NASS - Forecasting the Annual Almond Crop Production in California

Project No.: **14-ABCBOD1-Wang**

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

Project Cooperators and Personnel:
Dr. Neil H. Willits, Senior Statistician, UC Davis
Vic Tolomeo, USDA-NASS

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:

USDA-NASS provides an annual estimate of California almond production, which is based on almond counts and weights on randomly-selected trees, planting density, and estimates of crop acreage, across the major varieties represented in the almond crop. This project has identified a number of ways in which the estimates can be improved, primarily to reduce/eliminate several sources of bias in the estimates.

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 on the NORC secure server at the University of Chicago, 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 four 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 average weight of the nuts on a tree (*W*).

3. An estimate of the number of trees per acre (*T*) or planting density for almond orchards.
4. An estimate of the number of acres planted (*A*) in almonds, or in a given variety of almonds.

The total crop estimate is based on the product of these four terms, namely

Crop = W· N·T·A = f(W, 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 four 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 four 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 $f(X_1,X_2,X_3, X_4)$, then

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

With the exception of the data on estimated nut counts and the nut weights, each component variable comes from a different source. Moreover, the evidence for correlation between nut weights and counts is very weak, so it's reasonable to assume that the four component variables are uncorrelated, eliminating the need for the covariance terms in this equation, leading to the following equation:

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

In this equation, the contribution of each component to the overall variability of estimation is the square of the proportional error in each term. So if for the sake of argument, the error in the nut counts was 5% of the mean and the error in acreage was 15% of the mean, then the acreage would be responsible for 9 times as much of the total error in estimation.

Another implication of this formula is that systematic biases are a larger problem than mere imprecision of estimation, since a systematic bias in estimation will persist regardless of the sample size on which the estimate is based, whereas the precision of the estimate will improve if the sample size increases.

Of the four component variables, the least reliable are the estimate of the average number and weight of nuts on a tree, since they rely 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, the main focus of this research has been directed at these 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 random error in the estimate of this component is apt to be of lesser concern.

The most complicated of the four 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 branches 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

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

The validity of this assumption 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.

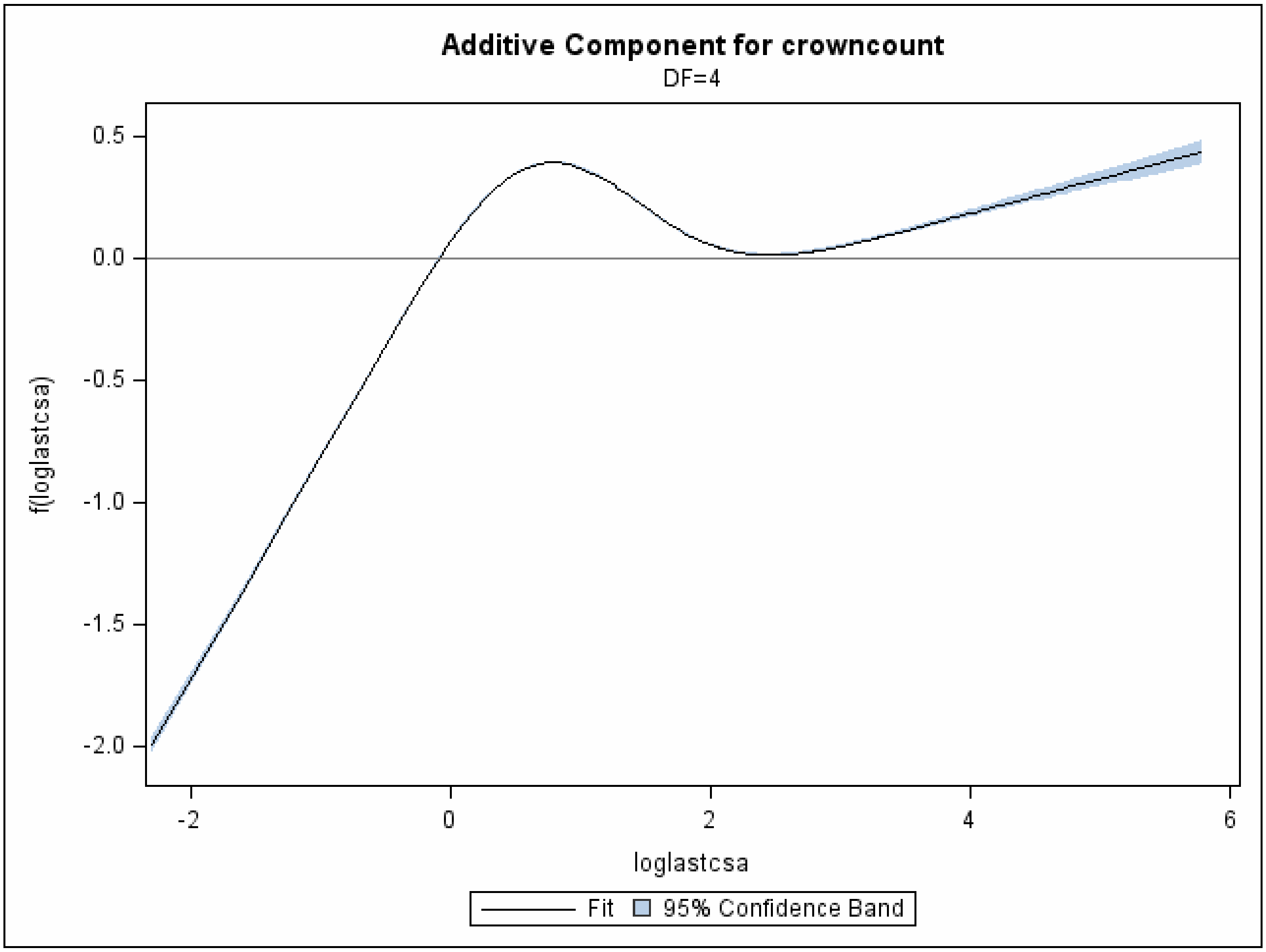
To assess whether this assumption is consistent with the data, a series of *generalized additive models* (GAMs) were run on the historical data, in order to look at the relationship between cross sectional area and the observed number of nuts on a branch. A GAM fits a model of the form

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

where *f* is a smooth, but otherwise fairly arbitrary, nonlinear function that's estimated using either spline or local smoothing methods.

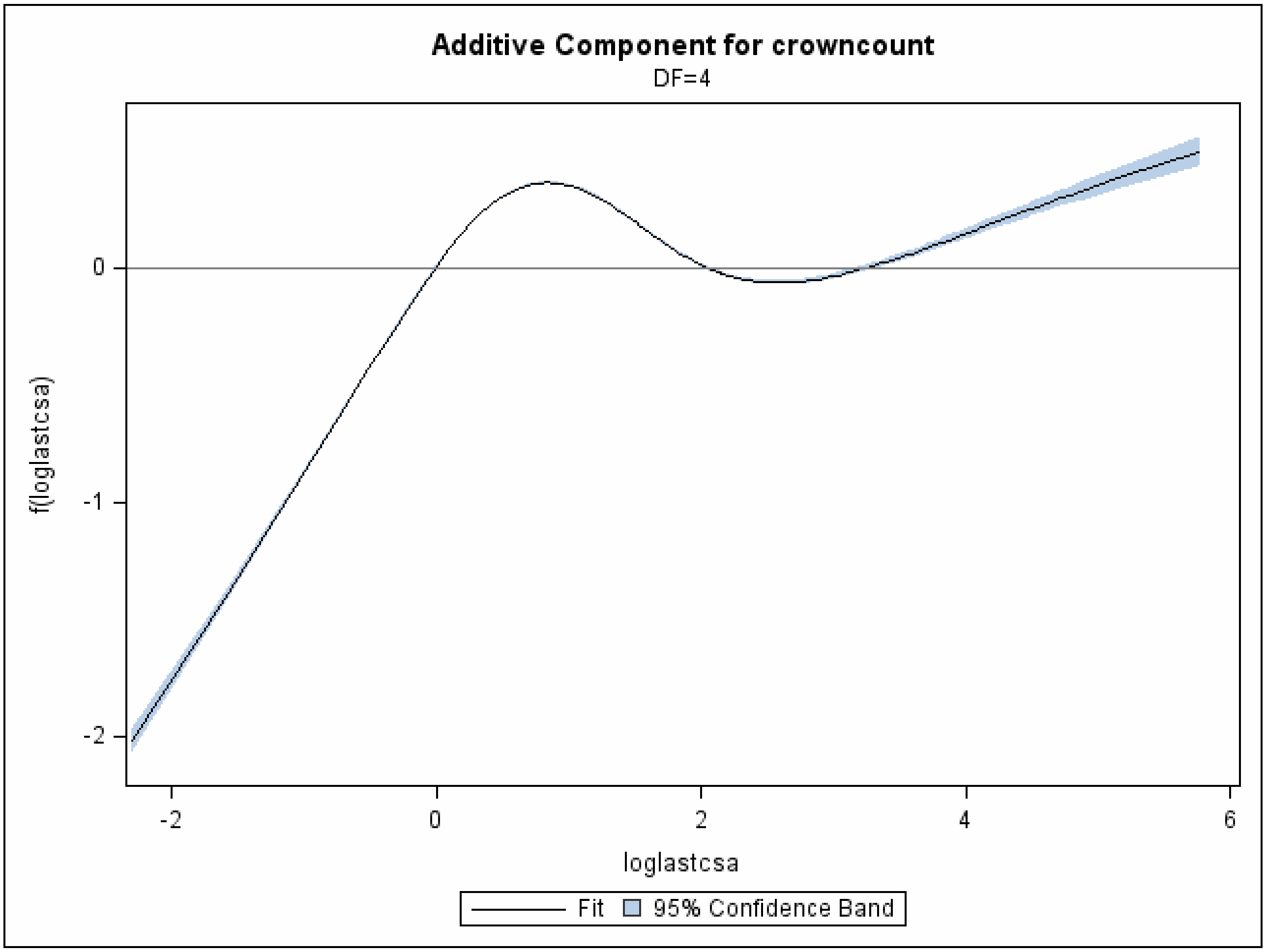
For the purpose of these analyses, the counts on different segments of the random path were examined separately. Rather than starting at the trunk and working outward, we decided to start at the end of the path (in the crown of the tree) and work inward. The main reason for doing this is because empirically most of the nuts are toward the outside (crown) of the tree.

One such plot is given next, looking at nut counts in all terminal branches across all almond varieties in the sample:



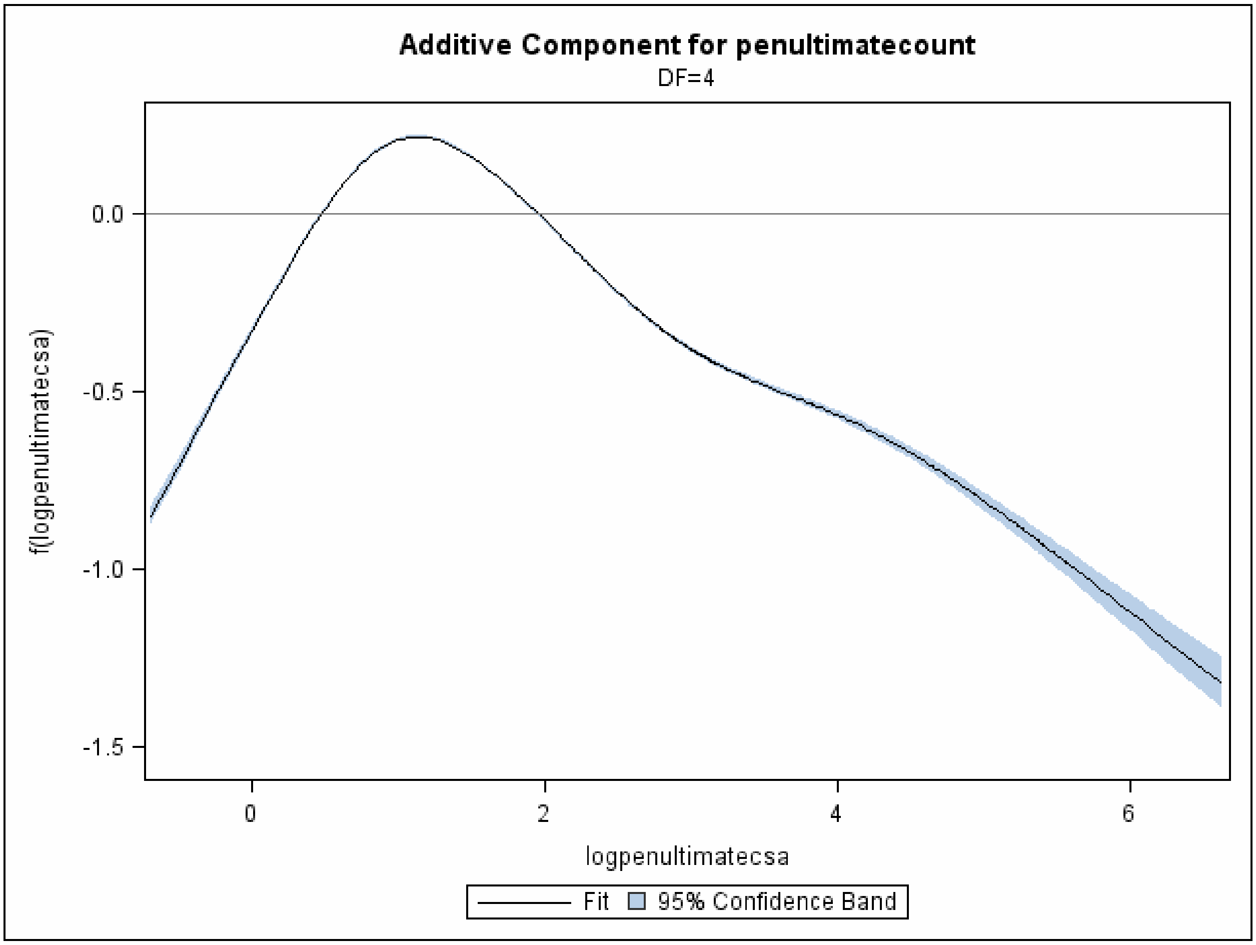
While this graph is approximately linear (with slope close to one) for small branch sizes, the observed number of nuts levels off for larger branch sizes. For trees in which the terminal branch had a large cross sectional area, any method that "scales up" the nut count from the random path to the entire tree assuming that CSA and nut counts are proportional to each other will be appreciably biased for large terminal branches.

Since one of the focuses of the research was to determine whether the Nonpareil crop estimate could be adjusted to improve the estimate, a similar analysis can be done that's restricted to the data from Nonpareil trees. The results of this analysis are qualitatively very similar, as seen in the following graph:



In fact, when similar analyses are done for a range of other varieties, the results were again qualitatively quite similar. This suggests that the bias in estimation is a problem in general, but that the extent of the problem is similar for Nonpareil trees as it is for other varieties.

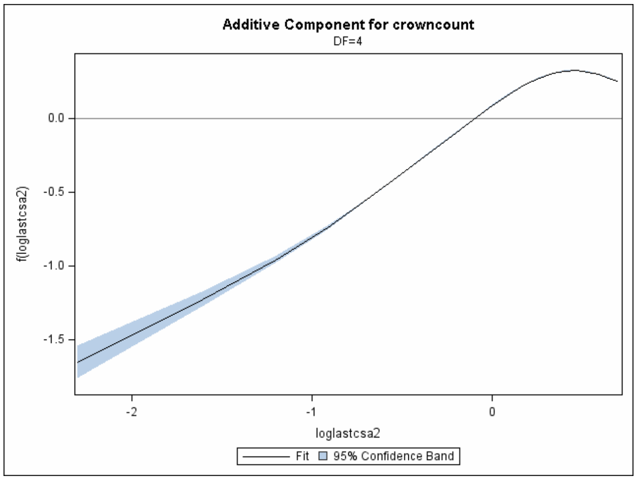
It's also possible to run similar analyses for branches that are in the interior rather than the crown of the tree. Here a different relationship is seen, in which there's a linear increase for *very* small branches, a peak and then the expected nut count ***decreases*** beyond that point. The cross sectional area where the relationship reaches its maximum is slightly larger than for the terminal branches, but since interior branches have larger cross sectional areas as a whole, the extent of the bias is larger here. As with the terminal branches, similar relationships are seen for specific varieties in penultimate branches (the ones before the terminal branch in a random path). Moreover, the relationships for branches that are even further from the crown of the tree are qualitatively similar to what's observed for penultimate branches. The following plot is an example of these relationships, calculated across all varieties, and focusing on penultimate branches:

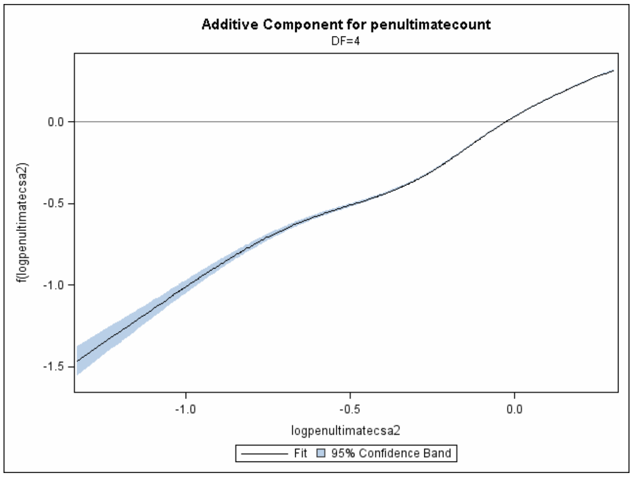


It's possible to modify the formula for "scaling up" the nut count on a random path to represent the entire tree. Very simplistically, this can be done by:

- Assuming a piecewise linear relationship that flattens out between ln(CSA) and ln(nuts) for terminal branches, and
- Assuming a piecewise linear increasing relationship that increases for small values of ln(CSA) and then decreases for values of ln(CSA) greater than a threshold value for interior branches.

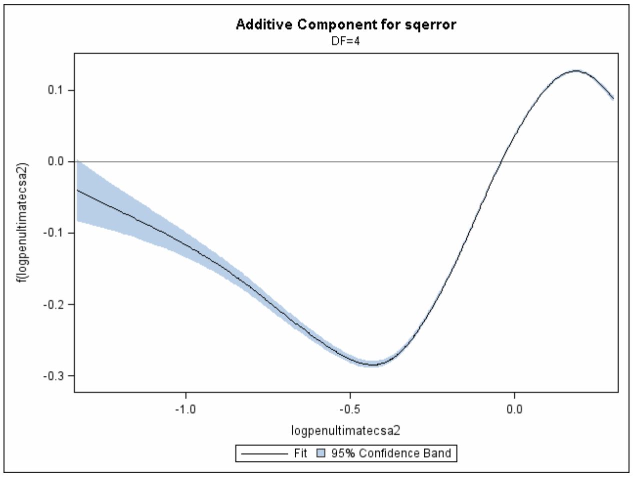
For both of these branch types, it's as if a modified cross sectional area is being used that takes the above piecewise linear form, after which it's assumed that ln(CSA*) and ln(nuts) are linearly related. If generalized additive models are fit using these new definitions, we get the following graphs:

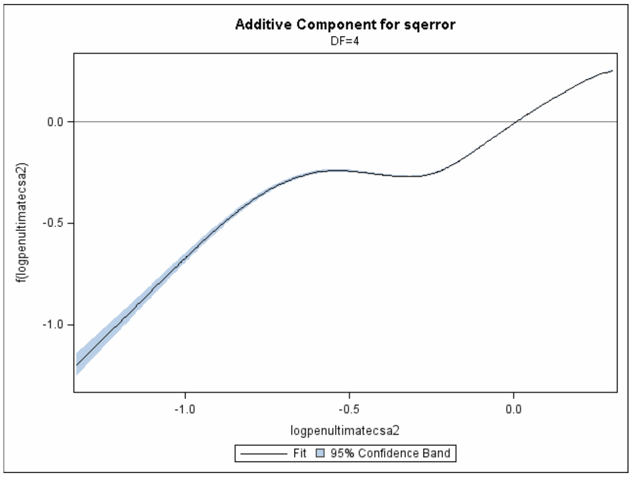




While these relationships aren't perfectly linear, they're much closer than the original graphs, and so they eliminate most of the bias that had been seen with the original method of estimation that USDA-NASS had been using.

Having addressed the bias in the nut count estimates, the next crucial question is how precise those estimated counts are. To look at this question, the squared errors from the previous modified models were fit as a function of the (modified) cross sectional area, using another set of generalized additive models. The results are shows in the following pair of graphs, one (on the left) for the terminal branches, and one (on the right) for interior branches:



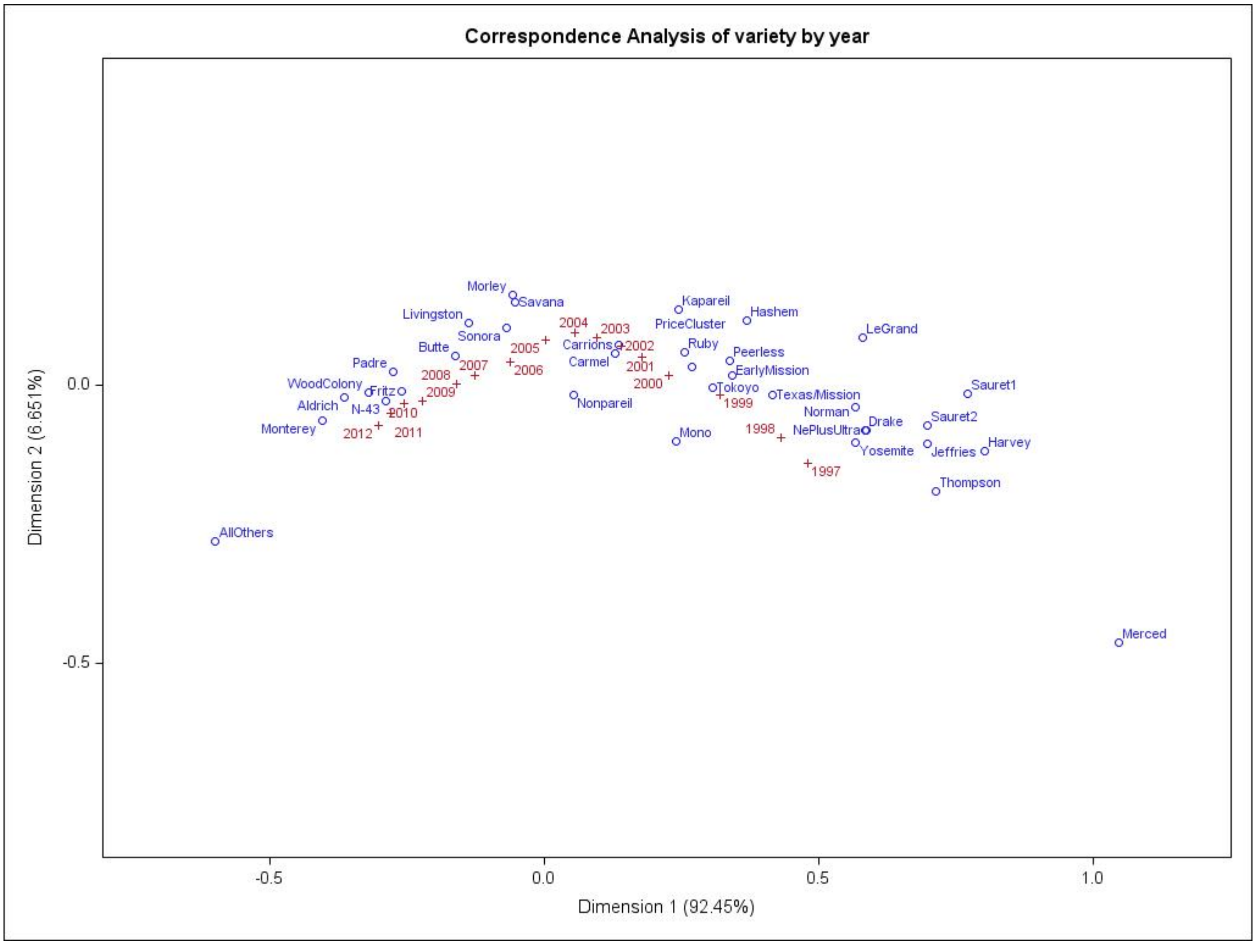


.

In the left-hand picture, we see that the errors in log nut estimates are greater for the smallest and for the largest branches. The large variability for small branches makes sense, since the proportional variability in a Poisson count is greater when the mean is small. The largest branches on this plot correspond to branches with cross sectional area near where the growth in nut counts flattens out, and the interpretation is that branches with this cross sectional area will comprise a mix of ones with continued growth in nut counts, and ones with greatly diminished nut counts. Since for a given tree, you can't be sure which you're observing, branches with diameters in this range should be down-weighted in created an overall nut count estimate.

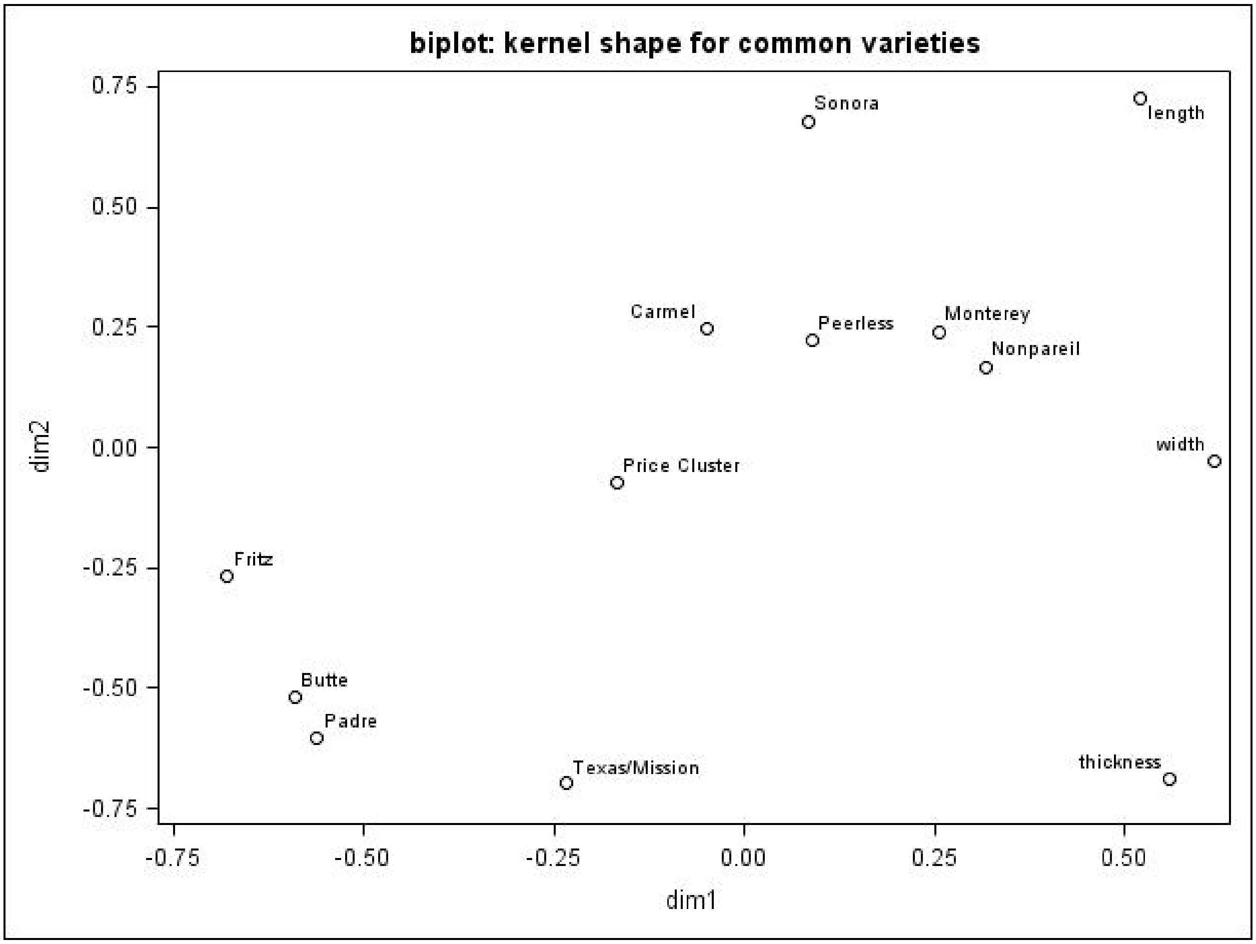
In the right-hand picture above, we see that the proportional variability in nut count increases as a function of branch size, meaning that branches with cross sectional areas close to the point where expected nut counts start decreasing are the least reliable. As before, trees with branches in this size range should be down-weighted when producing the overall estimate of nut production.

Analyses were also run on changes in planting acreage by variety and on nut sizes/dimensions by variety, to see whether Nonpareil characteristics are markedly different from those of other common varieties. A correspondence analysis was run on acreage numbers, and the results are seen in the following graph:



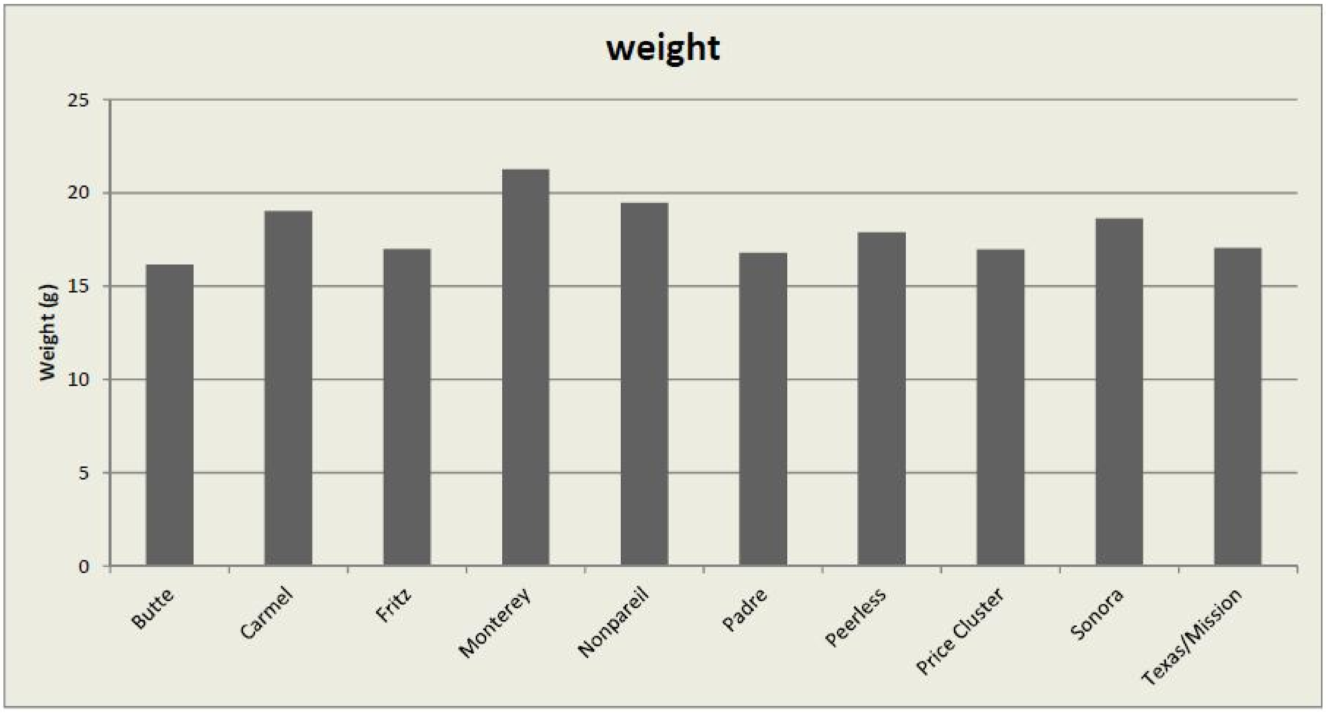
In this plot, the varieties that plot on the left are ones that have had the fastest increases in planting acreage, while the ones on the right have had the slowest increases (or even decreases) in acreage. It's instructive that Nonpareil is vaguely in the middle of this plot, meaning that an overall adjustment to the estimated acreage will serve the Nonpareil estimate will, while it will be less accurate for the varieties on either extreme.

For the nut size data, a multivariate analysis of variance was run on the average length, width and thickness of the nuts for which this data was recorded, and a biplot was created with those characteristics and the varieties plotted on the same axes. The plot was as follows:



Nonpareil plots on the right-hand side of the graph, along with the three size characteristics. Roughly, this means that Nonpareil almonds are larger than the other most common varieties, and so a crop estimate that assumes that the relationship between nut count and total crop weight is constant across varieties will tend to underestimate the total Nonpareil crop weight.

Similarly, an analysis of the available data on average nut *weight* shows that Nonpareil trees produce the second-heaviest nuts (second to Monterey) among the ten most common varieties (as seen in the graph that follows). The data on which this analysis was based was collected along random paths. This protocol is problematical, since a greater proportion of branch segments are counted in the interior of trees than in the crown of trees. This means that interior nuts are overrepresented in the current nut samples, and any systematic differences between interior and crown nut weights would bias the overall estimate of nut weight. NASS has started segregating the sampled nuts from the tree crown from the nuts from the interior, so when that data becomes available, it will be possible to assess the magnitude of any biases due to the method of sampling nuts without respect to location within the tree.



When these results are combined, the methodology for producing nut crop estimates will be improved by:

1. Revising the current assumptions about the relationship between branch CSA and nut counts,
2. Estimating nut counts separately for crown and interior branches,
3. Using a weighted crop estimate that recognizes when the nut count estimates for individual trees are apt to be inaccurate,
4. Adjusting for variety-specific differences in the year to year growth in planting acreage,
5. Recognizing varietal differences in the average weight per nut, and
6. Distinguishing between nuts located in tree crowns and those located in the tree interior in calculating average nut weights.

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 for their statistical content, and that the data on which the analyses are based are confidential and can’t be disseminated.