Evaluating data collection quality after protocol changes

Cherié J. Alf[*] Sarah M. Nusser[†] Veronica C. Lessard[‡]

Abstract

It is important to evaluate data quality in any survey, but especially after changes in protocols and measurement systems. When revisions have been made in a longitudinal survey, it can be particularly important to make estimates of error rates for the entire data collection process and possibly for specific domains associated with the process. We examine this problem from the context of the National Resources Inventory, a longitudinal survey of the nation's natural resources. Our goal was to evaluate the quality of the data collection process after protocol and organizational changes were implemented for the 2005 survey. Because estimates of error rates were of interest, we used a probability sample to select segments for quality review. The quality review sample design allowed for oversampling of sample units more prone to errors. We evaluate the design strata and illustrate how this approach can be used to assess potential problems in the data collection process.

Key Words: Quality Review, Process Quality

1. Introduction

It is important to evaluate data quality in any survey, but especially after changes in protocols and measurement systems. Lyberg and Elvers (2003) identify three different dimensions of quality associated with a survey: product quality, process quality, and organizational quality. Reviewing the data collection process after changes in protocols and measurement systems is a form of process quality evaluation.

This paper presents a case study for evaluating the process quality of data after changes in protocols and measurement systems. We focus on the quality of the data collection process for the National Resources Inventory (NRI) after major reorganization and protocol changes were implemented. The objective of our analysis is to estimate the error rate for survey variables and understand how the error rates vary in relation to factors that may affect data collection quality. We also evaluate the effectiveness of the sampling method used to acquire the quality review sample.

We begin by briefly reviewing literature regarding quality review methods. Then we give an overview of our motivating survey, the 2005 National Resources Inventory. Next we discuss the quality review sample design and how review data were used to estimate various kinds of error rates. Lastly we discuss our results and offer concluding remarks.

2. Background

2.1 Quality Review Methods

In almost all survey organizations some type of quality review is conducted for survey operations to ensure final results are of acceptable quality. Usually this is based upon some form of data inspection. The inspection method often involves post-selecting a sample of units and inspecting this sample for errors. Biemer and Caspar (1994) state that this traditional inspection method for quality review possesses limitations, such as additional time and costs, a failure of data collectors to take responsibility for the quality of their work (since quality is the job of the reviewer), and the perception that this process is unproductive for data collectors. Thus they advocate an alternative methodology that monitors the process quality throughout the data collection period.

Groves and Heeringa (2006) also note the necessity of continually monitoring the streams of process data and survey data. They suggest that a quality review design that monitors the process quality may make use of paradata. Paradata are data on the data collection process, such as keystroke files or call histories. Other kinds of information may be collected from a reviewer who monitors the data collection process as it occurs. When establishing a system for monitoring the process quality, Morganstein and Marker (1997) suggest a design process that involves identifying the critical data characteristics, understanding the data collection process and identifying the factors that have a

[*] Center for Survey Statistics and Methodology, Department of Statistics, Iowa State University

[†] Center for Survey Statistics and Methodology, Department of Statistics, Iowa State University

[‡] USDA Natural Resources Conservation Service

[0] This research was supported in part by cooperative agreement 68-3A75-4-122 between the USDA Natural Resources Conservation Service and Iowa State University.

large affect on the data characteristics as some of the important considerations when designing a quality review sample.

The information gathered during the quality review may be used to change the data collection process intermittently or in the development of future surveys, particularly surveys that are longitudinal. This may allow for a reduction in the variability of the data and survey estimates. The changes in the data collection process may consist of changes in protocols, instruction, or training. That is, we may implement changes intended to improve clarity in procedures for the data collectors, which will reduce the variability introduced by the data collector. If these changes are implemented intermittently, then the sampling design, data collection mode, or questionnaires should not be changed as these changes would affect the data and analyses.

While literature does not generally focus on using sample designs for quality review, a probability sampling design offers several advantages. A probability sample design offers an objective and scientifically defensible method of selecting sample units for review. In addition, it provides the basis for estimating error rates or other features of the data collection process. Auxiliary, historical or even freshly collected data from a sample unit facilitates oversampling of units that are potentially at higher risk for errors. One may also want to oversample units during the beginning or final stages of the collection process or for specific individuals, such as new interviewers.

2.2 National Resources Inventory

In our analyses, we use quality review data from the 2005 NRI. The NRI has been conducted since 1982 by the United States Department of Agriculture Natural Resources Conservation Service in collaboration with Iowa State University. Data are currently collected annually and provide information on status and trends of natural resources on all non-Federal lands.

The annual NRI sample design is a stratified two-stage cluster sample of 70,000 area segments. There are about 40,000 segments in a core panel that are observed every year and 30,000 segments in a rotation panel that changes with each survey. Each segment contains between one and three points. We focus on the 2005 NRI data, which is the first survey after major reorganization and protocol change. The initial data collection for both the segments and their points is conducted using remote sensing, primarily aerial photograph interpretation and administrative records. The data collection takes place in three regional data collection units by contract data collectors. The area data are recorded digitally using specialized survey software to delineate and attribute area features. The point data involve classification of conditions at each sample point. Administrative data for the points are collected by state staff through a standardized questionnaire.

The 2005 NRI was the first survey in which specialized survey software was used to digitally delineate boundaries and calculate the area of features. In the past, the boundaries of area features were drawn on a mylar overlay and then planimeter was used to obtain the acreage of each feature. These areas were then manually entered into a handheld computer. Data collectors recording feature and point data needed to apply special rules that were sometimes difficult to implement. Many of these rules were automated by the new software. In addition to the introduction of new measurement tools, some new codes were also developed to better match conditions of interest to pastureland and rangeland scientists. For example, previously the pasture codes consisted of grass, legume, or grass-forbes-legumes mixed. In the new protocol, these were designated as either grassland or scrub shrub.

In the past, NRI data were collected at dozens of locations across the U.S. In 2004, NRCS reorganized data collection into three data collection units where contract employees conduct the data collection under the direction of NRCS staff members. For the 2005 NRI, only the core panel of segments was observed for 2003, 2004, and 2005 conditions. For those segments selected to be in the quality review sample, an NRCS technical and natural resource expert evaluated the data collected for the area features and point attributes of the segments. This review was conducted through the use of a standardized questionnaire that focused on the accuracy of responses recorded by data collectors for specific data elements.

3. 2005 NRI Quality Review Sample Design

A formal sampling scheme was implemented to provide objective methods for selecting area segments into the quality review sample. The sampling rate for the quality review sample was 3% of the area segments in the total 2005 NRI sample. Although past NRI surveys have involved in-stream sampling of segments collected by each data collector, these systems were not in place in time for the 2005 NRI. Thus, the quality review sample was selected before the start of 2005 data collection and provided to each data collection unit.

When designing the quality review sample, we considered many of the characteristics outlined by Morganstein and Marker (1997). This included identifying the primary factors affecting the data quality, identifying the critical data characteristics, and understanding the data collection process. The primary factors that might affect the quality of the data collection process were the data collection units. While attempts were made to standardize procedures, there were still inconsistencies that occurred among the three data collection units, which resulted in variability of the data characteristics. Thus we sampled 3% of the area segments in each data collection unit for review. The individual data collection units were comprised of a number of states that can vary in segment conditions. As a result, we also wanted to spread the quality review sample across the states. One of our goals with the quality review was to assess the quality of the data after major changes in protocols and measurement tools. Segments and points involved in these changes were considered to have a higher risk for data collection error. Segments having a lower risk for data collection error were those that were homogeneous or had points with conditions that were unlikely to change over time. We wanted to oversample those segments with higher risk of data collection error.

The goal was to review 3% of area segments for each data collection unit. There were 39,628 area segments in the 2005 NRI sample and thus, the sample size for quality review sample was set to 1,180 area segments. The sample size for each data collection unit was calculated by multiplying the number of area segments within the data collection unit by 0.03. Hence, 479 area segments were selected from data collection unit A, 472 area segments from data collection unit B, and 229 area segments from data collection unit C. Our analyses focus on variables collected from sample points. There are 3,435 points in the review sample, with 1,382, 1,399, and 654 in data collection unit A, B, and C, respectively.

In selecting the area segments to be included in the quality review sample, we wanted to oversample those segments that were more likely to present challenges. To accomplish this, each area segment was assigned to one of three disjoint categories based on risk factors related to the potential for data collection errors. Category 1 contained area segments believed to present the highest risk for data collection errors. Highest risk areas were those where the protocol was substantially changed (urban segments) and where there existed heterogeneity in the segments. Thus, category 1 segments contained urbanized areas or points with a mixture of land cover/uses. For category 3, we focused on area segments that presented low risk for data collection errors. These segments had points with conditions that were unlikely to change over the 2003-2005 period. This included segments for which all three points fell on the same stable land type (e.g., forest, large water body, rangeland). All of the remaining area segments were classified in category 2, which contained the segments that were considered to be of medium risk for data collection errors.

To oversample area segments at higher risk for errors, we assigned unequal probabilities to each category. Segments in the high risk category 1 were 5/2 times more likely to be selected into the sample as those in the low risk category 3, and segments in the medium risk category 2 were 3/2 times more likely to be selected than segments in category 3. Thus, the size measures for the three sampling categories were 0.5, 0.3, and 0.2 for high, medium, and low risk categories, respectively.

Since data collection units were comprised of states, we defined the states as strata within data collection units and sampling categories were strata within each state. Data collection unit k contained N_k area segments in the original 2005 NRI sample. The number of area segments to be chosen in each state was defined to be proportional to the sum of the size measures over all segments in the state. The sample size of state j in data collection unit k was calculated as

$$n_{jk} \quad = \quad \frac{(0.03)N_k \sum_{i=1}^{3} p_i N_{ijk}}{\sum_{i=1}^{3} p_i N_{i.k}}, \tag{1}$$

where p_i was the size measure of category i (i.e., 5/2, 3/2, 1), N_{ijk} was the number of segments in category i in state j in data collection unit k, and $N_{i.k}$ was the number of segments in category i in data collection unit k. We also calculated the sample size for category i in state j in data collection unit k, which was given as

$$n_{ijk} \quad = \quad \frac{p_i N_{ijk} n_{jk}}{\sum_{i=1}^{3} p_i N_{ijk}}. \tag{2}$$

Since both the sample size and total number of segments are known, we calculated the inclusion probability for segments in category i in state j in data collection unit k as

$$\pi_{ijk} \quad = \quad \frac{n_{ijk}}{N_{ijk}}. \tag{3}$$

After the sample size for each state was determined, the area segments were sorted by county within each sampling category. Then the segments were selected using a systematic probability proportional to size (PPS) algorithm within each state.

To create estimates from the quality review sample, weights were calculated. Since there was no nonresponse, the segment's weight is simply its inverse selection probability. A segment's weight was defined as

$$w_{ijk} \quad = \quad (\frac{n_{ijk}}{N_{ijk}})^{-1}, \tag{4}$$

where n_{ijk} and N_{ijk} were as defined above. The weight for all segments in a given risk category, state, and data collection unit were the same. Finally for variables collected from sample points, we needed a point level weight, which is equal to the point's segment weight divided by the number of points in the segment.

4. Methods

To illustrate how probability-based quality review data can be used, we analyze results from the survey's most critical variable, land cover/use, which is collected at each sample point for the years 2003, 2004, and 2005. Land cover/use refers to categories of land cover and land use. Land cover is the vegetation or other kind of material that covers the land surface. Land use is the purpose of human activity on the land. There are many different land cover/use codes, including codes for a wide array of crops, forest, rangeland types, developed land, water bodies and streams.

To evaluate the process quality for this variable, we considered misclassification errors. A point is considered misclassified if the data collector records an incorrect determination for the land cover/use. Truth was determined by reviewers, who utilized a standardized questionnaire. Reviewers indicate whether or not they agree with the data collector's determination, and if they disagree, they then provide what they believe to be the correct determination. Some questions apply only to specific types of land or features. In these cases, we begin our analyses by excluding those segments or points that do not belong to this domain. For example, if we are analyzing the rangeland response variable, then we only want to include those segments and points that are rangeland. The variables under consideration are coded for base year 2003 and change years 2003-2004 and 2004-2005. In our analyses, we estimate error rates for land cover/use determinations. We summarize our findings for 2003, 2004 and 2005 conditions and for each data collection unit. We also examine review data relating to codes for grassland, scrub shrub, forest land, barren land, and other rural land, in part because protocols were altered for these codes and thus they were at risk for higher error rates. To consider effectiveness of sampling categories, we compare error rates for sampling categories.

To evaluate error rates, we calculate an estimate of the percentage of relevant points that were misclassified and the standard error of the percentage. To illustrate how we calculate the estimate of the percentage of errors, we focus on estimating error rates for a data collection unit, which represents a stratum in our review sample design. For point m in segment l in sampling category i in state j in data collection unit k, let Y_{ijklm} be an indicator variable defined as

$$Y_{ijklm} \quad = \quad \begin{cases} 1 & if\ error\ exists\ in\ the\ land\ cover/use\ code\ in\ a\ given\ year\ for\ point\ m \\ 0 & otherwise \end{cases} . \qquad (5)$$

Then the misclassification percentage for the land cover/use code in the given year for data collection unit k is given by

$$\hat{p}_k \quad = \quad 100 * \frac{\sum_{ijlm \in k} w_{ijk} Y_{ijklm}}{\sum_{ijlm \in k} w_{ijk}}, \qquad (6)$$

where w_{ijk} was the segment's weight defined as above. We used a similar estimator for error rates by sampling category and for specific land cover/use codes. SAS PROC SURVEYMEANS was used to compute the estimates and their standard errors. Data collection units and sampling categories were defined as the strata and the point weights above were defined as the weights. For the land cover/use variables, we used domain estimation methods.

For the land cover/use variable, there is a small set of points for which the reviewer responds that they do not agree with the data collector's determination but then proceeds to give the same identification to the point. These points are all in data collection unit A and all have the same reviewer, with one exception. Also, this exception occurs in 2004 and 2005 for only one point, but for 14 points in 2003. We are interested in the estimated misclassification percentages of those points where the reviewer's and data collector's identification differ. Thus we exclude the 14 points for 2003 and the single point for 2004 and 2005 and then proceed in our calculations for the land cover/use variable.

In the tables presented below, rather than provide actual error rates, we provide estimates that are scaled relative to the largest estimated percentage in the table. Each estimate and standard error is divided by the value of the largest percentage, which results in the largest percentage being equal to 1. For example, if the actual estimated percentages are 1.85, 5.38, 2.94, and 1.48, then we report the values of 0.34, 1, 0.55, and 0.28. Although error rates are generally quite low, the differences in the small error percentages appear amplified in relative error comparisons. This scaling was implemented because the error rates from this survey have not been reviewed for release and our main goal is to illustrate the usefulness of our approach.

5. Results and Discussion

Our results focus on the misclassification rates of points for land cover/use for specific features by data collection unit, year, and sampling category. By examining these results, we can compare the estimated percentage of points misclassified across the three years and data collection units. We can also determine which land cover/use codes are more difficult for data collectors to identify. Additionally, we may look at the effectiveness of the sampling categories in identifying problematic segments and their points.

For the 2005 NRI, a data collector uses the determination for 2003 conditions from the old protocols used in the prior 2003 NRI survey to establish the code for the 2003 conditions under the new protocol. The data collector then proceeds to record codes for 2004 and 2005 conditions in the same session. Thus as we compare error rates by year, we expect the rates to be consistent. Indeed, there is no difference among years (see Table 1). An examination of individual points in error each year indicates that the same points are generally in error each year. That is, there is an initial error from the 2003 determination that gets propagated for subsequent years. Thus we will focus only on 2003 conditions in the subsequent analyses.

Table 1: Land cover/use (year) - relative error rates and standard error proportional to the largest estimated error rate.

Year	Est.	SE
2003	0.94	0.10
2004	0.95	0.10
2005	1.00	0.11

Next we consider the estimated percentage of points with any misclassified land cover/use for each data collection unit (Table 2). Because the 2005 NRI was the first survey conducted by contract staff in each of the three data collection units, it is important to compare quality outcomes for the data collection units to evaluate whether discrepancies exist in implementing the survey. We find that data collection unit C has the highest estimated error rate. This trend holds for other point response variables as well. These higher error rates may be partly due to the land composition in data collection unit C, where it can be more difficult to distinguish between land cover/use codes. It is also possible that reviewers in data collection unit C differed from other data collection units in how they reviewed these land cover/uses. The estimated error rate for data collection unit C appears to be much greater than the other data collection units using the relative error rates reported here. This is due to the small differences in the actual number of errors, which are then amplified in relative error comparisons.

Table 2: Land cover/use (Data collection unit) - relative error rates and standard error proportional to the largest estimated error rate for 2003 conditions.

Data collection unit	Est.	SE
A	0.23	0.05
B	0.25	0.06
C	1.00	0.16

We also examined the estimated percentage of points with misclassified land cover/use for selected codes. The results for this analysis are summarized in Table 3 below. We find that grassland, scrub shrub, and forest land have higher misclassification rates than barren land and other rural land as well as all other kinds of land cover/use. Under the new protocols established in 2005, the codes for grassland, scrub shrub, and forest land were new to the survey. Thus, even though these were core panel segments and the prior year's determinations were available, no historical data were available for these land cover uses. This may contribute to the higher error rates. Another contributing factor is that it is rather difficult to distinguish between grassland, scrub shrub, and forest land. For example, points misclassified as grassland were usually more properly classified as scrub shrub or forest land. The land cover use codes for grassland and scrub shrub are used to define whether or not a point is rangeland and grazed land, two other point response variables. Thus, errors in grassland and scrub shrub codes may affect estimates associated with other classifications of the land.

Table 3: Land cover/use (kind of feature) - relative error rates and standard error proportional to the largest estimated error rate for 2003 conditions.

Kind of feature	Est.	SE
Grassland	1.00	0.25
Scrub shrub	0.86	0.21
Forest Land	0.66	0.18
Barren land	0.14	0.09
Other rural land	0.07	0.05

Returning to error rates for all land cover/use codes, we are interested in the effectiveness of the sampling categories in classifying risk levels of segments and hence points. Table 4 provides the estimated percentage of points with misclassified land cover/use in 2003 by sampling categories. We find that segments in the first category do have a higher error rate than segments in the third sampling category. This is as expected, which indicated that our use of sampling categories and oversampling those in the first category did identify more segments that were more likely to be in error.

Table 4: Land cover/use (sampling category) - relative error rates and standard error proportional to the largest estimated error rate for 2003 conditions.

	Est.	SE
Category 1	1.00	0.14
Category 2	0.80	0.13
Category 3	0.38	0.17

6. Conclusion

We used quality review data from the 2005 NRI to illustrate the value of using a probability-based review sample to monitor the quality of data collection in a survey, particularly after a major change in protocols. The probability design we used enabled us to spread the review throughout all counties in the survey. In addition, we were able to successfully target sample units that might be at higher risk for errors. We used a method that involved defining strata based on historical or auxiliary data correlated with potential error rates, and then setting different sampling rates for the strata to control spread of the sample across risk categories.

These review data were monitored throughout the data collection process for unusual levels of errors. Although the vast majority of variables observed by data collection units were reviewed, monitoring focused on variables expected to be problematic or that were critical to estimation. In this paper, we focused on the post-hoc analysis of review data, where it is possible to consider different kinds of questions such as whether the stratification was effective or what the error rates were for rarer land cover/uses that require review data from the entire process. We found that for the land cover/use codes that were affected by changes in protocols, higher misclassification rates occurred. This underscores the importance in evaluating the data quality after changes in protocols and measurement systems.

The 2005 NRI quality review sample design did not guarantee spread across data collectors. A design that addresses this goal in the context of continuous monitoring must rely on in-stream sampling. In an in-stream process, segments completed by a data collector are sampled as they are completed. Although a real-time sampling approach was very effective in the 2003 NRI, systems were not ready for the 2005 NRI and a pre-selected sample was used in its place. Considerations for in-stream quality review sampling are discussed by Ferraz (2004). They note that stratification over time can be used to increase inspection rates early and late in the process, when errors are more likely. In this setting, one can also use the freshly collected data to establish a size measure related to the error risk and give higher likelihood to sampling segments with higher risk. This approach supports the kinds of analyses that were conducted for the 2005 NRI to evaluate data collection process quality during and after the 2005 NRI.

References

[1] Biemer, P. & Caspar, R. (1994). Continuous Quality Improvement for Survey Operations: Some General Principles and Application. *Journal of Official Statistics*, **10**, 307-326.

[2] Biemer, P. & Lyberg, L. (2003). *Introduction to Survey Quality.* New Jersey: John Wiley & Sons, Inc.

[3] Eurostat. (2000). Assessment of the Quality in Statistics. Eurostat/A4/Quality/00/General/Standard report. Luxembourg.

[4] Ferraz, C., Nusser, S.M. & Opsomer, J.D. (2004). Sample designs for monitoring the data collection process of longitudinal surveys. In JSM Proceedings, Survey Research Methods Section. Denver, CO: American Statistical Association.

[5] Groves, R.M. & Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. textitJournal of the Royal Statistical Society, **169**, 439-457.

[6] Lessard, V. (2005). 2003 NRI Quality Control Data. White paper. Center for Survey Statistics and Methodology, Iowa State University. June 2, 2005.

[7] Lyberg, L. & Elvers, E. (2003). Survey Quality Issues During The Last 50 Years. Proceedings of the Section on Survey Research Methods. Alexandria, American Statistical Association.

[8] Morganstein, D.R. & Marker, D.A. (1997). Continuous quality improvement in statistical agencies. In Lyberg et al. (Eds.), Survey Measurement and Process Quality (pp. 475-500). New York: John Wiley & Sons.

[9] Nusser, S.M. (2003). Quality control (QC) sample algorithm for 2003 NRI data collection. White paper. Center for Survey Statistics and Methodology, Iowa State University. April 11, 2003.