# Modeling Nonsampling Errors in Agricultural Surveys

James Gentle[1], Charles Perry[2], William Wigton[2]
George Mason University[1], National Agricultural Statistics Service[2]
E-mail: jgentle@gmu.edu

## Abstract

Nonsampling errors present major problems in sample surveys. While sampling errors can be estimated and can be controlled by increasing the sample sizes sufficiently, neither the existence nor the magnitude of nonsampling errors can be predicted, and furthermore there is no simple means for controlling them. Increasing the sample size may actually increase the nonsampling errors. The existence of nonsampling errors often becomes apparent only when the results of two surveys disagree in excess of what could be accounted for by sampling variability. The National Agricultural Statistical Service of USDA conducts a census of US agriculture every five years in December, and conducts a survey of agriculture in June of every year. There is generally good agreement for most farms between the census and the surveys. Because of changes in farm operations over time, we expect a certain amount of deviation of the samples from the census. When the June surveys of the years 2002 through 2005 are compared with the census of 2002, however, for a small number of farms, the deviations exceed what would be expected due to ordinary changes in farm operations. Investigation of a sampling of the original records for the surveys confirmed the existence of nonsampling errors. We would like to be able to identify in advance the types of farms whose records are likely to contain nonsampling errors. In this work, we seek to apply methods of supervised classification to develop classification models for identifying records likely to contain nonsampling errors. The training samples for the classification analysis ultimately will consist of three classes of records: those in which the survey agrees with the census; those in which the survey does not agree with the census, but for which follow-up has not confirmed the existence of nonsampling errors; and those with known nonsampling errors. In our preliminary studies, we develop a measure of the magnitude of apparent nonsampling errors. Our initial classification studies assume binary classes, and then we analyze differences in frequencies of apparent error magnitudes for different classes of observations. At this stage the work is exploratory, but we believe that as we improve the error rates (both false positive and false negative) of our classification models, we can be in a better position to reduce the nonsampling errors.

**Keywords**: nonresponse, comparison of multiple surveys

## 1 Introduction: Nonsampling Errors

In statistical surveys we generally distinguish two kinds of errors: sampling and nonsampling. Sampling errors are the general differences between totals based on appropriately weighted sample averages and the true totals of the population. Sampling errors are not really "errors", in the sense of mistakes; rather, these errors are the result of observing a sample rather than the full population. Sampling errors are controllable by the sampling design. They are also controllable by sample size; as the sample size increases, the magnitude of the sampling errors decreases. One of the most important characteristics of sampling errors is that they are estimable. The general theory of sampling, together with the estimability of sampling errors, makes these errors predictable and manageable.

Nonsampling errors, on the other hand, lack a theory that enables estimation or control by sampling design.

### *Causes of Nonsampling Errors*

Two common causes of nonsampling errors are incorrect coverage and nonresponse. Coverage problems generally arise because of incorrect frames. Nonresponse problems occur for a variety of reasons, and are often closely related to the nature of the information being requested.

Another common cause of nonsampling errors is that the raw data that are gathered are incorrect. This can happen due to the respondent not understanding the question or having inadequate knowledge to respond correctly. If the data are collected by interviewers or surveyors, the person collecting the data may make mistakes. On the other hand, if the data are collected automatically, the scanners or sensors may be incorrectly calibrated.

In the case of nonresponse, if raw data are generated by imputation, either on the spot by the interviewer, or later by formulas involving other data, the raw data generated in this way are likely to contain errors.

Finally, nonsampling errors can occur due to data processing errors after the raw data have been collected.

### *Control of Nonsampling Errors*

The causes of nonsampling errors imply that any method for control of these errors must involve the improvement of the process. This may mean revision of questionnaires, better training of the field workers, general revision of the methods of data collection, revision of frames, and so on.

An important part of the process is awareness of which observational units or items are more subject to nonsampling errors than others. When such units or items are recognized, the implication for the survey process is to devote different levels of resources to different observational units.

While increasing the sample size decreases sampling errors in known ways, increasing the sample size may actually exacerbate the problem of nonsampling errors. This is because the process itself is more difficult to manage, and it is the control of the process that controls nonsampling errors. Larger sample sizes generally require more resources (money), and so increasing the sample size without commensurably increasing the resources likely increases the frequency of nonsampling errors. Conversely, decreasing the sample size may decrease the frequency of nonsampling errors that more than compensates for the increase in sampling errors.

*Identifying Nonsampling Errors*

Unlike sampling errors, which can be reliably estimated, nonsampling errors are difficult to identify or measure. If an item being sampled has known limits, or reasonable limits from which an arbitrary limit can be set, such as in the case of the age of a person who can be a parent, then we can identify nonsampling errors for any items not satisfying the limits. (A "parent" with a reported age of 5 years is a nonsampling error.) Standard data edits are often used to identify these errors, but they can only indicate the existence of an error and place lower bounds on its magnitude.

Another common method of identifying nonsampling errors is by use of repeated observations or redundant items. When two surveys are conducted over the same or overlapping populations, and some variables are the same on each, nonsampling errors may be identified by comparing identical respondents' answers.

Surveys often contain redundant items, such as the numbers of individual types of a particular class and then the total of all types of that class. Standard data edits are also used to assess the internal consistency of responses, but again, they can only indicate the existence of errors and place lower bounds on the magnitude of the errors.

In either repeated observations or redundant items, there is an inefficiency in the surveys; the repetitions or the redundancy comes with a cost whose only direct benefit is in identifying nonsampling errors.

Different surveys, however, often contain data on the same item because the surveys are conducted at different points in time, or because of overlaps in the frames. Atkinson (1997) and Bassi and Fabbris (1997) suggest ways of identifying nonsampling errors in multiple surveys, and Scali et al. (2004) describe ways of assessing and measuring nonsampling errors in a particular kind of data. Identification of nonsampling errors remains a major problem, however.

An important question is whether there are systematic patterns that are associated with nonsampling errors, and if so whether these patterns can be identified. This is the objective of this research for surveys conducted by the National Agricultural Statistical Service of USDA. In the work reported in this paper, we have addressed only simple patterns that merely rely on the values of certain items on the surveys.

## 2 Nonsampling Errors in Agricultural Surveys

Agricultural surveys are uniquely prone to nonsampling errors because of the diverse nature of the observational units. Some of these particular problems have been pointed out by Faulkenberry and Tortora (1981), Warde (1986), and Lesser and Kalsbeek (1999).

The first problem in any survey, of course, is defining the objectives of the survey and then identifying the relevant population to survey. If the objective in an agricultural survey is to estimate the total output of a particular agricultural product, for example, then the survey must include or sample all important producers. A significant amount of some products, however, is produced by a large number of operations that individually are very small. This situation creates a problem for agricultural surveys because the smaller operations are difficult to identify, and the operation is likely to change from year to year, as the operator chooses to discontinue or reinitiate the operation. By whatever definition is used, a "farm" last year may not be a "farm" this year. Likewise, a large garden and a few livestock that did not constitute a "farm", grow into a "farm".

*Frames in Agricultural Surveys*

A list frame provides a very simple approach to any survey, both for the person designing the survey and for the person actually collecting the data.

The problem is in maintaining a list. In agricultural surveys, a major problem in maintaining a list is the definition of a "farm". The most obvious definition of a farm would be based on some threshold of annual revenue or expected revenue from agricultural operations. As mentioned above, however, no matter what is the definition, there will be substantial annual movement into and out of the population by the marginal operators.

In practice, a list will almost always either undercover or overcover — or do both.

For surveys of farms, because a farm has a close relationship to geography, an area frame can provide total coverage.

## 3 Agricultural Surveys Conducted by USDA

The National Agricultural Statistical Service of USDA conducts various surveys in which the observational units are individual farms. Two important surveys, one a sample and the other a census, are

- June Area Survey; every year.
  - Based on an area frame that covers the entire United States.
  - Predominantly face-to-face interview conducted by a field representative who visits the farm and attempts to interview the farm operator or manager.
  - Conducted during first two weeks of June and asks for information as of June 1.
  - Purpose is to get a picture of the status of American agriculture early in the major crop-growing season.
- Census of Agriculture; every five years
  - Predominantly by mail to a list, with telephone follow-up.
  - Asks for information as of December 31 or for yearly totals.
  - Data collected during January through March of following year.
  - Purpose is to provide a broad structural view of American agricultural.

*Common Variables*

Out of over 800 items included on the two surveys, 25 to 30 are the same.

The variables that are common to the June Area Survey and the Census generally measure stable aspects of a farm, such as acreage in pasture, acreage under lease, and so on.

Total land in the farm is one of the variables on both the June Area Survey and the Census. In Figure 1, we show a scatterplot of the total land on the farms of each of each of four states as reported on the June Area Survey for 2002 and on the Census of 2002 (December).

We would expect most of these points (farms) to fall along 45° lines. We notice, however, that there are several farms in Iowa for which the total land variable reported on the Census of 2002 was quite different from what was reported on the June Area Survey for 2002.

*Discrepancies*

For each variable that is common to the June Survey and the Census, we defined a standardized categorical discrepancy variable based on

$$\frac{|X_{\text{Census}} - X_{\text{June}}|}{X_{\text{Census}} + X_{\text{June}}}.$$

We used a log transformation and discretized the measure into 5 categories, corresponding roughly to

- less than 5%,
- 5% to about 10%,

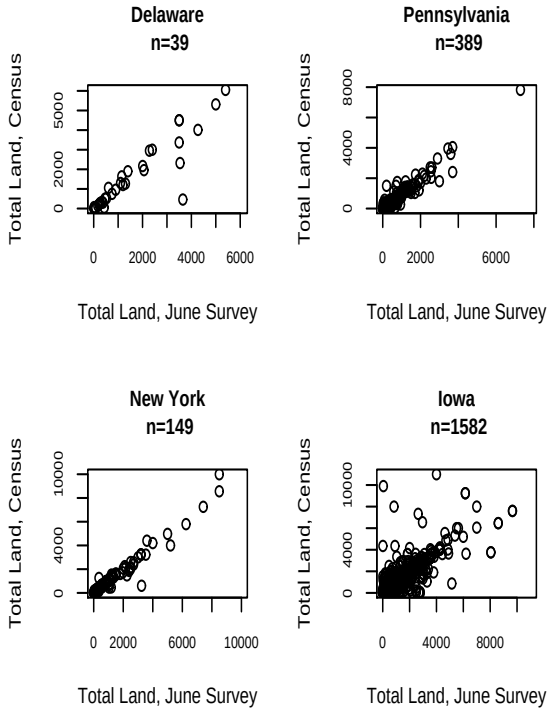


Figure 1: Total Land in Farms

- 10% to about 20%,
- etc.

We must emphasize that discrepancies do not indicate that a nonsampling error has occurred. As we have pointed out earlier, identification of a nonsampling error generally requires measurements or responses that are directly comparable. The responses for the same variables for a given farm on the June Area Survey for 2002 and the Census of 2002 could correctly be different because of changes of the farming operation. We expect these changes to be relatively small, however, so we proceed to analyze the discrepancies with respect to other variables.

*Analysis of Discrepancies*

The discretized discrepancies form the basis for various classification models. These models can be used directly as the levels in factors in linear models in which any of the other variables are the responses, or alternatively they can be treated as the responses in supervised classification.

Our analysis so far has just been exploratory. We are looking for apparent relationships between the discrepancies and other variables.

We have explored such things as differences among the states, differences among the discrepancy variables themselves, and possible relationships between one or more discrepancy variables and such obvious factors as who provided the responses to the survey.

There was generally good agreement for most farms between the census and the surveys. Because of changes in farm operations, we expect a certain amount of deviation of the samples from the census. Differences can be attributed either to actual changes over time over that six-month interval or to nonsampling errors.

In this project, we looked at the differences to see if we could identify any systematic patterns. We did not have information on whether the differences were due to nonsampling errors or to actual changes.

For each state, there are 25 to 30 discrepancy variables.

Some discrepancies, for example, in "grain storage capacity", tended to be larger than others.

- The first conclusion from this study, therefore, is that special attention needs to be given to the instrument that measures these variables (that is, the wording of the questionnaire, together with any auxiliary instructions).

We decided to focus on the total land variable and on the state of Iowa, from the June Survey of 2002 and the Census of 2002. We first did a binary classification tree on the discrepancy variable, allowing several of the other variables to enter the model as classifiers. The two-node tree shown in Figure 2 indicates that the two most significant variables (by the minimum deviance criterion) were the "respondent code" and the "reporting unit code". The respondent code variable" indicates who actually was interviewed (operator, spouse, partner, etc.), or that the operator/manager/spouse refused to cooperate, or that it was never possible to contact the operator/manager/spouse in person. The reporting unit code indicates the nature of the operation, whether it is owned by an individual, owned jointly by two partners, or has some other ownership structure.

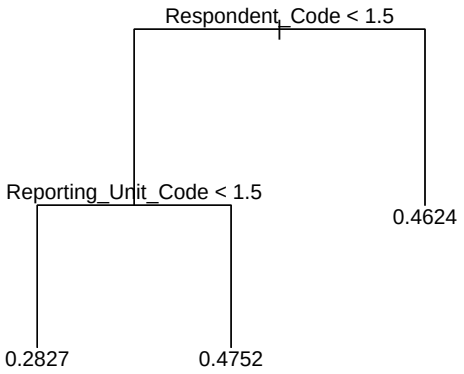


Figure 2: Classification Tree

A value of 0 in the binary classification variable indicated a discrepancy of less than 5% and a value of 1, a discrepancy greater than 5%. Both the respondent code and the reporting unit code had positive integral values. A respondent code of 1 meant that the respondent was the operator. A reporting unit code of 1 meant that the operation is owned by an individual legal entity. In Figure 2 we see that for the single individual reporting units for which the operator supplied the data on the June Survey, over 80% had discrepancies of less than 5%.

We do not expect any single variable, or even a small set of variables to be strong predictors of the amount of discrepancy, We are continuing these kinds of studies, nevertheless, but for the rest of this paper, we turn to analyses of the patterns of discrepancies within these two variables. We also consider the relationship of the discrepancies to two additional characteristics of the farm.

*Respondent Codes and Discrepancies*

The field representative who conducts the interviews for the June survey attempts to interview the farm operator or manager. It may not be feasible to interview the ideal respondent, however. It would seem likely that the discrepancies would be greater if the data were not provided by the ideal respondent. We grouped the respondents into four different types: "operator", "spouse", "partner", and "other", and investigated the relationship between the respondent code and the discrepancy for the total land variable for farms in the state of Iowa. In Figure 3, we show a histogram of the extent of discrepancy for the each type of respondent. A relationship can be seen in the shapes of the histograms of the discrepancy for the different types of respondent.

The bars in each histogram represent the six categories of discrepancy that we defined. The leftmost bar in each case represents less than 5% discrepancy.

It clear that the discrepancies are least when the respondent is "operator". The responses by the "spouse" appear somewhat less reliable; that is, in the histogram, there are greater frequencies of larger discrepancies. It is possibly surprising that the responses by "partner" showed even greater discrepancies. Possible reasons for this include the fact that the person listed as "partner" often has no direct involvement with the farm. These results suggest that greater effort should be spent on getting responses from the "operator", rather than from the spouse, partner,or other possible respondents.

*Discrepancy by Source of Response*

During the June Survey period sometimes it is not possible to get a response from the operator, spouse partner, or other person involved with the farm operation. In those cases, it is necessary for the field worker fills out a portion of the questionnaire relating to crops and land use on the tract by observation, and the rest of the questionnaire is completed in the state office using the latest available data to complete as much of the questionnaire as possible.

In Figure 4, we see that these estimated amounts tend to yield more discrepancies than amounts supplied in a face-to-race interview.

*Discrepancy by Type of Operation*

The simple classification model indicated that the type of reporting unit was a relatively good predictor of discrepancies.

As we saw in Figure 3, responses by partners in a farming operation show greater apparent discrepancies than responses by the operator or by the spouse. We may therefore expect to see greater apparent discrepancies in operations in which there is a partner, and in Figure 5, we see this is the case. Figure 5 shows histograms of the extent of discrepancy for the each type of operation. A relationship can be seen in the shapes of the histograms of the discrepancy for the different types of respondent.

*Discrepancy by Intensity of Agriculture in the Area*

We next looked at whether there were differences in the amount of discrepancies within various strata that are defined by the intensity of the agricultural operations on the land. We might expect that agricultural data would be of higher quality in regions where agriculture is more important; that is, where the land is more intensely cultivated.

In addition to discrepancies between the data for the same farm, in regions where the land is not intensely cultivated, we may also expect more differences in county totals between the June Survey and the December Census because of the phenomenon alluded to above; there is likely to be more marginal farms that lose their "farm" status, and other small operations that gain a "farm" status.

In Figure 6, we indeed see that there are more discrepancies in the regions with less cultivated land than in those with more cultivated land.

## 4 Discussion

As we have discussed, identification of nonsampling errors requires comparisons of data, either within a single survey or between two surveys. We have studied data collected in the June Area Survey for 2002 and on the Census Agriculture of 2002, and have compared common items from these two surveys. We formed a categorical measure of the discrepancies, and studied the extent of the discrepancies within various subgroups of the June sample. There could be true differences in these items because of a time difference in when the data are reported; hence, a discrepancy is only an indicator of a possible nonsampling error.

At this stage the work is exploratory. We seek to relate the magnitude of the discrepancies to other variables on the surveys. In this way, we hope to be able to identify records that are more likely than others to contain nonsampling errors.

*Conclusions and Current State of the Work*

We have seen that some variables tend to have larger discrepancies than others. We have seen that some states tend to have larger discrepancies than others.

We have also seen that different frequencies and different sizes of discrepancies are associated with different types of respondents and different ways the data are collected.

All of these findings have implications for further improvement of the process of data collection.

*Future Work*

As the identification of problem areas within the surveys leads to process improvements (hopefully!), the nature of the discrepancies will change.

As a result of previous analyses of nonsampling errors, NASS has already added some features to the survey analysis tools. These tools allow more immediate comparisons during the collection of the June Survey data, and analysts can supply annotation of differences.

As we continue this work we hope not only to identify "hot spots", but also to develop bases for data adjustments to reduce any systematic biases resulting from nonsampling errors.

## Acknowledgements

The views expressed here are not necessarily those of NASS or USDA. We thank Carrie Davies and Mike Bellow for their contributions to this study. We also thank Roberta Pense, Jaki McCarthy, and others at NASS for comments.

## References

Atkinson, Dale (1997). Assessing nonsampling errors in survey data through random intercept models, *ASA Proceedings of the Section on Survey Research Methods*, 401–406.

Bassi, Francesca, and Fabbris, Luigi (1997). Estimators of nonsampling errors in interview-reinterview supervised surveys with interpenetrated assignments, in *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, John Wiley & Sons, New York, 733–751.

Clusen, Nancy; Kasprzyk, Daniel; Schone, Eric; and Williams, Thomas (2002). Nonsampling error in a survey of department of defense health care beneficiaries, *ASA Proceedings of the Joint Statistical Meetings*, 585–590.

Faulkenberry, David, and Tortora, Robert (1981). Nonsampling errors in an agriculture survey , *ASA Proceedings of the Section on Survey Research Methods*, 493–495.

Lesser, Virginia M., and Kalsbeek, William D. (1999). Nonsampling errors in environmental surveys , *Journal of Agricultural, Biological, and Environmental Statistics* **4**, 473–488

Lessler, Judith T., and Kalsbeek, William D. (1992). *Nonsampling error in surveys*, John Wiley & Sons, New York.

Scali, Jana; Testa, Valerie; Kahr, Maureen; and Strudler, Michael (2004). Measuring nonsampling error in the Statistics of Income individual tax return study, *ASA Proceedings of the Joint Statistical Meetings*, 3520–3525.

Warde, William D. (1986). An investigation of nonsampling errors in U.S. Department of Agriculture surveys, *ASA Proceedings of the Section on Survey Research Methods*, 580–585.

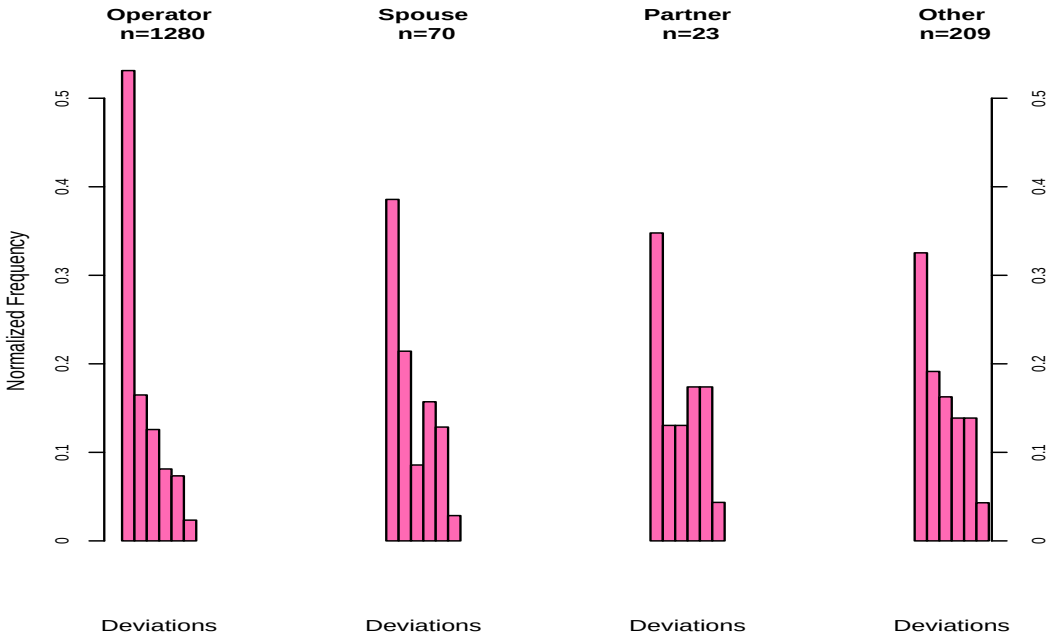


Figure 3: Discrepancy by Type of Respondent

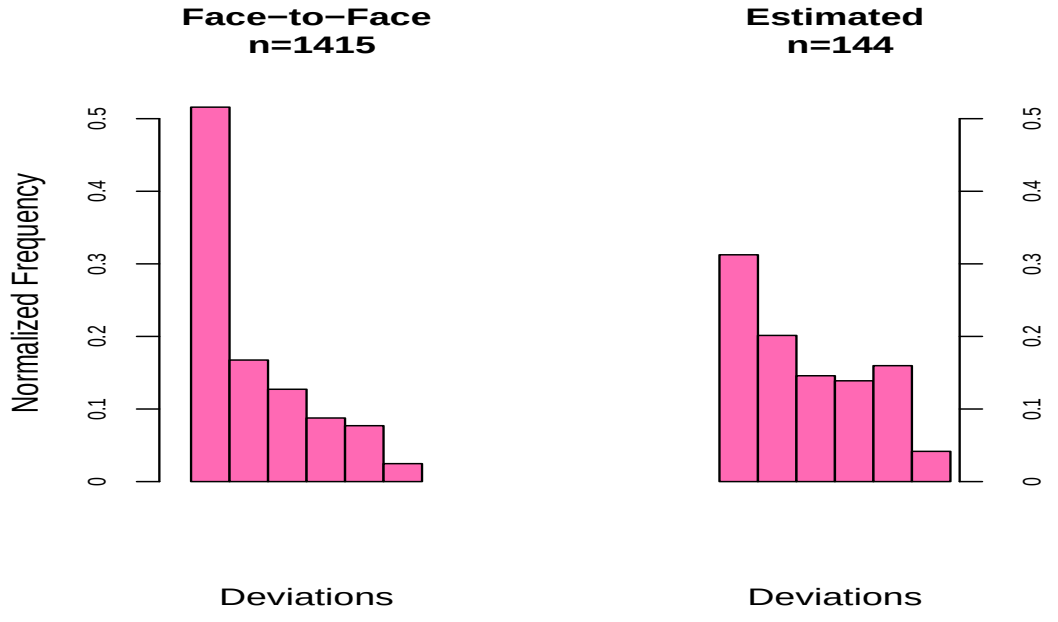


Figure 4: Discrepancy by Source of Response

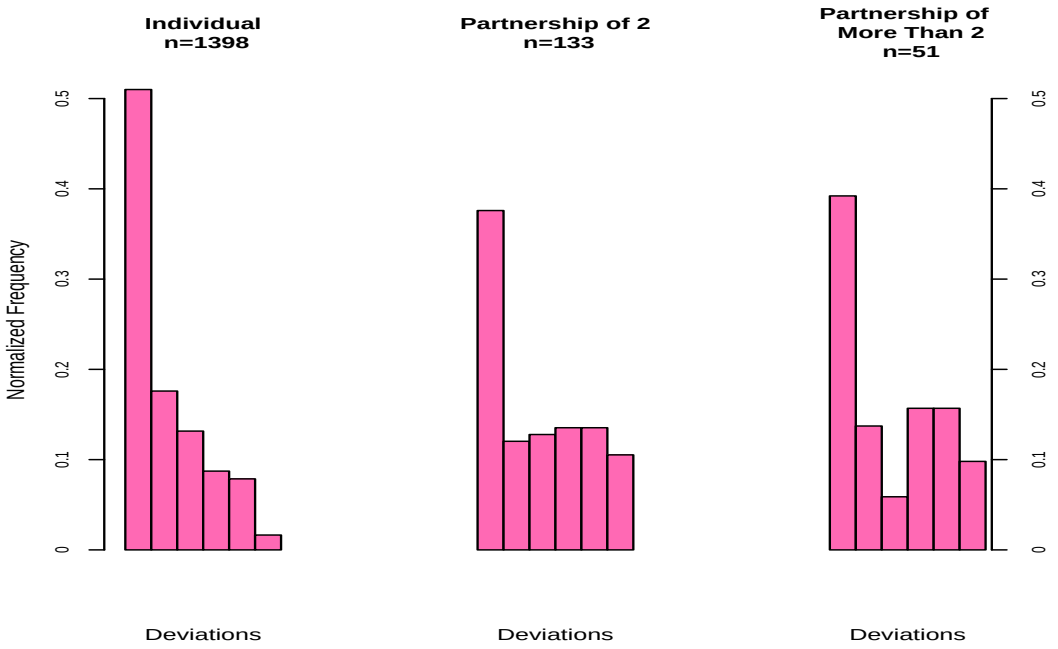


Figure 5: Discrepancy by Type of Operation

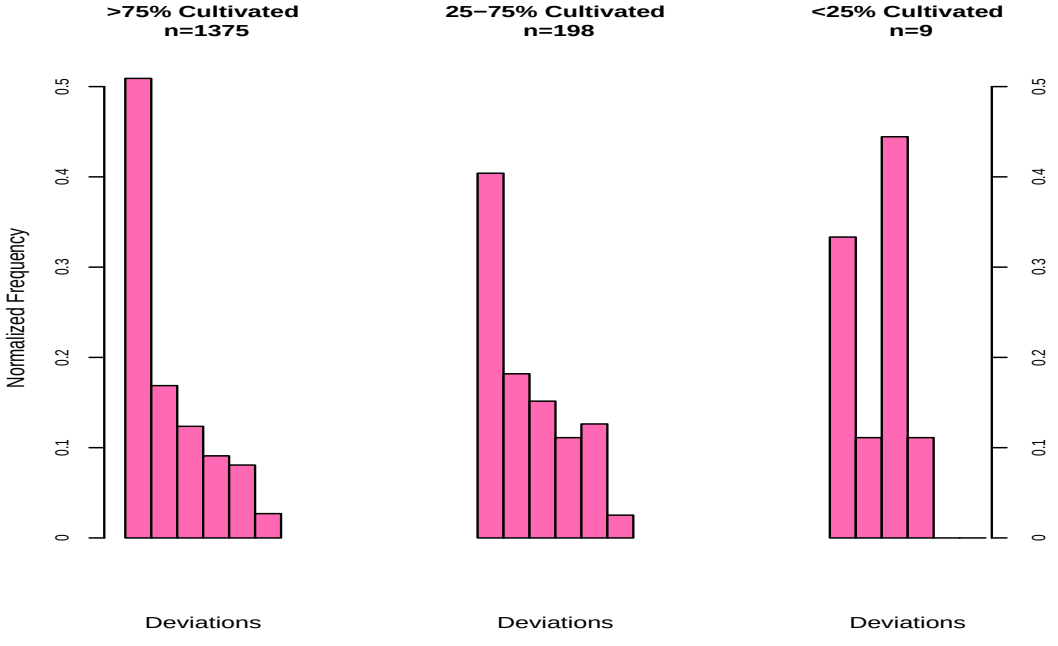


Figure 6: Discrepancy by Intensity of Agriculture in the Area