United States
Department of
Agriculture

National
Agricultural
Statistics
Service

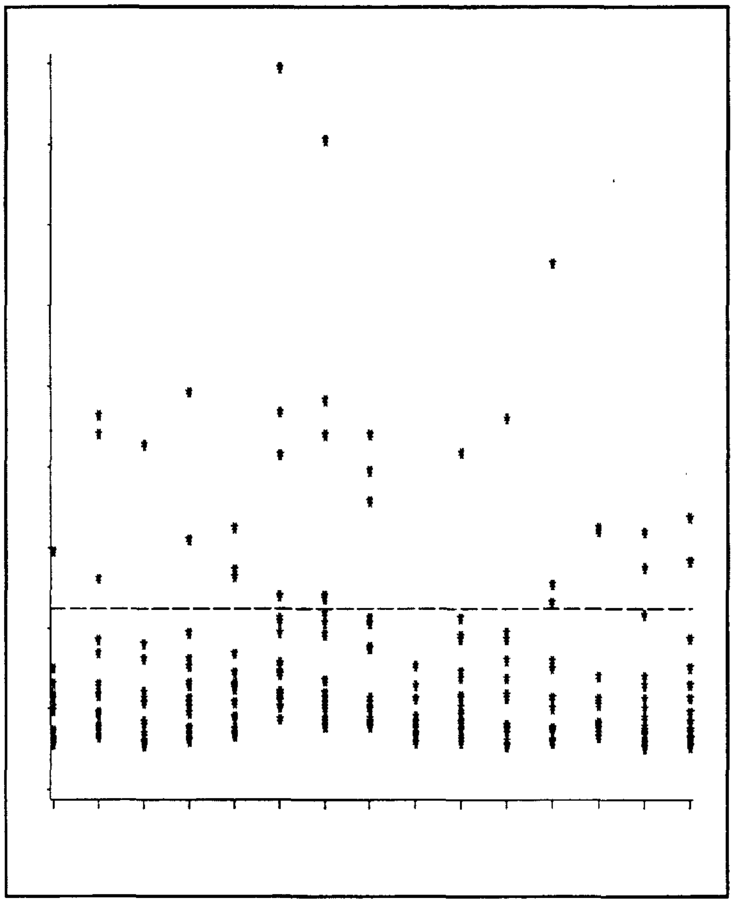
Research Division

SRB Research Report
Number SRB-95-02

July 1995

An Evaluation of Truncation Estimators for Improving State Estimates of Total Hogs

Susan Hicks
Matt Fetter
Susan Cowles



AN EVALUATION OF TRUNCATION ESTIMATORS FOR IMPROVING STATE ESTIMATES OF TOTAL HOGS, by Susan Hicks, Matt Fetter, and Susan Cowles, Sampling and Estimation Research Section, Survey Research Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D.C. 20250-2000, April 1995, Report No. SRB-95-02.

ABSTRACT

The National Agricultural Statistics Service (NASS) consistently strives to produce more reliable estimates of hogs at both the State and National levels through its Quarterly Agricultural Surveys (QAS). One contributor to high variances at the State level are outliers in the survey data. Outliers are defined as extremely large expanded observations. Outliers can be caused by large weights, large reported values of hogs, or a combination of both. In this report, we evaluate the efficiency of a class of estimators, called weight truncation estimators, for improving the reliability of the State level estimates of total hogs in the presence of outliers. The estimators truncate the weights of the large expanded observations so that the expanded value does not exceed a predetermined cutoff. Two types of truncation estimators are evaluated: data-driven estimators and fixed cutoff estimators. The fixed cutoff estimators hold the most promise for improving the State level estimates both in terms of ease of application and efficiency of the estimates.

KEY WORDS

Robust Estimation; Truncation-type Estimators; Winsorization; Minimum Estimated MSE Trimming; Iterative Fence Method.

This paper was prepared for limited distribution to the research community outside the U.S. Department of Agriculture. The views expressed herein are not necessarily those of the USDA or NASS.

ACKNOWLEDGEMENTS

The authors would like to thank: Don McKinney for programming the fixed cutoff estimator in the operational summary system under tight time constraints; Each of the 16 State offices that took the time and effort to provide suggested cutoffs for their States; And special thanks to Bill Iwig for reviewing the report and the cutoffs.

TABLE OF CONTENTS

SUMMARY

The National Agricultural Statistics Service (NASS) produces quarterly estimates of total hogs at the State and national level through its Quarterly Agricultural Surveys (QAS). Outliers are a recurring problem in agricultural surveys, especially livestock surveys, and can severely distort the estimates. The outlier problem can be approached as a sample design problem, an estimation problem, a nonsampling error problem, or some combination of these. In this report, we evaluate the efficiency of a class of estimators, called weight truncation estimators, for improving the efficiency of State level indications of total hogs. Two types of weight truncation estimators are evaluated: 1) data driven estimators that determine the optimal weight truncation value based on the sample data, and 2) fixed cutoff estimators that truncate the expanded values to a pre-determined cutoff, independent of the sample data.

The estimators were evaluated two ways: 1) a monte carlo simulation was developed to mimic the sample design of the non-overlap domain of the QAS, and 2) the most promising estimators were evaluated against final Agricultural Statistics Board numbers for five key hog producing States. The monte carlo simulation indicated that fixed cutoff estimators which used pre-determined truncation points were superior to data driven estimators in terms of MSE. These robust estimators produced relative efficiencies as high as 1.27. When evaluated against Board estimates for the five hog States, the fixed cutoff estimator generally outperformed the data-driven estimators whenever several large units occurred in the data.

Based on our analysis we recommend that the fixed cutoff approach be used to correct for outliers in the estimation stage. Additional gains could be realized by improvements in the sample design for estimating hogs or exploring methods for reducing nonsampling errors in the data collection.

AN EVALUATION OF TRUNCATION ESTIMATORS FOR IMPROVING STATE ESTIMATES OF TOTAL HOGS

INTRODUCTION

The National Agricultural Statistics Service (NASS) uses a multiple frame survey procedure for estimating hog inventories at the State and national levels. Estimates are published on a quarterly basis using data that are collected through the Quarterly Agricultural Survey (QAS) program. The QAS consists of two independent frames: a list sampling frame (LSF) and an area frame (AF). The LSF is a listing of known farm operations, while the AF is composed of segments of land. All land in the contiguous 48 States has a positive probability of selection in the area frame. Thus the AF is a complete frame and is used to measure undercoverage in the list frame. Sample units in the AF that are not on the LSF comprise the non-overlap (NOL) domain of the AF. The multiple frame estimate is formed by combining the estimate produced from the LSF sample with the estimate of the NOL domain derived from the AF sample. The proportion of the national hog inventory estimate that comes from the NOL has been about 12 percent since June 1993.

While the LSF is fairly efficient for hog inventory estimation, the AF is not. NOL hog operations can sometimes expand to such large values that they increase the multiple frame estimate beyond reasonable levels. These observations can be called "outliers." Although outliers do occur in the LSF sample they are primarily a phenomenon of the NOL. Because of their adverse effect on the estimates, some statistically sound procedure must be in place that treats them after the fact, such as robust estimation.

We will define a robust estimator as an estimator that is resistant to outliers. All of the robust estimators considered in this research are expansion factor-truncation type estimators. In simulation studies, many of these estimators were found to be more efficient than the unbiased alternatives. An evaluation of the performance of the most promising estimators was also made using real data.

THE OUTLIER PROBLEM

The outlier problem can be approached as a design problem, a nonsampling error problem, and an estimation problem. The focus of this research has been to approach the problem from the direction of estimation. Robust estimators treat outliers and influential observations after the fact rather than reducing their frequency or magnitude through design changes or by reducing nonsampling errors. Although outliers can and do occur at the national level, the focus of this research is to recommend a procedure to adjust for outliers at the State level.

<u>Outliers in the NOL Domain</u> -- The area frame uses a stratification scheme that is based primarily on cultivation intensity. Stratification based on cultivation intensity is quite efficient for estimating major crop acreage but is rather inefficient for estimating

hog inventories. This is due to the fact that in most States, hogs are a rare commodity and hog production is not highly correlated with cultivation intensity. In general, the NOL domain has larger sampling fractions than the list frame primarily due to cost considerations.

When large sampling weights are combined with an inefficient frame and a highly skewed rare population, extremely large expanded observations can occur. Sometimes these large expanded values can drive the estimate to unrealistically high levels. These exceptionally large observations are called <u>outliers</u>, because they lie far outside the bulk of the data. Outliers cause problems at both State and national levels. In addition to outliers, a more common type of observation is an influential observation. <u>Influential observations</u> are large but singly do not control the estimate.

At the time this report was prepared the NOL follow-on samples were selected as a subset of the NOL June sample. Under the old design, forty percent of the June sample was allocated for other surveys leaving only 60 percent of the original June sample to estimate NOL hog inventories for the follow-on surveys. For the follow-on surveys, NOL units were restratified based on what they reported in June. The restrata were defined to meet multiple needs and thus were not always optimal for hog estimation. Further subsampling was done in some of the restrata. This resulted in some rather large sampling weights. NOL sampling weights could range from around 30 to 1000. The basic form of the NOL expanded number of hogs for operation i was:

$$\text{EXPANDED HOGS}(i) = \text{EXPANSION FACTOR}(i) * \text{TOTAL HOGS}(i) \qquad (1)$$

where:

$\text{EXPANSION FACTOR}(i) = \text{SAMPLING WEIGHT}(i) * \text{TRACT ACRES}(i) / \text{TOTAL ACRES}(i)$

TRACT ACRES(i) = the number of acres sampled for operation i.

TOTAL ACRES(i) = the total number of acres for operation i.

SAMPLING WEIGHT(i) = inverse of the probability of selection for operation i.

In September 1994, the sampling scheme for the NOL portion of the QAS was changed. All NOL units which were found to be positive for hogs in June were interviewed in the follow-on samples. This should reduce the number and size of outliers that occur in the follow-on samples to mirror the June sample. The revised sampling scheme is not reflected in this report. The results still apply, although greater emphasis should be placed on results for June since the June survey will more closely resemble the new sample design of the QAS.

There are two important nonsampling errors that influence the number of outliers.

1) The incorrect determination of overlap status and

2) incorrect tract to farm ratio weights.

The errors in overlap status occur when a hog operation identified in the AF sample is incorrectly identified as an NOL operation. When treated as an NOL operation, its expansion factor can be substantially larger than its correct LSF expansion factor. If the operation has a large number of hogs, an outlier may be created. The second error involves the ratio used to allocate the total number of hogs associated with the operation to the proportion of the total farm acres actually sampled -- the tract to farm ratio. Because this ratio is less than or equal to one, it will reduce the effect of large expansion factors. An incorrect ratio can easily cause the expanded number of hogs for an operation to be many times greater than what it should be. Both of these nonsampling errors create expansion factors that are larger than they should be.

Note also that the kind of estimators we are evaluating do not discriminate between "true" outliers caused by finding a rare large observation in the NOL domain and outliers that result from nonsampling errors, because *the effect on the data is the same*. However, truncation estimators should not be viewed as a substitute for aggressively tracking and erradicating sources of nonsampling errors.

Outliers From The List Frame -- Due to the design of the QAS, most outliers come from the NOL domain. List frame expansion factors are typically less than 100. Because of this, the expansion factor plays a smaller role in the creation of outliers for list frame records than it does for NOL records. Large expanded values in the LSF are usually due to either a valid extreme operator with an expansion factor of 1.0, or a large operator classified into a stratum with a large LSF sampling fraction. Large extreme operators from self-representing strata are not considered outliers.

In this research, the emphasis will be on dealing with outliers in the NOL domain, although the same methods that are used for NOL outliers will be applied to the list frame outliers when the final estimator is adopted.

THEORETICAL BACKGROUND FOR TRUNCATION-TYPE ESTIMATORS

For estimating the mean of a population, Searls (1966) proved that under quite general conditions, there will always exist some cutoff point such that when reported values are truncated to this cutoff the resulting estimator will have minimum MSE over unbiased estimators. Further, Searls showed that gains are achieved for a wide range of cutoffs when the data originate from an exponential distribution. Ernst (1980) compared seven estimators of the sample mean that adjust for large observations. Four of the estimators were modifications of truncation estimators at a cutoff. The other three estimators were modifications of estimators which truncate a fixed number of observations. Ernst showed that for the optimal cutoff, the estimator which substitutes the cutoff for the sample values greater than the cutoff has minimum MSE.

These papers give theoretical results and provide insight into the performance of truncation estimators under simple random sampling, but do not provide information on how these estimators or other types of robust estimators would perform using more complex sampling designs. Work by Thomas, Perry, and Viroonsri (1990) applied empirical Bayes techniques where the data was truncated in many different ways. One conclusion of their work was that the best way to truncate the data was by fixed cutoffs.

The results from their research were further modified to create the robust estimator described by Perry and Keough (1991) which is applied at the national level. From the national indication, an outlier component is computed by quarter based on records that expand above a certain fixed cutoff on a State by State basis. This outlier component is then averaged over several years by quarter and the current quarter's outlier component is replaced with the average outlier component. This estimator is not used to produce indications at the State level. Due to the sparseness of the NOL domain at the State level, an estimate of the average outlier component would be less stable than at the national level.

GENERAL FRAMEWORK FOR TRUNCATION-TYPE ESTIMATORS

All of the robust estimators considered in this report are expansion factor truncation-type estimators. The expansion factor for the NOL domain is defined in (1). For LSF operations, the expansion factor is more complicated due to the nonresponse adjustment. It will not be explicitly stated here. Interested readers should refer to Fetter (1992).

Outliers will be defined to be records whose expanded values exceed some cutoff, T. All of these estimators truncate the expanded value back to T by reducing the size of the expansion factors for records that exceed the cutoff. The portion of the weight that is truncated is then "smoothed" over all records within that domain. Thus, these expansion factor truncation estimators are all downwardly biased. A general expression for all robust estimators in this study will now be presented.

Define:

w_i = untruncated expansion factor

$\hat{Y}$ = $\Sigma_i w_i y_i$, the unbiased estimate of the population total

$$w_i' = w_i \quad \text{if } w_i y_i \leq T$$

$$= T/y_i \quad \text{if } w_i y_i > T$$

Note that $w_i' \leq w_i \ \forall \ i$

Define the truncation adjustment for domain D as:

$$A_D = \Sigma_D w_i / \Sigma_D w_i' \qquad (2)$$

The general robust estimator can now be expressed as:

$$\hat{Y}_t = \Sigma_D A_D \Sigma_i w'_{Di} y_{Di}$$

The domain is defined as the final nonresponse adjustment cell for list frame records. All NOL records in a State collectively define the single NOL domain. For list frame records the truncation adjustment is applied to all records within the post-stratum used for nonresponse adjustment (Fetter 1992) from which the outlier record came. For the NOL domain, the truncation adjustment is applied to all NOL records.

The problem that remains is how the cutoff value, T, should be determined. There are many different methods that can be used to determine this cutoff. Nearly all of these methods can be put into either one of two broad categories: fixed cutoff methods and data driven methods.

FIXED CUTOFF ESTIMATORS

Fixed cutoff estimators, also referred to as winsorization at a fixed cutoff, define the level which separates the outliers from the non-outliers prior to data collection. This method may produce any number of outliers on any given survey, including zero. Theoretically, if we knew the distribution of the sample universe from which the data were generated, we could derive the optimal cutoff. From the simulation study, which was based on a fixed known universe, we were able to derive the optimal cutoff for that universe. However, in practical applications the population distribution is not known. Thus, historic data and expert judgement are the best tools for defining the optimal fixed cutoff.

One advantage of the fixed cutoff estimators is that they seem to be more effective at identifying true outliers. Because the cutoff is determined before the sample is selected, they are also completely robust against samples that are not typical of the population (bad samples). However, determining the optimal cutoff for future samples often involves a combination of guesswork and expert judgement and the performance of the cutoff must be reviewed periodically.

We evaluated the fixed cutoff estimators two ways. First, based on the simulation we evaluated the efficiency of a range of different cutoffs around the optimal cutoff. Second, we evaluated the fixed cutoff estimator compared to the Agricultural Statistics Board on real survey data for five key States. The cutoffs applied to real survey data were determined by using a combination of historical data review and expert judgement. This is discussed more thoroughly in a later section.

DATA DRIVEN ESTIMATORS

Data driven methods refer to methods where the cutoff value is not known until the survey data have been collected. Thus, the cutoff value is a function of the observed data. One advantage of the data dependent methods is the objectivity used to determine the cutoff; an algorithm is used instead of expert judgement. The disadvantage is that the cutoff must be rederived with each new sample. All data dependent procedures rely on the assumption that the sample distribution reflects the population distribution within the sampling domain (i.e. stratum, cluster, etc.). In other words, these methods are not robust against bad samples. Our studies have shown that data dependent cutoffs fluctuate considerably from sample to sample. This occasionally results in cutoffs that are unrealistically low.

There are many different types of functions which can be used. Most of these are either a function of order statistics such as once winsorized, iterative fence method, or of sample moments such as Minimum Estimated MSE. These methods will be briefly described below.

Winsorization at an Order Statistic -- Winsorization at r order statistics will be defined as using the r+1st largest expanded value in the entire sample (list and NOL) as the cutoff value. Thus for the once-winsorized estimator, the cutoff value is defined to be the second largest expanded value appearing in the sample. The largest observation is then truncated back to the second largest observation. Unlike the fixed cutoff estimator, winsorization at r order statistics results in a variable truncation point but truncates a fixed number of observations. In the June simulation study, the performance of winsorization at one, two three and four order statistics were evaluated. The once-winsorized estimator gave the best performance of the winsorized estimators over all simulations. The once-winsorized estimator was also evaluated on the follow-on simulation and on historical survey data.

<u>**Minimum Estimated MSE Trimming**</u> -- Minimum Estimated MSE Trimming (MEMSE) is a variation of winsorization. The idea behind the MEMSE procedure is to determine the optimal trimming level or the optimal number to trim based on an estimate of the MSE. It is thus an iterative approach, that is highly dependent upon how representative the sample data is to the sample universe. In Hicks and Fetter (1993), we evaluated MEMSE as a tool to determine the optimal trimming <u>level</u>. The results were not promising. In this application, we evaluated MEMSE as a tool to determine the optimal <u>number</u> of observations to trim.

The data was winsorized at one, two and three order statistics using the procedure described in the previous section. At each iteration, an estimate of the MSE was computed. The order statistic that minimized the estimated MSE was then used as the cutoff value. If the estimated MSE was minimized with no truncation, then the data were not truncated.

The estimator of the MSE that was used in this procedure is derived from the following relation:

$$E(\hat{Y}_t - \hat{Y})^2 = Var(\hat{Y}_t) + Var(\hat{Y}) - 2\ Cov(\hat{Y}_t,\hat{Y}) + [E(\hat{Y}_t) - E(\hat{Y})]^2 \tag{3}$$

$$MSE(\hat{Y}_t) + Var(\hat{Y}) - 2Cov(\hat{Y}_t,\hat{Y} \tag{4}$$

$$MSE(\hat{Y}_t) = E(\hat{Y}_t - \hat{Y})^2 - Var(\hat{Y}) + 2Cov(\hat{Y}_t,\hat{Y}) \tag{5}$$

where:

$\hat{Y}_t$ = *the truncated estimate*

$\hat{Y}$ = *the untruncated unbiased estimate*

If we assume that the correlation between the truncated and untruncated estimate is equal to 1.0 then the MSE($\hat{Y}_t$) can be expressed as:

$$MSE(\hat{Y}_t) = E(\hat{Y}_t - \hat{Y})^2 - Var(\hat{Y}) + 2[Var(\hat{Y}_t)Var(\hat{Y})]^{(1/2)} \tag{6}$$

This assumption is useful because of the difficulties in estimating the covariance between the truncated and untruncated estimator from one sample. It is a conservative assumption in the sense that it gives the upper bound for the MSE($\hat{Y}_t$). Though unconditionally unbiased estimators of the first two terms in (6) are fairly straight forward, estimating the third term presents some difficulty. The method of estimating the third term that was employed in the simulation study was to compute the estimate of the variance of the truncated estimator as if the truncated expansion factors were the untruncated expansion factors.

<u>**The Iterative Fence Method**</u> -- This method is based on Tukey's inner and outer fence definitions. Tukey defines the <u>inner fence</u> as:

$1.5 * (Q_{75} - Q_{25}) + Q_{75}$, where Q_{75} is the 75th percentile and Q_{25} is the 25th percentile over the entire sample (list and NOL). The outer fence is defined as: $3.0 * (Q_{75} - Q_{25}) + Q_{75}$. Tukey defines "far out" values as those falling beyond the outer fence. These values might be construed as being outliers.

The iterative fence method is used by Statistics Canada and was first applied to QAS data by Flores-Cervantes (1993). It is a procedure that iteratively derives fences that are modifications of the fences defined by Tukey. In this study, these fences are computed using Tukey's definitions but are subject to the restriction that no more than seven fully expanded observations fall outside either the inner or outer fence. If more then seven expanded observations fall outside either fence, that fence is recomputed, applying Tukey's definitions to those values that fall outside the fence being computed. Once the fences have been determined, the cutoff level is determined to be the midpoint between the two fences. All expanded observations falling outside this cutoff are defined as outliers and are subjected to expansion factor truncation. This method may produce any number of outliers from zero to seven.

EVALUATING THE EFFICIENCY OF THE ESTIMATORS WITH A MONTE CARLO SIMULATION

Although applying these estimators to survey data sheds some light on the performance of the robust estimators across time compared to the Agricultural Statistics Board estimates, it gives us little indication of the relative performance of these estimators in terms of their true mean square error (MSE) for a given population. To investigate the efficiencies of the robust estimators under consideration we developed a simulation study. Because of the complexities of the multiple frame design and because the major source of outliers and sampling variability is from the NOL, we restricted the simulation to the NOL domain of the area frame. Iowa and Georgia were chosen for the simulation. Georgia was chosen because it had many potential outliers during the evaluation period. Iowa was chosen because it is the largest producer of hogs. The results for Georgia were presented in Hicks and Fetter (1993). This paper presents the simulation results for Iowa.

Modeling the NOL Domain of the QAS -- Each State has its own Area Sampling Frame composed of segments of land. Each segment is defined by artificial or natural boundaries that are easily identifiable at ground level. The segments are stratified based on the intensity of cultivation within the segment. Each segment is further subdivided into parcels of land based on operating arrangement. These parcels of land are called tracts and represent the reporting unit for the QAS.

To form a base for each model, positive weighted NOL tract data for a series of June base surveys were pooled together. This was done in order to gain as much information about the distribution of the NOL hog tracts as possible. For the earliest June QAS data set, all positive NOL tracts were used. For subsequent June data sets, only NOL hog tracts from incoming replicates were used. This was done to eliminate multiple representation of the same tract that might occur if a tract was in the NOL domain for more than one survey year. The tract level weight is a product of the stratum sampling weight and a tract adjustment factor. The adjustment factor prorates an operation's reported value back to the tract level for operations that only partially reside within the sample segment.

Although the tract is the reporting unit, the segment is the sample unit. To deal with this problem, segments with multiple NOL hog tracts were summed to the segment level. It was then possible to develop a model based on segment level data.

A separate parametric distribution was fit to the positive segment level data for each land use stratum that contained hogs. For Iowa three strata were used based on percent of cultivated land: >75% cultivated, 25% - 75% cultivated, and <25% cultivated. These were the 1300, 2000, and 4000 strata. The distributions were all Gamma density functions. Due to the sparseness of the data it was difficult to validate the models. However, our main interest was in developing a reasonable, highly skewed distribution rather than developing highly accurate models of the NOL sample domain. The number of segments having no NOL hogs were modelled by estimating the probability that a sample segment would be positive for hogs based on the number of positive segments to total segments from the QAS data sets. This was done by strata.

From the parametric distributions, we generated a fixed population of segments such that the populations had the same number of segments as the actual populations for that State. With the zero segments included, the result is a highly skewed population with a large spike at zero and a long right tail. From the fixed universe, we drew 1000 stratified simple random samples with replacement of size 420 to mimic the June sample design.The follow-on survey samples were selected from the June sample by modeling the two-phase sample design. First, the June sample segments were delineated into tracts. Then the sample tracts were assigned to second-phase strata using a probability mechanism and subsampling within the second-phase strata. Each sample was "treated" by applying each of the estimators under review. The true MSE of each estimator was estimated over all 1000 samples for both the June simulation and the follow-on simulation. The efficiency of the estimators was estimated as the ratio of the MSE of the unbiased estimator to the MSE of the new estimator.

Simulation Results -- The simulation studies show that fixed cutoff estimators will outperform (in terms of reduction in true MSE) data driven estimators even when the cutoff chosen is far from optimal (on the high side of optimality). See Table 1. These results support Searle's findings that improvements can be made over the unbiased estimator for a wide range of values for T. The simulation also showed that the optimal cutoff is larger for the follow-on samples than the June sample. This is not particularly surprising, since the expansion factors are larger for the follow-on sample than the June sample.

The data-driven estimators were barely more efficient than the unbiased estimator. The MEMSE estimator, in particular, was strongly biased downwards when winsorizing to an order statistic. Recall that the MEMSE estimator minimizes the estimated MSE not the true MSE. The estimated MSE, while unbiased over a large number of samples, is highly variable from sample to sample and

Table 1. **Simulation Results for Iowa**

June Survey		Follow-on Survey	
Estimator	**MSE Ratio**	**Estimator**	**MSE Ratio**
Iterative Fence	1.00	Once-winsorized	1.06
Minimum Estimate MSE	.99	Iterative Fence	1.03
Winsorization at R order statistics		Winsorization at a fixed cutoff	
One	1.07	125,000	.86
Two	1.00	180,000	1.22
Three	.98	230,000	1.27
Four	.96	255,000	1.26
		270,000	1.25
Winsorization at a fixed cutoff		350,000	1.19
100,000	.90		
120,000	1.08		
140,000	1.14		
160,000	1.13		
180,000	1.10		
200,000	1.08		

for different winsorization order statistics within a sample. The highly data dependent nature in which this estimator is derived can cause it to truncate many observations in periods with few outliers and few observations in periods with many outliers. Of course, in some samples it truncates the optimum number, but in general the number and amount truncated is highly unpredictable.

APPLYING THE BEST ESTIMATORS TO HISTORICAL SURVEY DATA

Three robust estimators investigated in the simulation were chosen to be applied to QAS data sets for Georgia, Illinois, Indiana, Iowa, and North Carolina. These estimators were chosen based on their performance in the simulation study and/or their familiarity to NASS personnel. The fixed cutoff method, the once-winsorized method, and the iterative fence method were chosen. For a simple example of the calculation of the fixed cutoff estimator see Appendix A.

Determining the Optimal Cutoff Based on Historical Data -- To evaluate the performance of the fixed cutoff estimator on historical data we first had to determine the optimal cutoff for each of the States evaluated. Based on the results of the simulation, we looked for cutoffs that truncated on average one to three records in a State per survey. In general, we placed more emphasis on the follow-on survey than the June survey. This followed directly from the simulation results which showed that cutoffs greater than the optimal were preferable to lower cutoffs and the optimal cutoff for the follow-on survey is greater than the optimal cutoff for the June survey. Thus by optimizing the cutoff for the follow-on survey we provide a conservative cutoff for the June survey. Realizing that States vary in the occurrence of outliers and the effect of outliers on the data, we enlisted the help of the five State offices in setting the optimal cutoff for their States. Appendix B contains the memo we sent to the States requesting input on the optimal cutoffs for their State. We reviewed the recommended cutoffs against historical data before setting a final cutoff for each State. The optimal cutoffs for hog inventory for each State are:

Iowa	225,000
Illinois	125,000
Indiana	80,000
North Carolina	50,000
Georgia	25,000

The QAS Results -- Because all of these robust estimators concede some bias in order to reduce variance, we must take care not to concede too much bias. Thus, a good robust estimator will be close to the untruncated estimator **most of the time**. (The untruncated estimator refers to the estimator described by Fetter(1992), commonly called the Revised Estimator by NASS statisticians.) However, when unusually large observations occur in the sample, the robust estimator should reduce the effect of these observations on the indication. Thus, the role of the robust estimator is not to "tweak" the indication every quarter, but to bring the indication down to meaningful levels when unusually large observations push the indication to unrealistically high levels.

The graphs in Appendix C show the performance of each of the robust estimators compared to the untruncated estimator and the Agricultural Statistics Board. A note about the graphs. The data series for the iterative fence estimator and the once-winsorized estimator begins in June 1988 and continues to June 1993, while the data series for the fixed cutoff estimator begins in June 1991 and ends in June 1994. The first two estimators were computed as part of the research project along with several versions of the fixed cutoff estimator, using different cutoffs. The final fixed cutoff estimator was recalculated after carefully reviewing the cutoff levels by State. In the interest of brevity, we have only included the final fixed cutoff estimator in the graphs. As far as anticipating the effect of the fixed cutoff estimator for time periods not covered, you can assume that it would track very closely to the data driven estimators.

The effect of outliers on the untruncated estimates is apparent for States that had significant outliers in the periods studied. Georgia's unbiased indication was severely effected by a single large NOL unit in the 1989 survey year. A similar situation occurred in Indiana for the 1989 survey year. In the 1992 survey year, Iowa experienced several large NOL tracts in the sample including one tract that expanded to a particularly large number of hogs. North Carolina had some large list frame units that were essentially placed in the wrong stratum in the 1991 survey year. These units thus received larger expansion factors than what they would have received if placed in the correct stratum.

For States with few outliers in the periods studied, none of the robust estimators had much effect of the estimators. In periods where there were outliers, all of the robust estimators reduced the expansion factors to fall more in line with what the correct expansion factors should have been. When the cause of the unusually large estimate was due to a single outlier driving the estimate, each of the robust estimators performed equally well. However, when the cause was due to several large units driving the estimate, the fixed cutoff estimator will generally outperform the data-driven estimators.

RECOMMENDATIONS

The appeal of the data-driven estimators is the perceived objectivity with which the cutoff is derived. However, the fixed cutoff estimators perform as well as or better than any of the data driven estimators when the source of variability is not only the size of outliers but the number of outliers included in the sample. The fixed cutoff estimator has the added advantage that it is simple to apply and does not require multiple passes of the data files to determine the cutoff. A disadvantage of the fixed cutoff estimator is the inflexibility of the cutoffs to changes in the sample design, because the cutoffs are derived based on historical data.

With those caveats in mind, we make the following recommendations:

1) **We recommend the fixed cutoff estimator over the data-driven estimators for improving estimates of hogs at the State level.**

   Based on this recommendation, the Statistical Methods Branch began operational testing of the fixed cutoff estimator in December 1994. Detailed specifications are included in Hicks (1994). The top 16 hog producing States are included in this test. A memo similar to the one shown in Appendix B was sent to the other States requesting their recommended cutoffs. We extended the scope of the original project to include truncation levels for sows farrowed as well as total hog inventory. The State recommended cutoffs were reviewed by looking at the distribution of the largest 20 expanded values for a number of historical surveys. An example of these graphs for the five hog States is included in Appendix D. A summary of the State recommended cutoffs as well as the final cutoffs is included in Appendix E.

2) **If the operational test of the fixed cutoff estimator in December 1994 and March 1995 is promising, we recommend that the Statistical**

Methods Branch provide the top 16 hog producing States this estimator starting in June 1995.

3) **The cutoffs should be evaluated at regular intervals.** To facilitate this review, the summary system should create an output file for each survey of the top 20 expanded values for hog inventory and sows farrowed by State. This file should be appended to previous files to maintain a historical record of the effect of the cutoffs on the data.

4) **We recommend that the fixed cutoff estimator be tested for other NASS surveys where outliers cause extreme fluctuations in survey expansions.**

REFERENCES

Ernst, L.R. (1980), "Comparison of Estimators of the Mean Which Adjust for Large Observations," Sankhya: The Indian Journal of Statistics, 42, 1-16.

Hicks, S. and Fetter, M. (1993), "An Evaluation of Robust Estimation Techniques For Improving Estimates of Total Hogs", presented at International Conference on Establishment Surveys.

Flores-Cervantes, I. (1993), "Handling Outliers with Fences", presentation to Research Division, National Agricultural Statistics Service.

Hicks, S. (1994), "Specifications for Revising the SPS Summary System to Include a Truncation Estimator for Hogs, Sows Farrowed, Pig Crop, and Pigs per Litter," Internal NASS memorandum.

Lee, H. (1991), "Outliers in Survey Sampling," prepared for the Fourteenth Meeting of the Advisory Committee on Statistical Methods.

Potter, F. (1988), "Survey of Procedures to Control Extreme Sampling Weights," Proceedings of the Survey Research Methods Section of the ASA.

Rumburg, S. (1992), "Characteristics of Directly Expanded Hog Data Outliers," NASS Staff Report, No. SRB-92-02, U.S. Department of Agriculture.

Searls, D. T. (1966), "An Estimator for a Population Mean Which Reduces the Effect of Large True Observations," Journal of the American Statistical Association, 61, 1200-1204.

Thomas, D. R., Perry C.R., and Viroonsri, B. (1990), "Estimation of Totals for Skewed Populations in Repeated Agricultural Surveys: Hogs and Pigs," NASS Staff Report, No. SRB-90-02, U.S. Department of Agriculture.

U.S. Department of Agriculture (1983) : Scope and Methods of the Statistical Reporting Service. Publication No. 1308. Washington, D.C.

Appendix A -- Example of the Calculation of the Truncated Estimator

For this example the cutoff is 100,000.

Record	Post-stratum	W_{nr}	DAF	y	Exp_value
1	1	20.0	1.0	500	10,000
2	1	75.0	1.0	2000	150,000
3	1	25.0	1.0	250	6,250
4	1	25.0	1.0	700	17,500
5	1	15.0	.9	0	0
6	2	75.0	1.0	200	15,000
7	2	70.0	1.0	0	0
8	2	100.0	.7	325	32,500
9	2	125.0	.8	400	40,000
10	2	150.0	1.0	800	120,000
11	3	1.0	1.0	200,000	200,000
Untruncated estimate after nonresponse adjustment					581,500

Observations 2 and 10 will be truncated because they exceed the cutoff. Record 11 exceeds the cutoff, but will not be truncated because it is a Prob-1 EO. The truncated weights for records 2 and 10 are calculated as follows:

Record 2

$$W_{trun} = \frac{100,000}{2000*1.0} = 50$$

Record 10

$$W_{trun} = \frac{100,000}{800*1.0} = 125$$

The adjustment factors are calculated for each post-stratum as follows:

Post-stratum 1

$$ADJ = \frac{20 + 75 + 25 + 25 + 15}{20 + 50 + 25 + 25 + 15} = 1.18518519$$

Post-stratum 2

$$\text{ADJ} = \frac{75 + 70 + 100 + 125 + 150}{75 + 70 + 100 + 125 + 125} = 1.05050505$$

The final weights are then calculated as W_{trun} * ADJ and the truncated expanded value is calculated as W_{trun} * ADJ * DAF * y.

Record	Post-stratum	W_{trun}	ADJ	DAF	y	Truncated Exp_value
1	1	20.0	1.18518519	1.0	500	11,851.85
2	1	50.0	1.18518519	1.0	2000	118,518.52
3	1	25.0	1.18518519	1.0	250	7,407.41
4	1	25.0	1.18518519	1.0	700	20,740.74
5	1	15.0	1.18518519	.9	0	0.00
6	2	75.0	1.05050505	1.0	200	15,757.58
7	2	70.0	1.05050505	1.0	0	0.00
8	2	100.0	1.05050505	.7	325	23,898.99
9	2	125.0	1.05050505	.8	400	42,020.20
10	2	125.0	1.05050505	1.0	800	105,050.51
11	3	1.0	1.0	1.0	200,000	200,000.00
Truncated estimate after nonresponse adjustment						545,245.79

Note that after truncation and adjusting the weights the truncated expanded values may exceed the cutoff depending on how much of the expansion was due to the weight and how much was due to an unusually large reported value. Operationally, the post-strata will have more records than in this example. Thus we expect the effect of this adjustment to the final weights to be much smaller than shown in this example. We expect the adjustment factors to be more in the range of 1.03 than 1.18.

Appendix B -- Memo Requesting Cutoffs from State Offices

TO: Deputy State Statistician
State Office

SUBJECT: Hog Outlier Cutoff Values

Research Division has been investigating several Robust estimation techniques which adjust for "outliers", with the goal of improving survey estimates of total hogs at the State level. The estimator that shows the most promise is called winsorization at a fixed cutoff. The algorithm for the estimator is as follows:

1- For each State we choose a cutoff, say c.

2- Within the State we identify all the records within both the NOL and list domains that have expanded values greater than c (excluding prob-one EOs). We truncate the sampling weights (i.e., expansion factors) of those records so that the expanded values equal c.

3- The excess weights are then "smoothed" over all records within that domain. For the NOL domain, the excess NOL weight is smoothed over all records in the NOL domain. For the list records, the excess list weight is smoothed within weighting class (stratum). Usually, this smoothing has the effect of increasing all weights by less than 3 percent.

The key to this technique is determining the best cutoff for each State. We've investigated several techniques for determining the best cutoff at a State level based on the distribution of the data within each quarterly survey. However, we haven't been overly impressed with any particular technique. We need your input.

Based on your knowledge of the QAS total hog estimates in your State we'd like you to recommend reasonable cutoffs for the June and the follow-on surveys. Notice that this estimator is, on average, downwardly biased -- it can only lower the hog indication. Ideally, we'd like a cutoff that only truncates extreme outliers to minimize biasing the State indication. It would be better to have a cutoff that is too large than one that is too small. Over an infinite number of surveys, our analysis indicates we would expect to truncate on average 1 - 3 records in a State per survey. This could vary by State, depending on the likelihood of outliers occurring in the NOL or large list operations being mis-stratified. Some States might average 0 records truncated, and other States might average 5 - 10 records truncated. The number of truncated records could also vary by survey year, since each set of samples are randomly selected, so experience from just the past one or two years may not provide a good indication. It is also possible that a different cutoff is needed for the June survey than for the follow-on surveys, where the NOL records tend to have larger expansion factors. Your recommendation should reflect your best judgement of the largest expanded value that would be "reasonable" from the QAS (excluding prob-one EO's), recognizing that any expanded value larger than this cutoff will be truncated to the cutoff.

Each State currently has an outlier cutoff value used in the SPS Summary to calculate the outlier component of the hog expansion (the sum of all expanded values greater than the cutoff). This outlier component is also used in a Robust estimator calculated at the U.S. level by the Livestock Branch. For the States in this study the current cutoffs are:

Iowa	80,000
Illinois	50,000
Indiana	40,000
North Carolina	30,000
Georgia	25,000

You may want to use these cutoffs as a starting point, but do not feel constrained to chose cutoffs close to them if you feel that they are either too large or too small for this type of estimator. Our preliminary analysis indicates that the Iowa cutoff may be too small, and that the Georgia cutoff may be slightly too large. We appreciate your expert review.

We could conduct extensive analysis on historic data from each State to determine these cutoffs, but this would be time-consuming and quite possibly inconclusive. Or we could make our own best guesses on these cutoffs, but we feel your expert opinions will better serve our project. We will use your recommended cutoffs in analysis of this winsorization method applied to previous survey data sets from your State. If this approach continues to look promising we could ask all other States to provide recommended cutoffs. Your thoughts and comments on this process are important. Please return your recommendations by September 1 using the attached form. If you have any questions please feel free to contact either Susan Hicks, Matt Fetter or Bill Iwig at 703-235-5213. Thanks for your assistance. We will try to provide you with periodic updates on the status of this project.

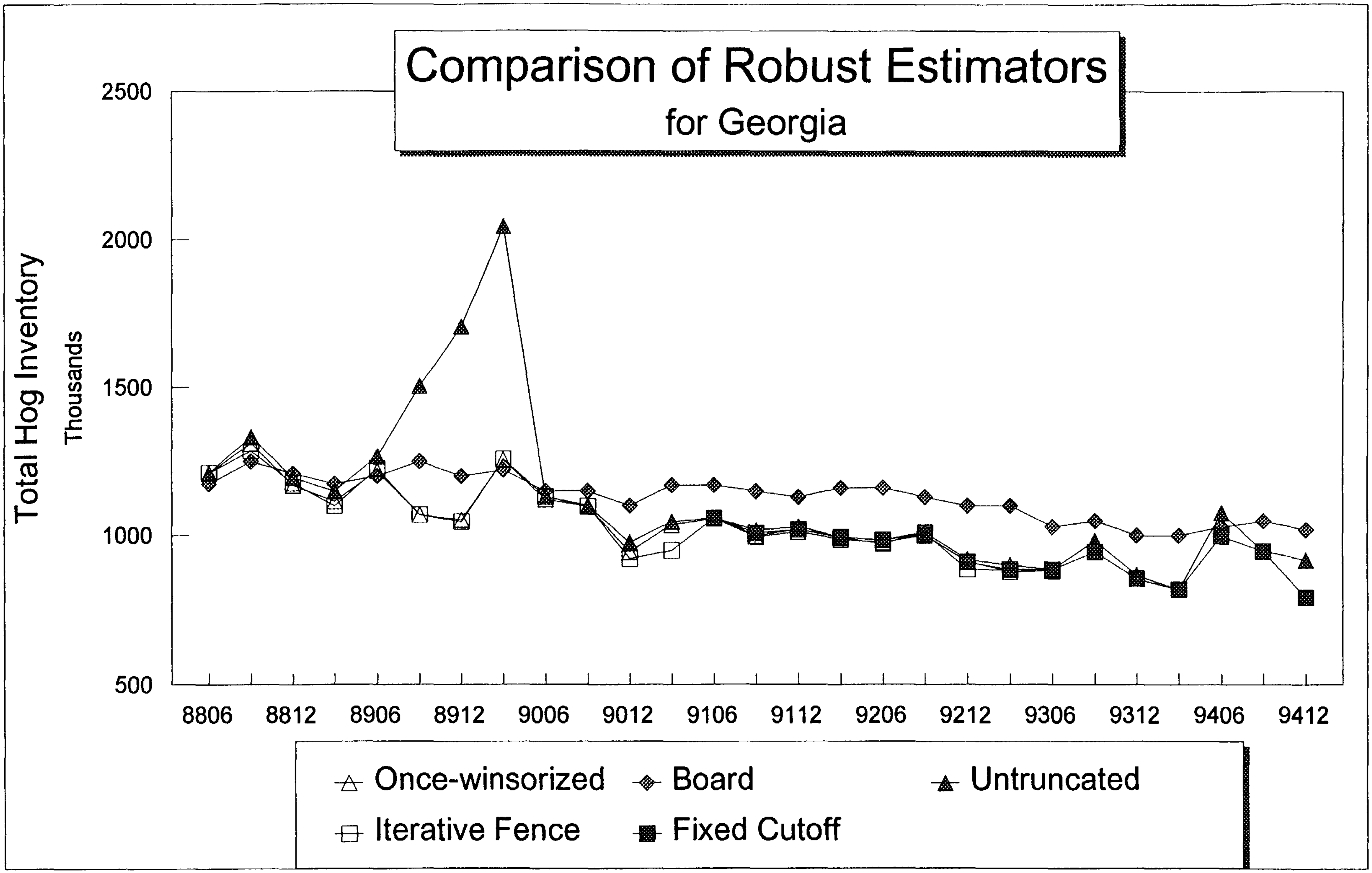
What would be your recommended cutoff for the June surveys?

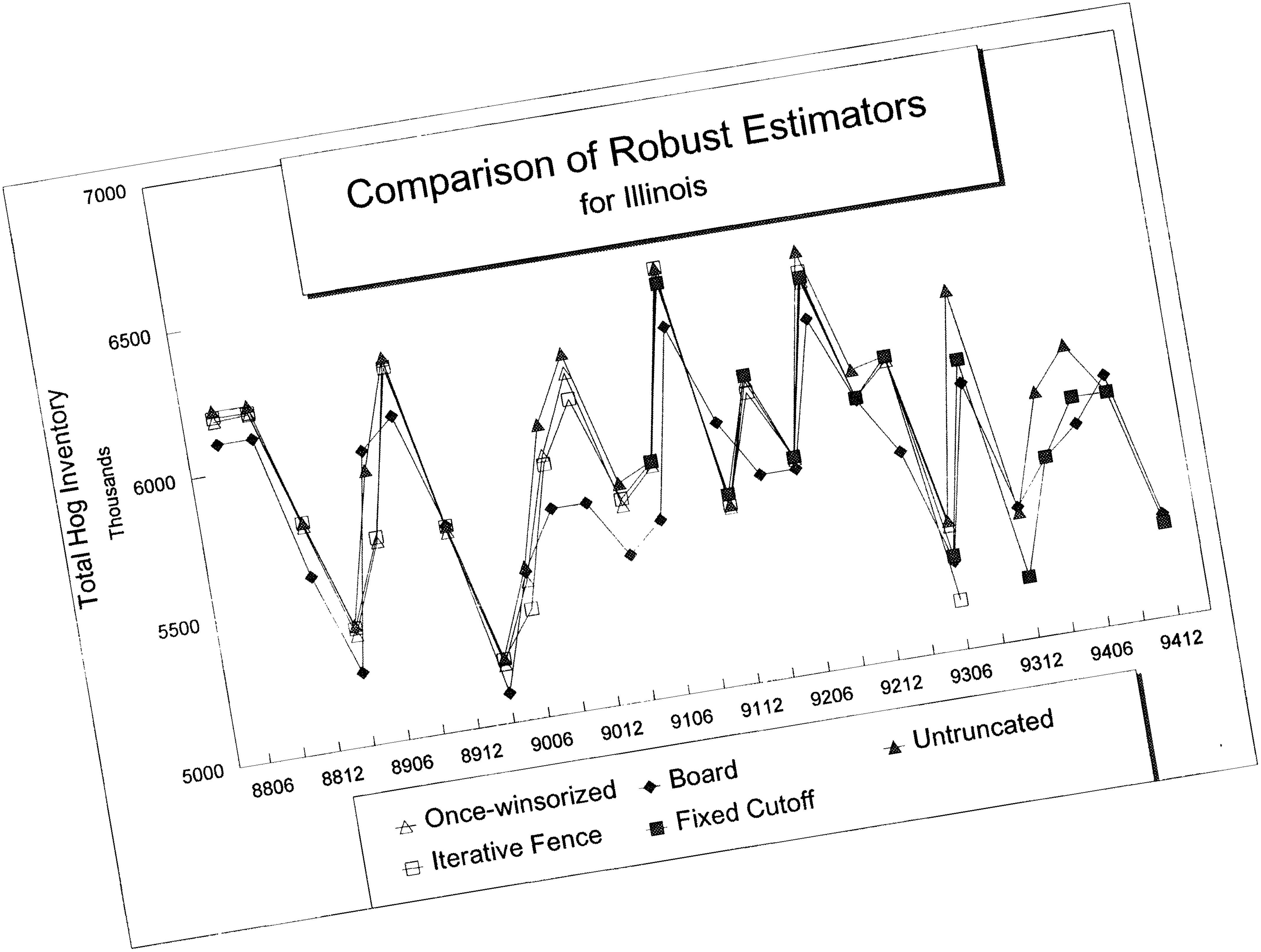
What would be your recommended cutoff for the follow-on surveys?

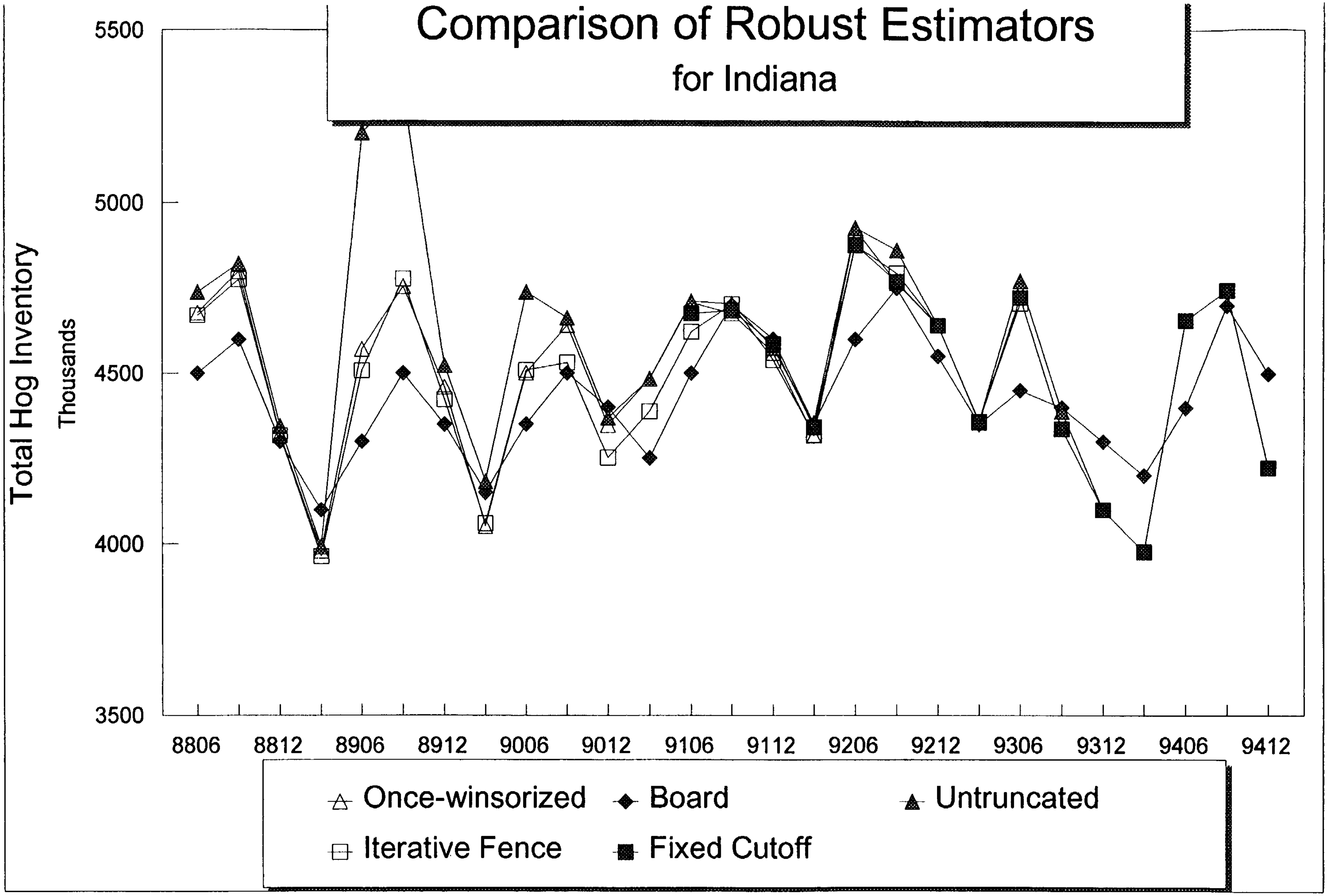
How often do you think these cutoffs should be revised or reevaluated?

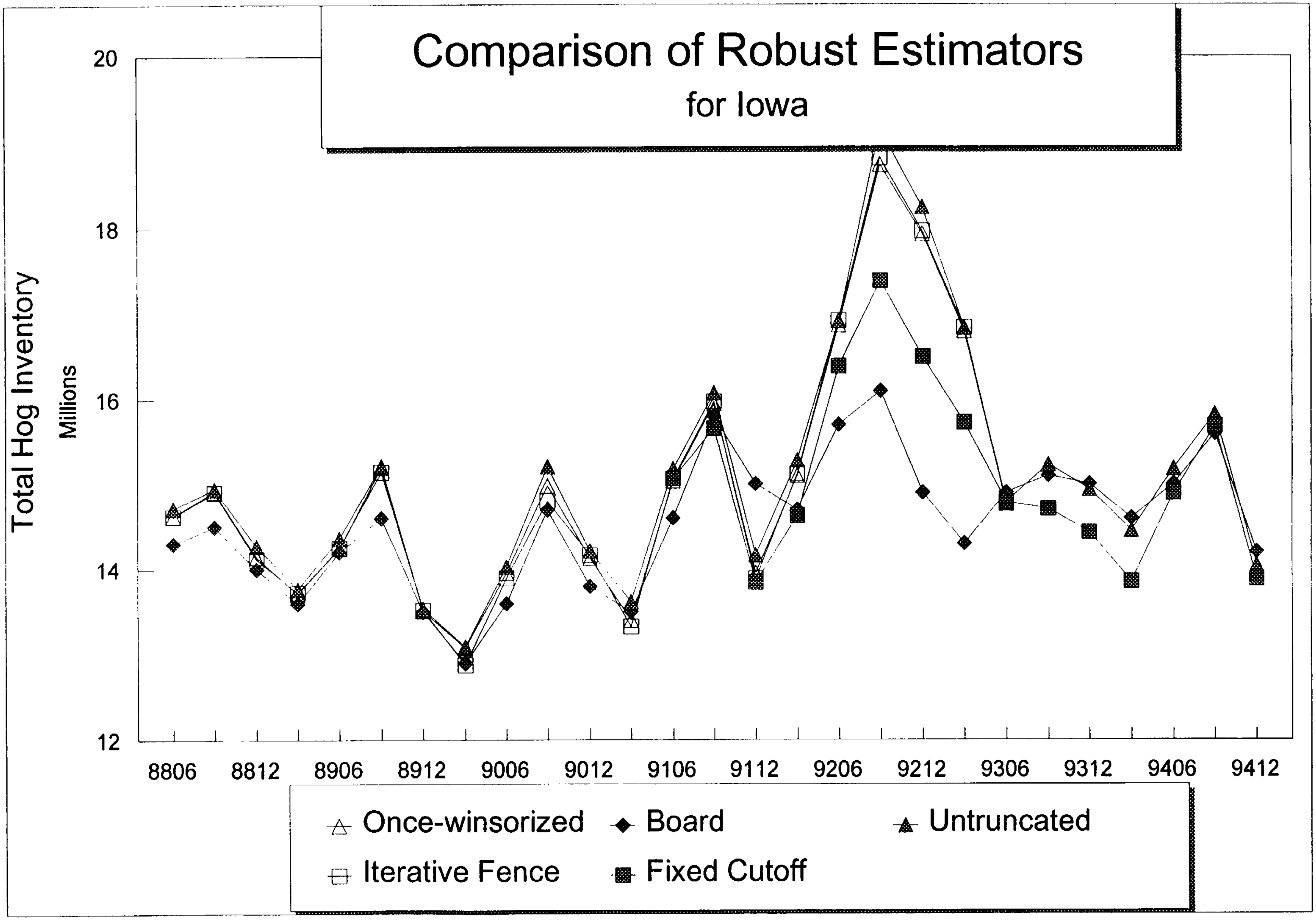
What criteria did you use to determine the cutoffs you recommended?

How do you currently adjust for outliers in your State recommendation?

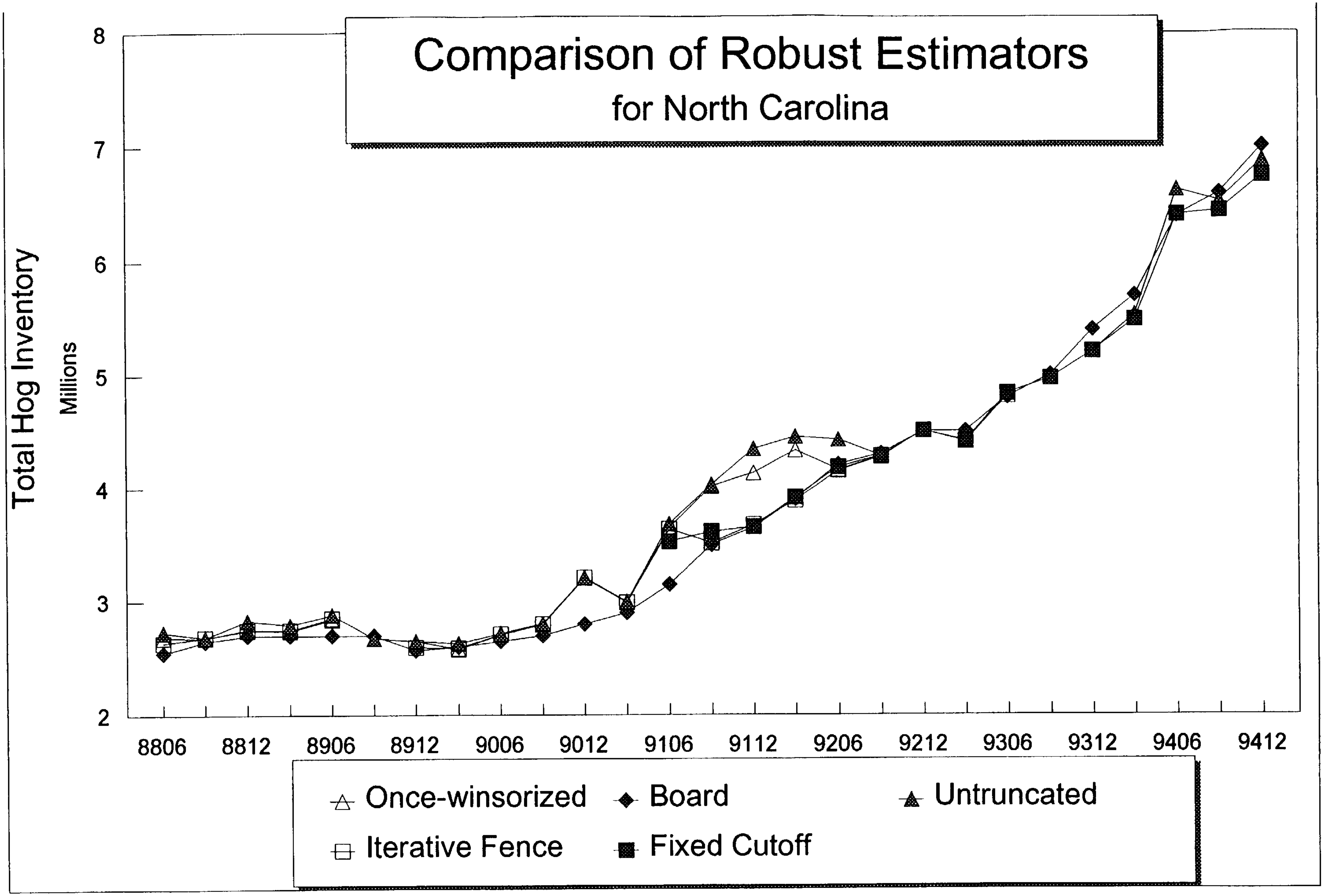

Comparison of Robust Estimators
for Georgia
Total Hog Inventory
Thousands
2500
2000
1500
1000
500
8806
8812
8906
8912
9006
9012
9106
9112
9206
9212
9306
9312
9406
9412
Once-winsorized
Board
Untruncated
Iterative Fence
Fixed Cutoff


Comparison of Robust Estimators
for Illinois
Total Hog Inventory
Thousands
7000
6500
6000
5500
5000
8806
8812
8906
8912
9006
9012
9106
9112
9206
9212
9306
9312
9406
9412
△ Once-winsorized
◆ Board
▲ Untruncated
□ Iterative Fence
■ Fixed Cutoff


Comparison of Robust Estimators
for Indiana
Total Hog Inventory
Thousands
5500
5000
4500
4000
3500
8806
8812
8906
8912
9006
9012
9106
9112
9206
9212
9306
9312
9406
9412
Once-winsorized
Board
Untruncated
Iterative Fence
Fixed Cutoff

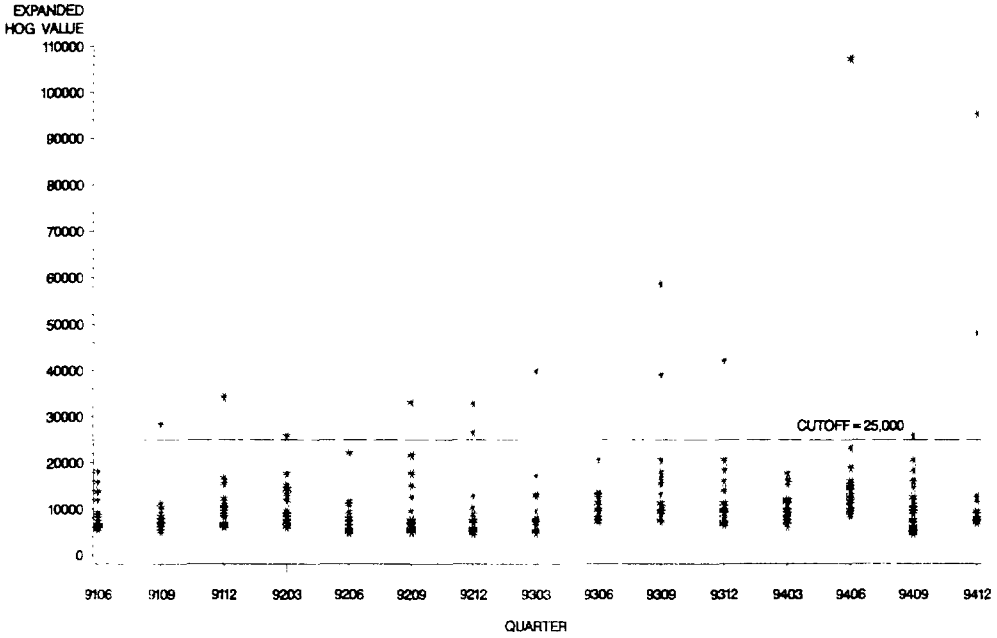


Comparison of Robust Estimators
for Iowa
Total Hog Inventory
Millions
20
18
16
14
12
8806
8812
8906
8912
9006
9012
9106
9112
9206
9212
9306
9312
9406
9412
Once-winsorized
Board
Untruncated
Iterative Fence
Fixed Cutoff

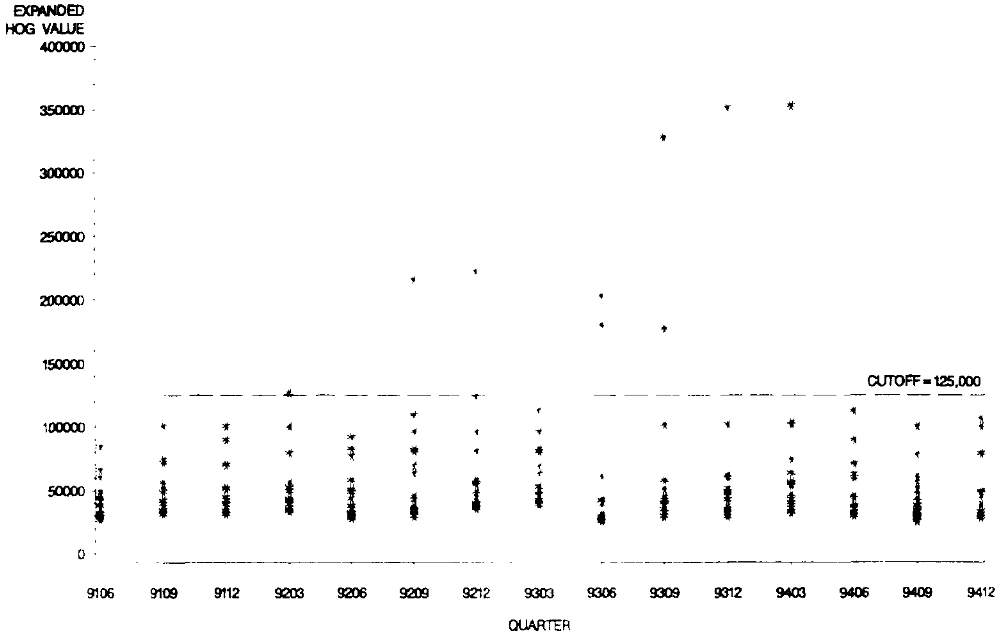

Comparison of Robust Estimators
for North Carolina
Total Hog Inventory
Millions
8
7
6
5
4
3
2
8806
8812
8906
8912
9006
9012
9106
9112
9206
9212
9306
9312
9406
9412
Once-winsorized
Board
Untruncated
Iterative Fence
Fixed Cutoff

Appendix D -- Graphs of Top Twenty Expanded Values vs Cutoffs

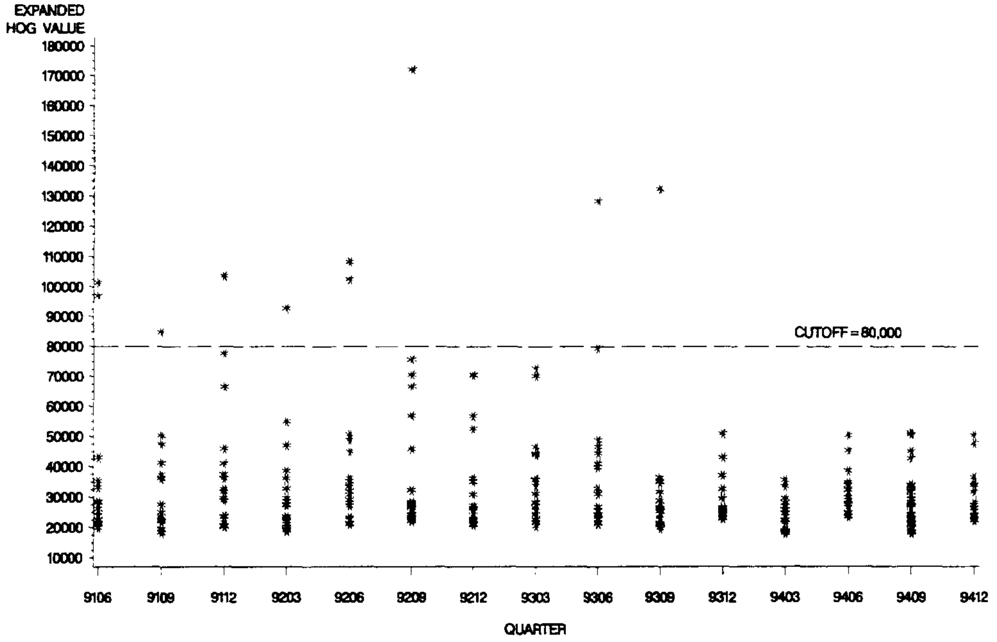
GEORGIA
TOP 20 EXPANDED HOG VALUES EACH QUARTER



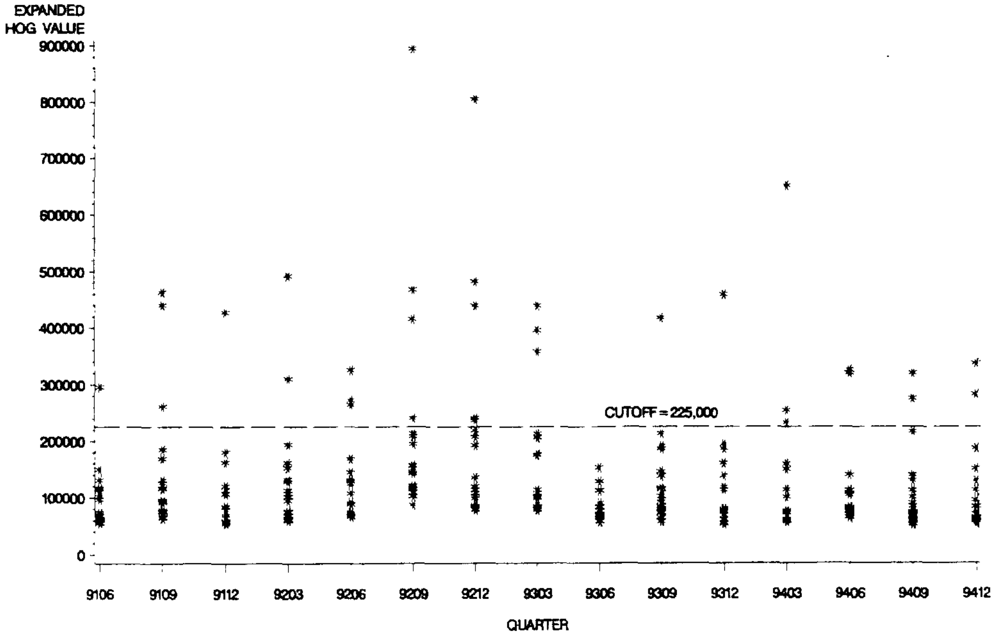
ILLINOIS
TOP 20 EXPANDED HOG VALUES EACH QUARTER

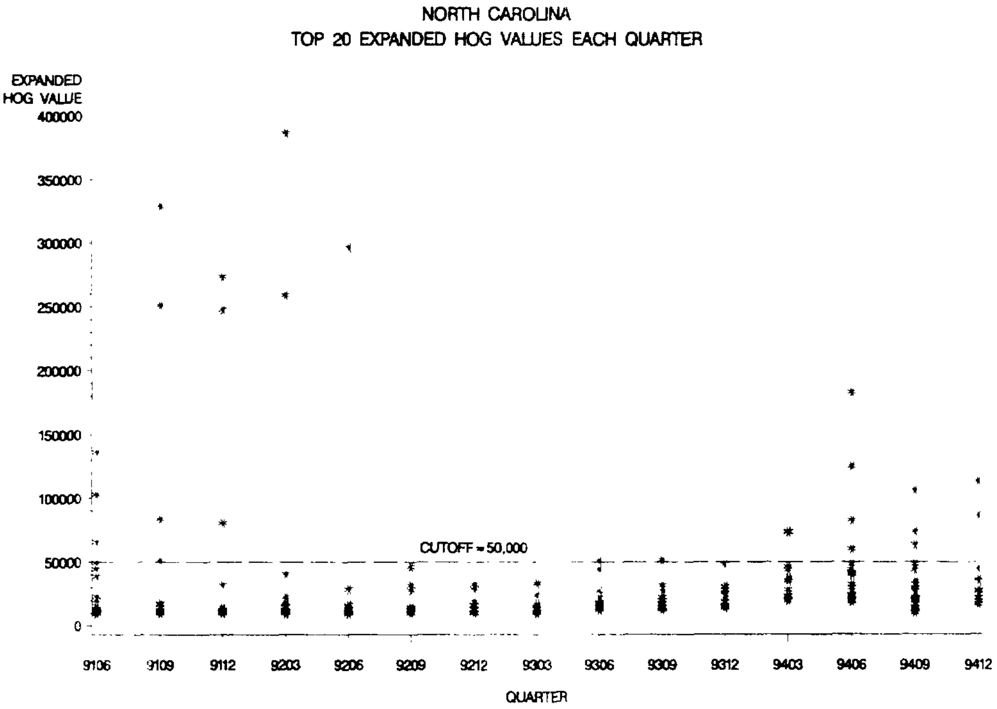


INDIANA
TOP 20 EXPANDED HOG VALUES EACH QUARTER



IOWA
TOP 20 EXPANDED HOG VALUES EACH QUARTER




NORTH CAROLINA
TOP 20 EXPANDED HOG VALUES EACH QUARTER
EXPANDED HOG VALUE
400000
350000
300000
250000
200000
150000
100000
50000
0
CUTOFF = 50,000
9106
9109
9112
9203
9206
9209
9212
9303
9306
9309
9312
9403
9406
9409
9412
QUARTER

endix E -- SUMMARY OF HOG OUTLIER CUTOFF DETERMINATIONS

ollowing table shows the cutoffs recommended by the States and cutoffs determined by examining data 3 Quarterly Agricultural Surveys: March 1992, December 1992, and June 1993. In most cases cutoff s are the same or nearly the same. The average number of values exceeding the cutoffs (requiring ation) is also shown.

	Expanded Hogs Cutoff			**Expanded Sows Cutoff**		
	State	Research	Research Avg. Truncated	State	Research	Research Avg. Truncated
rgia	25,000	25,000	1.00	2,000	2,000	0.67
ois	100,000	125,000	1.33	5-7,000	7,000	0.67
ana	100,000	80,000	0.67	5,000	5,000	0.67
a	125,000	225,000	2.33	20,000	20,000	1.67
sas	20,000	30,000	0.33	4,000	3,000	0.67
tucky	30,000	30,000	0.67	2,000	2,500	0.33
higan	30,000	30,000	1.00	3,000	3,000	1.33
nesota	100,000	100,000	1.33	6,000	7,000	1.00
souri	30-50,000	40,000	1.00	5-7,000	4,000	0.00
raska	200,000	125,000	0.67	10,000	10,000	0.67
th Carolina	50,000	50,000	1.00	5-8,000	5,000	0.67
o	50,000	60,000	1.33	5,500	6,000	0.33
nsylvania	100,000	50,000	2.33	5,000	5,000	0.33
h Dakota	50,000	50,000	0.67	5,000	5,000	0.33
nessee	25,000	35,000	0.67	1,500	2,500	0.33
consin	50,000	60,000	1.00	6,000	6,000	0.00

alues in the following table are from the 12 Quarterly Agricultural Surveys making up the 1991, 1992 and survey years.

	Number of Values Truncated Using Research Cutoffs				**Average Number Values Truncated**	
	NOL	List	Total	Average	Hogs	Sows
rgia	6	10	16	1.33	0.83	0.50
ois	14	6	20	1.67	0.75	0.92
ana	20	2	22	1.83	0.83	1.00
a	42	0	42	3.50	2.25	1.25
th Carolina	9	20	29	2.42	1.50	0.92