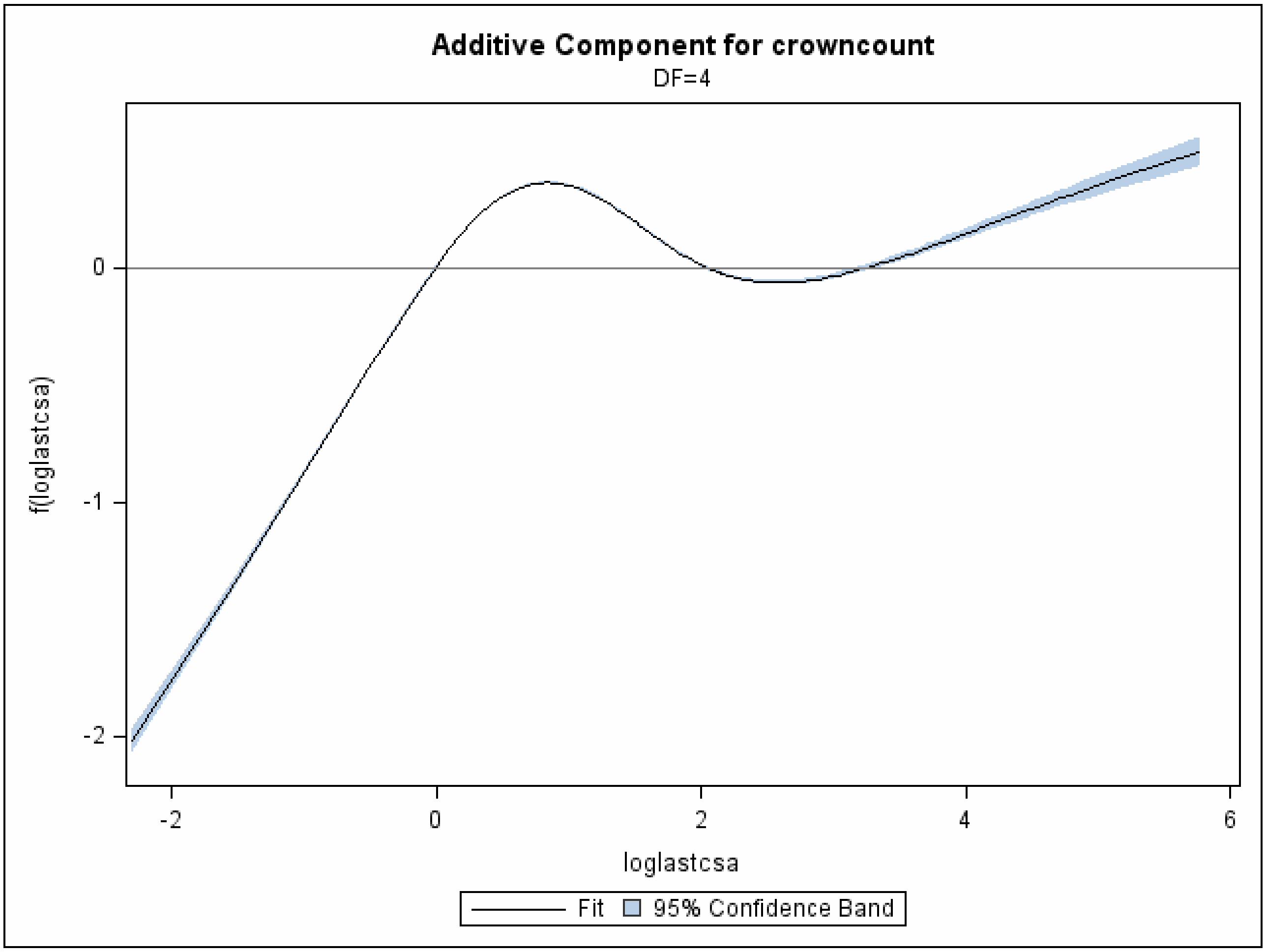
The method used here for fitting a nonlinear relationship was a cubic spline, and a reasonable question to ask is how sensitive the results are to this choice of a smoothing technique. For contrast, the following graph gives a smoothing component plot that was generated from the same data, using a different method, namely LOESS (locally weighted sum of squares). This method produces curves that appear less smooth, but while this graph may superficially look quite different, its broad features are quite similar. For example, there's a roughly linear increase in the curve for small cross sectional areas, a small dip for moderate values, after which the response flattens out:



A question that statistical analysis *can't* address is why the relationship between CSA and nut numbers takes this form. Part of the answer may lie in the fact that a terminal branch with a large CSA isn't typical, since you'd expect there to be additional branch points further along on the branch. The fact that there *aren't* may mean that the branch has been pruned in a way that eliminates those subsequent branch points. Since this question can't be resolved based on statistics alone, NASS may want to try to identify cases for which a terminal branch is surprisingly large in CSA, and investigate the situation further to see what's causing this aberration.

Briefly, additional analysis has been done to look at similar relationships for non-terminal branches. The results (the graphs for which are not presented here) show that for *penultimate* branches (ones that branch *into* the terminal branch), the broad features of this relationship remain, with an initial near-linear increase, followed by a dip, and then a flattening of the relationship. However, for penultimate branches, the near-linear increase applies only to the smallest branches, the dip is somewhat less pronounced, and the response is nearly flat for all but the smallest branches. This implies that while the proportionality assumption is approximately valid for the smallest branches, the proportion of penultimate branches for which this is the case is much smaller than had been the case for terminal branches.

An important question that needs to be addressed is whether these patterns differ in substantial ways for Nonpareil trees. In the analyses done to date, the differences seem to be fairly subtle. For example, for terminal branches, the fitted relationship between CSA and nut counts looks quite similar to the curve estimated based on the data from all varieties:



Still, a more careful (and quantitative) examination of this relationship needs to take place, to see if the adjustments to the “scaling up” procedure need to be calibrated separately for the data from Nonpareil trees, relative to the other varieties.

The next phase of the research on random path sampling is to see whether an adjusted method of calculating nut totals can be extended to provide an improved overall estimate of the almond crop (or the crop for particular almond varieties). This will be done initially on historical data, to get a retrospective sense of whether the revised overall estimates represent an improvement.

Another question that will be addressed in the coming year is whether the estimates of the number of acres planted in a given variety can be improved through the use of Time Series methodology, such as ARIMA (autoregressive integrated moving average) models. In addressing this question, it will be crucial to note whether there are outliers in these predictions, such as those that might be caused by economic upheavals, which would limit the applicability of approaches that are used primarily for stationary time series.

Research Effort Recent Publications:

No publications have been submitted based on this research, due to the fact that the statistical analyses don’t warrant publication and that the data on which the analyses are based are confidential and can’t be disseminated.

References Cited:

None