DATA EDIT ERROR ANALYSIS

David Megill and Sandra Rowland
Bureau of the Census [1]

1. Introduction

Although certain types of response errors cannot be detected without an expensive reinterview or validation program, it is possible to obtain an indication of the general quality of the data by examining the output of the edit. A high rate of edit changes for a particular questionnaire cell would indicate that the quality of the incoming data for that cell is poor. The examination of the frequency and types of errors identified during the edit provides valuable information concerning questionnaire design, training of interviewers, field supervision, coding, editing, and objective criteria for the comparison of error types and error rates among surveys.

This paper will examine the procedures used and the results obtained in data edit error analyses of rural household surveys taken in the Dominican Republic in 1976 and in Peru in 1984. The Dominican Republic Cost of Production Survey collected data on land area, crop production, animal inventories, processed products, employment patterns, farm income and family income from 1800 farm households. The Peru Rural Household Survey collected similar data from 6069 farm and nonfarm rural households.

The experience with data edit error analysis from these surveys has shown that it is an economical and objective measure of certain kinds of errors in survey data including inconsistent entries, invalid (impossible) entries and entries outside established limits of tolerance and item nonresponse. The information required for the analysis is readily available from computer tapes and error correction sheets generated during the hand and/or computer edit. Data edit error analysis is a very useful addition to the concept of nonsampling error analysis and is recommended for use in the control and evaluation of survey data.

2. Dominican Republic Cost of Production Survey

2.1 Definitions of Measures Used in the Analysis

Error Counts by Cell

The counts of changes by cell for all questionnaires were computed by comparing the original unedited data tape with the final edited tape. If a cell was changed one or more times during the edit, it was counted as one error. The error count thus gives the number of errors by questionnaire and allows for the computation of the number of errors by cell for all questionnaires. The reader should understand that the error count is a tally of changed cells, without regard to the reasons for the changes or to specific sources of errors. The error count reflected that, due to any number of unspecified problems throughout the survey, some response values were considered to be invalid and had to be changed. This fact should be kept in mind throughout the following discussion.

Base Counts

The base counts are the number of cells that should have non-zero response in the questionnaires. The computation of the base is very important because it tells us the actual length of the questionnaire in terms of response rather than the length of the questionnaire in terms of the total number of cells in the questionnaire. Although the questionnaire contained 1275 cells, the average number of cells in the base was only 206 (valid non-zero responses) per questionnaire.

Error Rate

The error rate is the ratio of the error count to the base. The mean error rate for all cells in the 1670 questionnaires was 6.05 percent. We shall keep this figure in mind as we proceed to examine the relationships between error rates and other factors.

Imputation Rate

During the edit, the editors were unable to correct unreasonable responses under two conditions: (1) if the value violated the maximum or minimum allowable limits or (2) if the correct answer could not be determined through the evaluation of other responses in the questionnaire. These uncorrectable values were marked "not available" by a field of nines. Later, a few of the fields that had violated the reasonable limits were replaced with the original values, but most of the nines fields were replaced with national averages, that is, imputations calculated from the reasonable responses to the questions.

2.2 Analysis of Error Rates by Section of the Questionnaire

Table 1 represents the distribution of questionnaires by error rate based upon the computation of the ratio for each questionnaire. It indicates that 66 percent of the questionnaires had an error rate of six percent or less. The average error rate (ratio of the means) for all questionnaires was 6.05 percent.

Error Rates for Sections of the Questionnaire

The questionnaire has eight sections:

SectionI:	The Producer and Location of the Farm
Section II:	Area, Land Utilization, Tenancy, and Type of Land
Section III:	Crops Cultivated on the Farm During the Last 12 Months
Section IV:	Inputs and Production of Crops
Section V:	Animal Inventory
Section V:	Animal Inputs
Section VI:	Processed Products
Section VII:	Net Worth and Administration
Section VIII:	Non-Farm Income

The error rate fluctuated a great deal from one section to another and did not show a gradual increase from Section I to Section VIII of the questionnaire. The questionnaires were administered from beginning to end without skipping from one section to another and back; therefore, a gradual increase in the error rate from the beginning to the end of the questionnaire would have implied fatigue on the part of the interviewer or the respondent. Fluctuations, on the other hand, imply that certain sections of the questionnaire were more difficult to administer or to respond to. For example, the section on eggs had a 14 percent error rate while the section on draft animals which follows it had an

error rate of only one percent. Furthermore, the extremely high error rate for Section VI, or Processed Products, (17 percent), was followed by 3 and 4 percent error rates in Sections VII and VIII, respectively.

However, in both Section IV with five repeating subsections on crop production and inputs and Section V with three repeating subsections on animal inputs, the error rate increased from the first to the last subsection implying that detailed information on crops and livestock becomes more difficult to obtain as the number of specific crops and livestock to be dealt with increases.

2.3 Questions with High Error Rates

Those questions which contributed to the high error rates in certain sections of the questionnaire, particularly questions with an error rate over 15 percent, were carefully evaluated. However, the analysis of error rates by question is not included in this paper. An example of such analysis is given in Section 3 for The Peru Survey.

2.4 Error Rates for Crops

Of the 47 crops examined, 18 had error rates over six percent and seven had error rates over 10 percent.

The permanent crops had an error rate almost twice as high as that for temporary crops due to the high error rate of questions on number of trees and area in production, and to the problems with coffee and cacao. Other permanent crops such as oranges and grapefruit had rather high rates due to the conversion of their production to units rather than pounds.

Of the temporary crops, only garlic, onions, yams, yautia, eggplant, and sesame had error rates over six percent. They also had base counts well below average and may therefore be considered to be rare crops. The high error rate for these crops was probably due to their novelty.

2.5 Error Rates for Animal Inputs and Processed Products

The error rate for animal inputs was high for pork, draft animals, beef cattle, and "other animals." Of these, only pork was quite common in the survey. Its high error rates were due primarily to the overestimation of feed costs which was due, in turn, to the monthly time frame used for animal inputs.

The high rates for other animals were probably due to the rareness of their occurrence in this section. It is also clear that the rates are quite high for most processed products. In fact, 11 out of 13 processed products have error rates of over 10 percent.

2.6 Error Rate by Base Count and by Interview Length

This section discusses the effect on error rates of questionnaire duration (base count) and interview length. The average questionnaire length was 206 base cells, the average interview length was 78 minutes. Both distributions were positively skewed, with the majority of cases below the mean values. In addition, the correlation coefficient between questionnaire length and interview length was 0.64, which indicates a strong positive linear relationship for 1647 cases. The error count totals, the base count totals, and the error rates were calculated for eight questionnaire-length groupings and nine interview-length groupings. The error rates were relatively close to the overall error rate mean. In other words, the error rate remained fairly constant, regardless of how many responses were made or how much time was devoted to the interview. This suggests that questionnaire length and interview duration may have little effect on the error rate.

2.7 Error Rate by Interview Date

Most of the interviews were administered during 4 weeks, from the second week in March through the first week in April 1976. The interviewers administered an average of 464 questionnaires each week for the first 3 weeks before their productivity slackened the final weeks as they completed their assigned segments.

Nationally, the error rate declined steadily from the first week of interviewing (6.6 percent) through the fourth week (5.4 percent) before climbing to 6.2 percent during the last week, when a few teams hurried to finish their interviews. This decreasing trend is also evident for teams. Seven out of ten teams have decreasing rates, accounting for the nearly 20 percent decline in the national rate from the first to the fourth week. This implies that interviewer morale did not wane as the survey progressed. The experience from their first interviews became personalized extensions of their interviewer training, preparing them to confront a variety of situations in the field. More importantly, as the interviews were submitted, some errors were reviewed and the results were immediately returned to the teams in the field. This immediate feedback was a major factor in improving team performance in the field over time.

2.8 Error Rates by Farm Size, Number of Crops, and Type of Respondent

The average base count per questionnaire was 206. Previously we noted that while the error count was correlated with the base count, the error rate was not. This is reflected once again as we review the error counts and error rates by farm size. The average error rate did not differ by farm size, in spite of the fact that the average base count for small farms was 181 while that of medium and large farms was around 235.

The average base count per questionnaire also rose as the number of crops recorded in Section III of the questionnaire rose. For example, questionnaires with five crops or less had an average base count of 181 while those with 11 crops or more had an average base count over 338. This time, however, there was a perceptible increase in the error rate for questionnaires with more than 11 crops. It increased from 6 percent for questionnaires with 10 crops or less, to eight percent for those with 11 to 15 crops.

Questionnaires for which the respondent was the farm manager (rather than the producer, the wife or son of the producer, or any other type of respondent), had an error rate of 9 percent. This may be due to the small number of observations in that category; to the reluctance of farm managers to divulge information; or to their unfamiliarity with actual sales and purchases.

2.9 Error Rates by Interview Team and Interviewer

The quality of the information gathered in a survey is highly dependent upon the quality of the interviewers administering the survey. The

practice of evaluating error rates by interviewer is especially worthwhile for projects which require a permanent staff of interviewers. It provides the most objective criterion for judging the quality of field supervision and enumeration and, therefore, helps direct training efforts to those who need it most.

There was a one-to-one relationship between workzones and interviewing teams. This allowed us to examine error rates by workzone as a proxy for interviewing teams and supervisors. Its major drawback was that the workzones are also geographically based so that the effect of geographical differences on error rates cannot be controlled.

We concluded that the interviewer teams for workzones A, D, H, I, and J did a comparatively good job (error rates below average), those in workzones B, E, and G did an average job (error rates near average), and those in workzones F and C were the target teams to which future training programs should pay particular attention (error rates above average).

It appears that the workzones with the lowest error rates (A, B, D, and H) had good supervision because all of the interviewers had error rates close to or less than the national average (6.02 percent). It should be noted that the quality of interviewer's work is also affected by factors other than supervision, such as workload and cooperation of the respondent.

2.10 Analysis of Errors by Correction Class

The source of information used in this Section was the output from the hand edit which contained the number of changes made during the hand edit by correction class and edit cycle. Each change made by an editor was accompanied by the correction class which, in the judgment of the editor, best described the cause of the error. The number of changes made by the editors was higher than the error count referred to previously because it was possible for a cell to be changed more than once during the edit.

The types of errors which were edited fall within the categories defined below:

a. Keypunch: errors caused by inaccurate transfer of the data from the questionnaires to the computer data files.
b. Coding: any error caused by the use of the incorrect codes for crops, livestock, processed products, etc.
c. Missing: any error caused by invalid cell values of zero.
d. Sum: any error caused by incorrect summation or disaggregation.
e. Conversion: any error caused by incorrect conversion of production units to standardized units.
f. Incomplete conversion: any error caused by the inability to convert production units to standardized units due to lack of conversion rates.
g. Consistency: any error caused by inconsistent information between questions and sections of the questionnaire.
h. Enumeration: any error caused by the interviewer's failure to correctly administer the questionnaire.
i. Max-min: any error caused by unreasonable cell values.

Enumeration Errors

Table 2 shows the number of errors by correction class which were edited during the entire hand edit. We see immediately that almost half of them were, in the opinion of the editors, enumeration errors. This is understandable given that all errors due to the incorrect enumeration of processed coffee and cacao were included in this category. The "enumeration" correction class was also used when errors could not be classified in any of the other correction classes. It, therefore, served the purpose of a "catch-all" or "other" category. This is unfortunate because much information concerning the causes of errors was lost in this category. More categories could have been created when the need for them was felt during the edit but the correction transcription sheet and the master correction computer program were designed for only one-digit correction classes. Therefore, the "enumeration" correction class could include any error caused by reasons not defined in the other eight correction classes.

Missing Data, Inconsistencies, and Max-Min Violations

Referring to Table 2 again we find that 36.5 percent of the errors were placed in the following correction classes: missing, consistency, and max-min. Errors due to inconsistent information between questions and sections of the questionnaire are the smallest error category of this kind (5.9 percent) and violations of maximum and minimum limits are second with 7.1 percent of the errors. The largest category of recall errors is, therefore, missing data which accounted for 23.5 percent of all errors dealt with in the edit.

Summing, Keypunch, Coding, and Conversion Errors

Fourteen percent of the errors detected by the edit were due to incorrect summing, keypunching, coding, and conversion. While summing errors were responsible for 2.3 percent of the total errors checked, each of the other causes was responsible for approximately four percent.

3. **Peru National Rural Household Survey**

The editing and imputation procedures for the Peru National Rural Household Survey provided for changing the value in a questionnaire cell whenever an out-of-range or inconsistent entry was found. Unfortunately, the type of error was not recorded as part of the editing procedures, so it is not possible to know whether most of the error was due to problems with the data collection, office coding or data entry. In the case of particular cells with a high error rate, the nature of the problem was investigated further. The edit error analysis for the Peru survey was facilitated by the use of CONCOR, a generalized editing package which automatically produced a diary of all errors identified and corrected during the edit.

There were basically three ways the entry for a particular questionnaire cell would be changed when it was found to be invalid. One way was for the office editors to fill out an error sheet specifying the valid value if it could be determined from the questionnaire. In cases where the valid value was unknown, the cells were flagged with a "$" during the edit phase and were later imputed using a "hot-deck" or "cold-deck" method.

The third type of change was an automatic imputation during the CONCOR consistency edit, carried out mostly for screening questions (with yes/no answers).

3.1 Methodology

For the purposes of the error analysis, the error rate for a particular questionnaire cell was defined as the total number of errors for that cell divided by the number of cases where the cell had a value greater than zero (i.e., the number of applicable cases, or base). It should be noted that the errors measured in this study are only a detectable subset of the overall population of errors and does not include response errors which fell within acceptable ranges. Therefore, the error rates reported here should be considered minimum values for the overall relative nonsampling error.

The number of errors for each cell was obtained from the summary data from all the error sheets (i.e., manual corrections) at the end of the edit plus a summary of all the automatic changes carried out by the CONCOR runs. There were cases where some individual cells were changed more than once during the edit, in which case they would be double-counted in the total number of errors, thus increasing the error rate. This estimation procedure was used in the error analysis mainly because of the nature of the computer edit and the edit diaries generated. However, considering that multiple changes to a cell generally warrant less confidence in the quality of the corresponding data, it appears reasonable to include all such changes in the error rate. On the other hand, it is sometimes also of interest to examine the error rate based on simply counting each cell once regardless of how many times it was changed. In this case the error rates could be determined by comparing the original unedited data tape to the final edited version to determine the number of cells changed, as in the case of the Dominican Republic Cost of Production Survey.

3.2 Results of Error Analysis

The overall error rate for the questionnaire cells was 6.1 percent, which appears quite reasonable given the nature of the survey. This compares to an overall error rate of 6.05 percent for the Dominican Survey. Considering that the latter rate was based on counting each cell with a correction only once, regardless of how many times it was changed, the comparable rate for the Peru survey was probably even smaller. Therefore, the overall quality of the survey data appears to be quite good, although the error rates vary considerably by questionnaire cell, and caution is certainly warranted when using data from cells with an error rate of 20 percent or higher. Table 3 shows the frequency distribution of the 404 unique questionnaire cells by error rate. It can be seen in this table that over 50 percent of the questionnaire cells have an error rate of less than 5 percent, and only 17.1 percent have an error rate of 20 percent or higher.

3.3 Questionnaire Cells with High Error Rates

In order to examine more closely the more problematic variables, the 69 questionnaire cells with an error rate of 20 percent or higher, with the corresponding number of automatic corrections, total number of errors, applicable number of observations (base) and the overall error rate for each, were closely examined. The results of the edit were reviewed in order to understand the nature of the major problems related to these variables. The following observations were made concerning the questionnaire cells with error rates of 20 percent or more:

1. Most of the cells with a high error rate have a relatively small base (as low as 1 or 2 observations). This is the case with cells related to the purchase or sale of land. In some cases the number of errors is actually larger than the base because of multiple changes to individual cells (in such cases the error rate is shown as 100 percent). Given the low number of observations for land purchases and sales and the high error rate, no reliable estimates can be expected for these variables even at the national level.

2. The question that inquires about the household's percentage share in the profits (losses) from the farm operations held in partnership had a high error rate. It appears that the respondents did not understand the question very well, and the interviewers did not probe sufficiently. Therefore, the data for this cell are questionable, although given the small number of cases involved, it should not affect the overall quality of the income data.

3. The quality of the data related to sales of crops appears to have suffered due to a misunderstanding on the part of the respondents and a lack of interviewer care in probing. The problem may also be partly related to the questionnaire design for this section, as the space assigned for the answers to these questions had been divided into two, in order to indicate whether the sale was on a cash or credit basis. These data should therefore be used with caution. However, the number of credit sales was relatively small (about seven percent of the total crop sales), so this should only have a minor affect on the overall quality of the crop sales data.

4. In the case of the cell corresponding to the question on the month of planting, the most likely source of error is a recall problem on the part of the respondents, since the planting could have occurred up to 18 months before the interview date. Because of the large amount of imputation, this data should be used with caution.

5. The very large number of changes for the cell on the type of seed used resulted from the large number of respondents who answered that no seeds were used, when in fact it was necessary for some type of seeds to be used. This indicates a lack of probing on the part of the interviewer.

6. The large number of imputations of blanks for the cell which screened for cows and goats milked indicates that many interviewers did not follow the skip pattern, although the answers were correct and the quality of the data was not directly affected.

7. There were many inconsistencies in the information on the number of weeks cheese and butter were produced, resulting in a large number of changes. Apparently the interviewers did not probe sufficiently to resolve such inconsistencies. In the case of the cell corresponding to expenses related to the production of cheese and butter, many respondents apparently could not

account for such expenses, resulting in a high level of imputation.

8. The three cells related to rent paid for land had few observations (197) and relatively high error rates, (33.5 percent, 27.9 percent and 46.2 percent, respectively). Therefore, the corresponding data do not appear to be very reliable.

9. For the cell corresponding to expenses related to processed products, there were relatively few observations (227) and the error rate was 26.9 percent, indicating that this data should be used with caution. Apparently the high level of imputation was due to the fact that many farmers do not systematically account for their expenses when they process agricultural products.

10. The data in the section on technical assistance also appear to be of poor quality given the low number of observations and the high error rates. The cells corresponding to reasons for not applying recommendations should definitely not be used for any tabulations or analysis, since there were only nine observations for all these cells combined, and most of these had been changed.

11. In the case of the cell on vehicle rental expenses for transporting crops, there were 187 imputations due to inconsistencies, indicating a deficiency in interviewer probing.

12.The main reason for the high error rates for cells on use of other agricultural machinery or implements was that if there was an indication in another cell that insecticide was used for a crop, they were imputed to indicate that an owned or borrowed insecticide applicator was used.

13. The section on farm household members belonging to an agricultural association suffers from a small number of observations and many changes, especially the questions regarding the reason for leaving the association. Therefore, much caution should be used in analyzing this data even at the national level.

14. The cell corresponding to income from various other sources only has 64 observations, with a total of 15 changes, so the inference from these data is very limited. However, this would not have much effect on the overall income data. In the case of the cell on income in kind, the problem is more serious, as there is a larger base (520) and a very high error rate (79.4 percent). Apparently, it is difficult for the interviewers to obtain good information on income in kind, as the respondents in many cases do not know the value of the trade. This could result in a significant bias on the income of smaller subsistence farmers who rely more on trade. However, given the relatively small value of income in kind, this should not have a very serious effect on the overall average farm household income.

The general conclusion from this error analysis was that the overall quality of the data appears to be quite good, with an error rate of only 6.1 percent across all cells. However, data for individual cells with a high error rate, such as those discussed above, should be used with caution; most of these cells have a relatively small number of observations, also limiting the inferences which could be made from these data.

REFERENCES

Control and Evaluation of Data Quality Methodological Working Document No. 3, U.S. Agency for International Development, Bureau for Latin America, Washington, D.C., June 1978.

Documento Methodologico, Encuesta Nacional de Hogares Rurales, Instituto Nacional de Estadistica, Ministerio de Agriculture, Lima, March 1986.

[1] This paper reports the general results of the research undertaken by Census Bureau staff. The views expressed are attributable to the author(s) and do not necessarily reflect those of the Census Bureau.

Table 1: Frequency of Questionnaires by Error Rate

Error Rate	Relative Absolute Frequency	Cumulative Frequency (Percent)	Frequency (Percent)
0	30	1.8	1.8
1- 2	323	19.3	21.1
3- 4	452	27.1	48.2
5- 6	292	17.5	65.7
7- 8	217	13.0	78.7
9- 10	118	7.1	85.7
11- 15	148	8.9	94.6
16- 30	87	5.2	99.8
31- 100	3	0.2	100.0

Variable Rate

Mean	6.050
Variance	24.339

Table 2: Number of Errors by Correction Class

Correction Class	Number of Errors	Percent of Total
Keypunch	869	3.8
Coding	823	3.6
Missing	5,426	23.5
Sum	525	2.3
Conversion	243	1.2
Incomplete Conversion	734	3.2
Consistency	1,368	5.9
Enumeration	11,415	49.5
Max-Min	1,645	7.1
TOTAL	23,057	100.0

Table 3: Frequency Distribution of Questionnaire Cells by Error Rate

	Total	Error Rate (Percentage)							
		0-1.9	2.0-4.9	5.0-9.9	10.0-14.9	15.0-19.9	20.0-29.9	30.0-49.9	50.0+
Number of Cells	404	112	96	85	26	16	17	9	43
Percentage of Cells	100	27.7	23.8	21.0	6.4	4.0	4.2	2.2	10.6