Evaluation of Using CVs as a Publication Standard

Jason Bell* Wendy Barboza†

Abstract

Each year, the National Agricultural Statistics Service (NASS) conducts a County Agricultural Production Survey (CAPS). The purpose of this survey is to estimate the acreage and production of specific commodities at the county level within each state. To avoid disclosure issues, NASS has two criteria for their publication standards. Either the county must have 30 usable, positive reports, or it must have a certain number of usable, positive reports and the reported data must account for at least 25% of the county-level estimate. In the past, Coefficients of Variation (CVs) were not available because CAPS was not a probability-based sample. This has since changed, thus the CV can now be used as a publication standard.

In order to evaluate this, plots were generated to display the relationship between usable, positive reports and the CVs of the final survey estimate for each county. In general, the plots revealed that there is a negative relationship between the two variables. More importantly, the CV of some publishable counties was unexpectedly high, suggesting that a minimum CV might be an appropriate criterion for determining whether or not a county is published.

Key Words: NASS, publication standards, CV, coefficient of variation, graphics, county estimates

1. Introduction

The National Agricultural Statistics Service (NASS) is an agency under the United States Department of Agriculture (USDA) that provides timely, accurate, and useful statistics in service to U.S. agriculture. NASS conducts many establishment surveys each year to provide data relating to U.S. agricultural products, such as national and state-level crop production and yield estimates.

There is also a great demand for county-level data (i.e. corn production and yield in a given county). To determine these estimates, NASS conducts the County Agricultural Production Survey (CAPS or County Estimates survey) each year. Prior to 2009, CAPS was not sampled using a probability structure. NASS began its implementation of Multivariate Probability Proportional to Size (MPPS) for CAPS by sampling 5 states using MPPS in 2009 and 2010. In 2011, NASS sampled all states using MPPS for CAPS.

With a probability sampling scheme in place, we now have meaningful coefficients of variation (CVs). These allow us to compare a normalized measure of variation across counties and commodities as they relate to both the quality of point estimates and the number of reports from which the point estimates are determined.

2. Motivation

NASS currently has a set of publication standards to determine whether or not we will publish estimates for a county. In order to publish, we need at least

*National Agricultural Statistics Service, jason.bell@nass.usda.gov

†National Agricultural Statistics Service, wendy.barboza@nass.usda.gov

3 reports with positive data (planted, harvested, production) with no individual report accounting for $p\%$ (p is a confidential parameter) or more of the final estimate for that county. This condition must be met or the county cannot be published. Furthermore, we need at least 30 reports with positive data or at least 25% coverage (the harvested acres from reports divided by the current year's harvested acreage estimate) in the county.

We wanted to examine the CVs with these publication standards in mind in order to answer some questions. First, can and should the CVs be used as a publication criterion? Second, once we reach a certain number of reports in a county for a given commodity, should we stop collecting data in that county? This will save resources, and also allow us to focus on other counties and commodities so that we can publish as many counties for as many commodities as possible. Finally, how well are our current publication standards performing?

3. Analysis

To answer these questions, we examined the CVs graphically. We started with the 2009 and 2010 data by looking at the five states that were sampled using MPPS. The commodities examined were Corn, Soybeans, Winter Wheat, and Barley (in 2010 only). We plotted the CVs against the number of usable, positive reports for counties that had at least 3 reports. Each point represents an individual county. The area of the bubbles is proportional to the percent coverage (defined above). Plotting characters 'X' and 'O' represent unpublished and published counties, respectively. A Loess smoother (with 95% confidence bounds) was used to show a trend line.

In 2011, we generated the same plots for all States and all commodities. The national data was broken up by geographic regions (see Figure 1). We looked at the data first at the national level, then at the regional level. At that point, it was clear that we needed to look at a few states individually to further examine counties that were both published and had high CVs. To do this, we plotted the reported yields for each individual record against the county where the area of the bubbles is proportional to the number of harvested acres.

Figure 2 shows the national data for soybeans. The plot on the left defines the

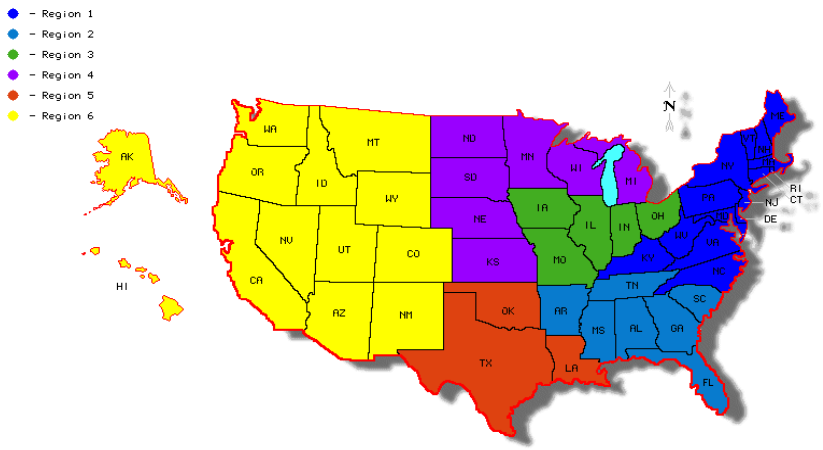

Figure 1: Regions

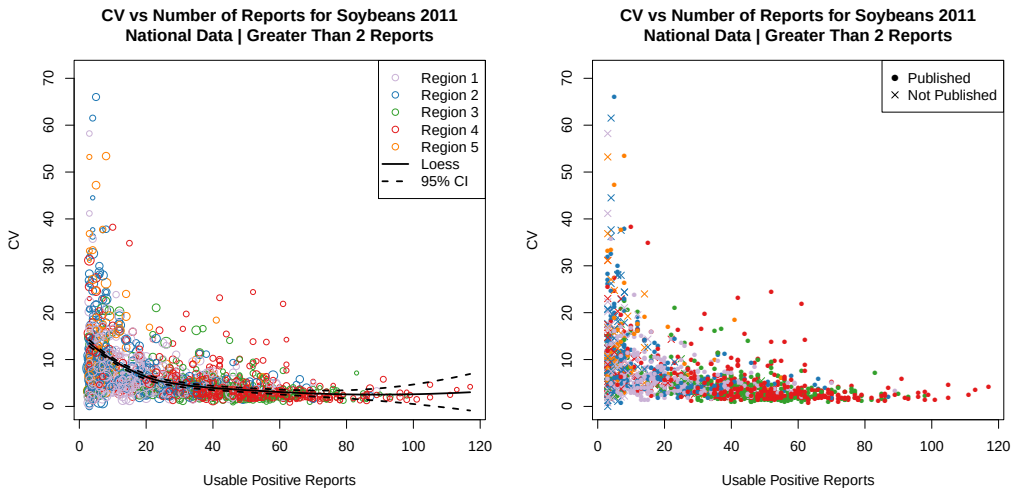

Figure 2: National Data for Soybeans

regions by color and shows a Loess smoother with 95% confidence bounds. The plot on the right shows published vs. unpublished counties. A quick look reveals the overall relationship between the CVs and the number of usable, positive reports, which is to say a strictly decreasing relationship. Upon closer examination, we see that the line appears to stabilize around 40 reports. At this point, we might wonder if it is really necessary to obtain 120 reports, or even 60+ reports in order to publish a high-quality point estimate. On the other side of things, we notice a few published counties that had very high CVs, which is cause for concern and needs to be further examined.

Figure 3 shows the national data for winter wheat, and gives us another example of what we saw in Figure 2, although the published counties with high CVs aren't quite as high. We see, again, that the Loess line seems to taper off at around 40 reports, yet there's a county wherein we collected around 140 usable, positive reports. We see the same thing in Figure 4, the national data for sorghum, only this time there's a published county with a CV of over 100%. We saw further evidence of published counties with high CVs and relatively large numbers of usable, positive reports in the remaining commodities, although they are not discussed here.

We generated Figure 5 to show the total counts of published counties with CVs over 20, 30, and 40% for all commodities. The height of the bar represents the total count for each commodity, and the colors represent the number of published counties in each category. For example, there were just above 70 published counties for corn across the nation that had large CVs. Of those, just under 50 were between 20% and 29%, about 20 were between 30% and 39%, and of about 5 counties were published with CVs 40% and above.

We now revisit Sorghum in Texas. Figure 6 shows data from Texas only, as opposed to the national data. Letters A through I label published counties with CVs higher than 25%. The plot on the right shows the individual reports for those counties. County E had 3 reports. One large operation reported a yield near 0 bushels per acre, and two smaller operations reported yields around 60 bushels per acre. This is causing the CV to soar above 100%. The question is, should we have published this county? Did this operation truly have a yield *that* small? The current estimate published in county E was 9.4 bushels per acre.

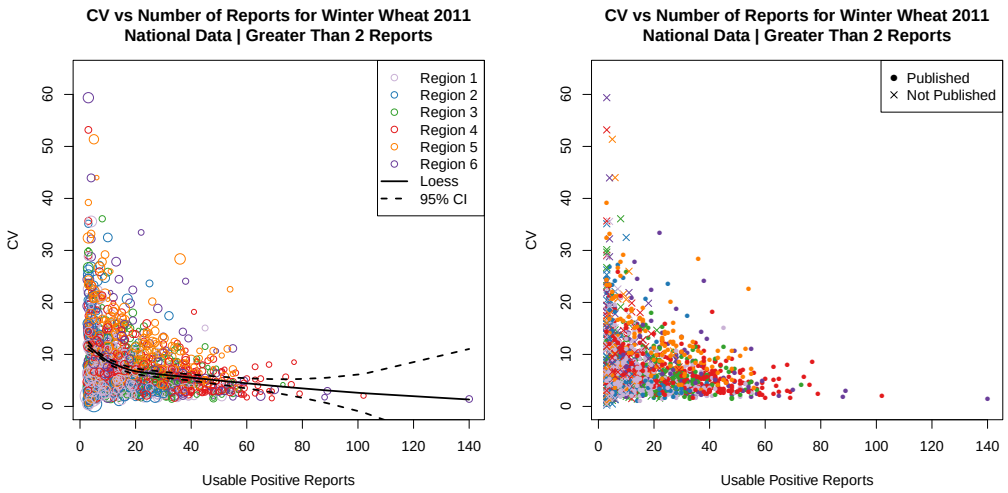

Figure 3: National Data for Winter Wheat

Another big issue that revealed itself in these plots is that of irrigated vs. non-irrigated. In some states, we ask whether or not a commodity was on irrigated or non-irrigated land. We see a distinct separation within the counties (Figure 6). This is causing higher CVs in those counties, yet the point estimates *should* be published.

We see in Figure 8 another cause to high CVs: spread in data and outliers. County E has 4 reporters with similarly sized farms. Three of the reporters have similar yields and the fourth is an outlier. The number we published in County E was 63.5 bushels per acre. Should we have published this county? Since the outlier is causing the CV to jump up around 25% and clearly influenced the point estimate to be artificially high.

4. Conclusion

We can now begin to answer the questions we asked going into the research project. First, should the CVs be used as a publication standard? We have 2 publication criteria in place that we follow religiously, but are they enough? The CVs provide us with additional information about the quality of the point estimate. Although variation is naturally going to decrease with additional usable, positive reports, the CVs give us some comparison within counties and between commodities, and should definitely be looked at as an additional criterion for publication. For example, a county *should not* be published with a CV above $x\%$. We may also want to consider publishing all counties meeting the 2 publication standards already in place, but give a disclaimer for published counties with CVs over, say, 20% and caution those using the data to do so at their own risk.

The other question we wanted to answer is a little trickier. Can and should we discontinue collecting data in a county once a threshold number of reports has been achieved? The short answer is yes. If we are concerned with saving resources (which we are), we should stop calling in a county for a given commodity when we can assure the quality of the point estimate. However, there are many reasons why we may not be able to do this. When focusing on minor commodities, such as sunflowers, we will get reports for the major commodities, such as corn, for free.

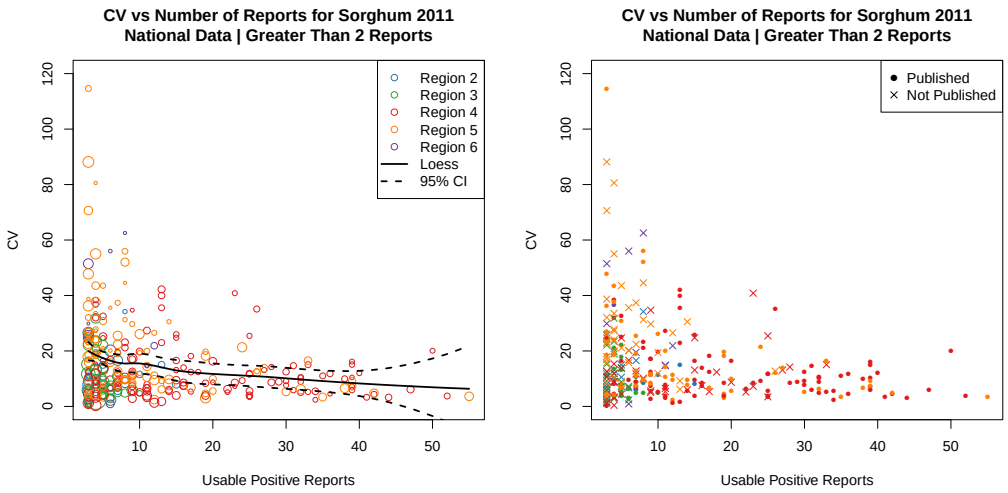

Figure 4: National Data for Sorghum

Furthermore, constructing the logic to "assure quality" of point estimates can be a very tricky problem. A NASS team is looking further into this, and with the addition of meaningful CVs will hopefully be able to arrive at a solution.

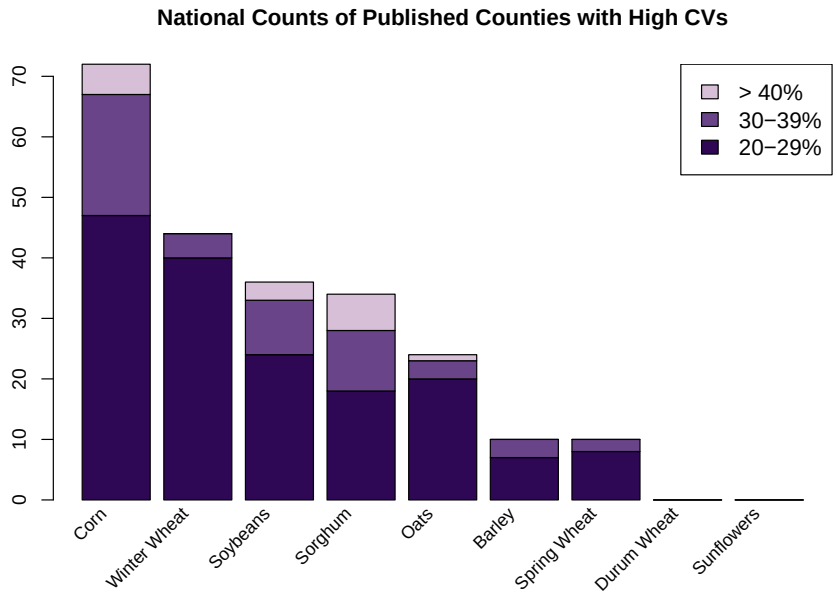

Figure 5: National counts of published counties with CVs above 20%.

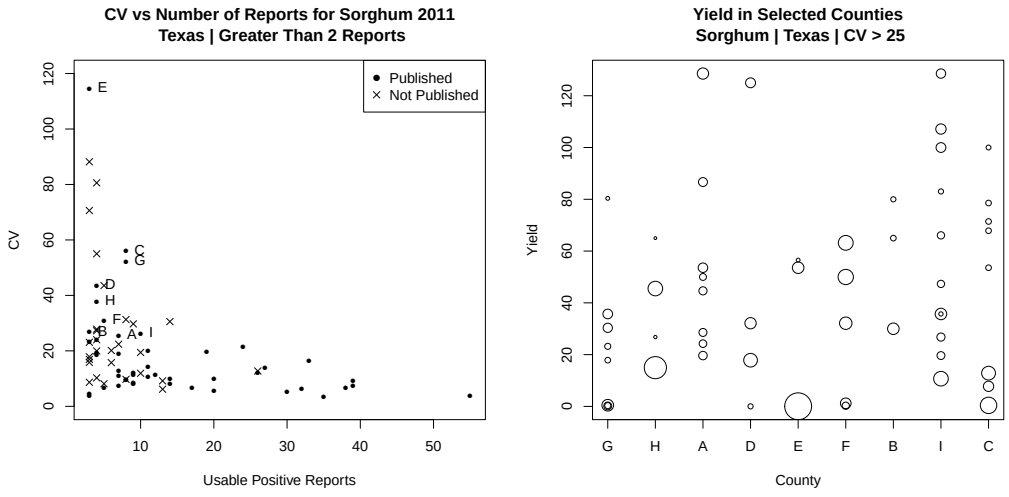

Figure 6: Example of Large Outlier and Few Reports — Sorghum in Texas

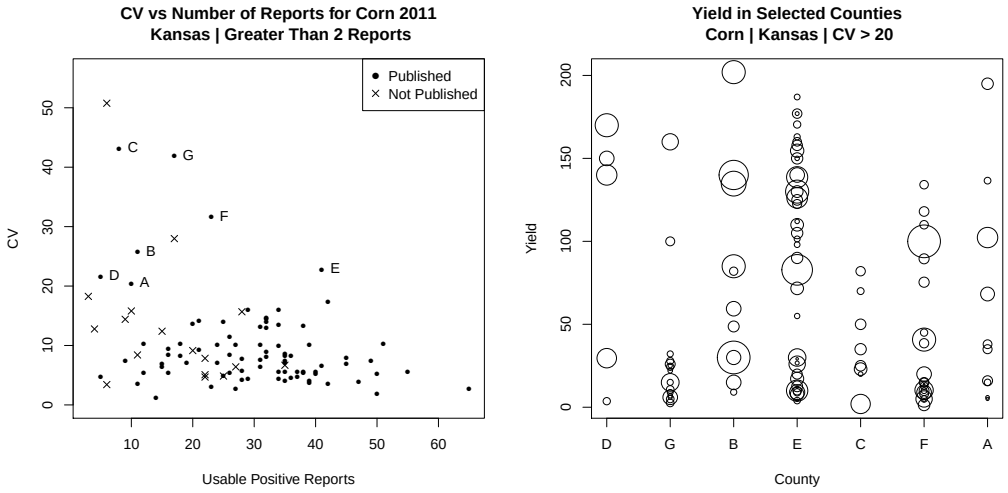

Figure 7: Example of Irrigated vs. Non-Irrigated — Corn in Kansas

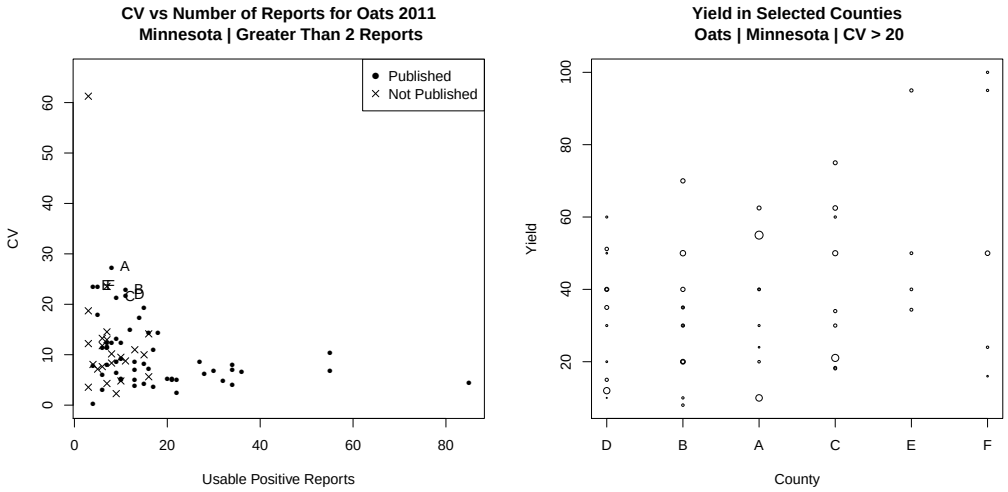

Figure 8: Example of Spread in Data and Outlier — Oats in Minnesota