United States
Department of
Agriculture

Economic
Research
Service

**Agriculture
and Rural
Economy
Division**

Correcting for Nonresponse in Transition Matrices Calculated from Longitudinal Data

R. Neal Peterson
Fred Gale

It's Easy To Order Another Copy!

Just dial 1-800-999-6779. Toll free (in the United States and Canada). All other areas please dial 301-725-7937.

Ask for *Correcting for Nonresponse in Transition Matrices Calculated from Longitudinal Data* (AGES 9113).

The cost is $4.00 per copy. For non-U.S. addresses (including Canada), add 25 percent. Charge your purchase to your VISA or MasterCard, or we can bill you. Or send a check or purchase order (made payable to ERS-NASS) to:

ERS-NASS
P.O. Box 1608
Rockville, MD 20849-1608.

We'll fill your order by first-class mail.

Correcting for Nonresponse in Transition Matrices Calculated from Longitudinal Data. R. Neal Peterson and Fred Gale, Agriculture and Rural Economy Division, Economic Research Service, U.S. Department of Agriculture. Staff Report No. AGES 9113.

Abstract

Longitudinal data suffer from the same statistical problems as cross-sectional data. True estimates of means and sums require adjustments for sampling rates or, in the case of census data, nonresponse rates. Whereas in cross-sectional data the standard method of adjustment is the attachment of weights to individual observations, this method does not work in the case of transition matrices calculated from longitudinal data. The reason for this is the unavoidable misclassification of farms as exiters, entrants, and continuers that arises from nonresponse. The method presented here offers a simple algorithm based on four assumptions for making the necessary correction. The algorithm is easily implemented with standard spreadsheet software.

Keywords: Transition matrices, longitudinal data, nonresponse adjustment.

1301 New York Avenue, NW
Washington, DC 20005-4788

March 1991

Contents

Correcting for Nonresponse in Transition Matrices Calculated from Longitudinal Data

R. Neal Peterson
Fred Gale

The Problem

Unadjusted statistics from U.S. Census of Agriculture cross-sectional data do not yield true estimates of population means and totals. The main source of error is nonresponse. Every census year questionnaires are mailed to all identifiable agricultural operations, some of whom answer and some of whom do not. Overall, in any given year approximately 10 percent of operators fail to respond, with the percentage being roughly inverse to the size of the operation. The self-selection process in the decision by operators either to respond or not respond to the questionnaire both reduces the statistical accuracy of estimates and introduces bias. For these reasons the Census Bureau stratifies population of farms according to size, calculates nonresponse rates for each stratum, and uses these rates to make statistical adjustment. Thus, reported totals are weighted sums and reported means are weighted averages. (For a discussion of the method and impact of weighting for nonresponse in the census see 1987 Census of Agriculture, Volume I, Appendix C, Statistical Methodology, U.S. Dept. Comm., Bur. of Census, Nov. 1989.)

The Bureau of the Census has developed a U.S. Census of Agriculture longitudinal file that can be used by economists and other analysts to study the process of structural change in the farm sector. The file links consecutive census years 1978, 1982, and 1987 by matching census records that possess the same Census File Number, which is a unique identifier that tends to be preserved across census years as long as the same operation remains under management of the same operator. In instances when an operation is found not to exist in a particular year, it is nonetheless carried in the file with zero values for that year's variables. Thus, the longitudinal file enables analysis of change in individual operations from the time of a farm's entry, across the years of its continued operation, until its exit.

Census longitudinal data suffer from the same problem of nonresponse as cross-sectional data since the longitudinal data consist of census records that have been matched across different years. This paper reports a method that logically corrects for nonresponse in transition matrices that are calculated from longitudinal data. Although the need for this adjustment arose in the course of analyzing Agricultural Census data, its applications are not restricted to that data. Transition matrices calculated from longitudinal data of any sort that exhibit variable response rates or sampling rates among strata can be corrected using the adjustment method below.

The Transition Matrix

A transition matrix is simply a cross-tabulation of the subject population into a set of classes in two different periods. Each c_{ij} entry in the matrix contains the number of operators reporting in class j in the second census who reported in class i in the first census. Logically, the universe of the names and addresses to whom the Census Bureau mails out its questionnaires is decomposable into three "response" categories: "absent" (nonexistent farms or places misidentified as being farms), "no response" (existing farms that fail to respond), and "response" (existing farms that respond). These three categories partition the transition matrix into nine submatrices (fig. 1).

The fact that some operations are missing from the longitudinal data as a result of nonresponse means that some observations in the initial transition matrix will be misclassified. Farms in submatrix G, classified as "exiting" farms because they responded to census I but not to census II, are in fact continuing farms. Similarly, farms in submatrix E, classified as "entry" because they responded to Census II but not to Census I, are also continuing farms.

In addition to misclassified farms are farms that were missed in both censuses because of operators' failures to respond in both years (submatrix D). These farms do not enter the totals of any cell of the transition matrix. When nonresponse is treated the same as absence, the definition of exiters, entrants, continuers, and potential farmers is misspecified, as shown in figure 2, and the transition matrix is incorrect. The transition matrix is correctly specified when response and nonresponse categories are grouped together in the definition of exiters, entrants, continuers, and potential farmers (fig. 3).

Assumptions

A procedure for correcting for nonresponse must make some assumptions about the act of responding/not responding by farm operators to the census questionnaires. We chose the following four assumptions.

Figure 1. Submatrices formed by the three response categories.

Census I \ Census II	Absent	No response	Response
Absent	0	A	B
No response	C	D	E
Response	F	G	H

Figure 2. The incorrectly specified transition matrix.

Potential Farmers	Entrants
Exiters	Continuers

Figure 3. The correctly specified transition matrix.

Potential Farmers	Entrants
Exiters	Continuers

1. For all size classes, the probability of an operator failing to respond to the census questionnaire is accurately estimated by the observed rate of nonresponse.

2. For all size classes, the probabilities of nonresponse between the different censuses are independent. That is, the probability of nonresponse by operator X in the first census is unrelated to operator X's probability of nonresponse in the subsequent census.

3. For all size classes, the probability of nonresponse by farms that later exit is no different from the probability of nonresponse by farms that continue in operation.

4. For all size classes, the probability of nonresponse by farms that have entered is no different from the probability of nonresponse by farms that continued in operation.

Realistically, assumptions 2, 3, and 4 cannot be completely true. For instance, operators who deliberately choose not to answer a questionnaire in one census year would presumably be less disposed to answer later census questionnaires than operators who did respond. Similarly, operators who expect to quit farming soon may be less inclined to answer a questionnaire than farmers who intend to continue farming. Without better information, however, these assumptions seem the best that can be made.

The Procedure: An Example

An algorithm for correcting the unadjusted transition matrix for nonresponse must estimate each of the submatrices A through H (fig. 1) from information that is contained in the published census volumes, the longitudinal file, and the four assumptions above. For the sake of clarity, the explanation that follows employs an example to illustrate the method (tables 1 and 2).

Table 1--Example: The uncorrected transition matrix

Item	Census II farms			Row totals
	Missing	Class 1	Class 2	
		Number		
Census I farms:				
Missing	0	21,164	3,136	24,300
Class 1	26,708	47,411	3,322	77,441
Class 2	3,311	2,099	10,905	16,315
Column totals	30,019	70,674	17,363	118,056

Table 2--Example: Calculation of nonresponse rates

Item	Census I		Census II	
	Class 1	Class 2	Class 1	Class 2
			Number	
Published volume totals	87,432	17,258	80,520	17,963
Longitudinal file (responses)	77,441	16,315	70,674	17,363
Nonresponses	9,991	943	9,846	600
Nonresponse/response ratio	0.1290	0.0578	0.1393	0.0346

In our example, total farms declined from 93,756 in the first census to 88,037 by the time of the second census, and farms are sorted into two size classes. Missing farms are the sum of absent farms and nonresponding farms. The purpose of the algorithm is to discover what portion of the missing farms were nonresponses and what portion were absent, and to allocate the nonresponses into the two different size classes. Although for simplicity this example uses only two size classes, the procedure generalizes to any number of classes as the reader may verify while examining the procedure's six steps. Rounding was to the nearest integer.

Step 1: Continuing Farms that Responded in Both Censuses

To begin, there is a set of entries that may be inserted directly into the expanded transition matrix without alteration. These are the farms that responded to the census in both years, matrix H. It consists of four entries which are the four different combinations of farm classes in the two years. Thus, at step 1 the expanded transition matrix is as follows:

O	A	B
C	D	E
F	G	47,411 3,322 2,099 10,905

Step 2: Continuing Farms That Responded in Only One Census

The next two steps fill in the remaining portions of the continuing farm population. By assumption that nonresponse rates from different censuses are independent, the continuing farms that failed to respond in one census (matrices E and G) may be easily calculated from the continuers that responded both times (matrix H). Beginning with G (the matrix of continuing farms that responded in census I but not in census II): The class-1 farms that remained class-1 farms (the upper left entry of G) equals 13.93 percent of 47,411, since 0.1393 is the census-II class-1 ratio of nonresponse to response. Likewise, the upper right entry of G equals 13.93 percent of 3,322. The lower left entry equals 3.46 percent of 2,099. And the lower right entry equals 3.46 percent of 10,905. In general then, using matrix notation, the matrix G is the product of the responding continuers matrix H and the census-II nonresponse ratios:

$$\begin{bmatrix} 47{,}411 & 3{,}322 \\ 2{,}099 & 10{,}905 \end{bmatrix} \begin{bmatrix} .1393 & 0 \\ 0 & .0346 \end{bmatrix} = \begin{bmatrix} 6{,}605 & 115 \\ 292 & 377 \end{bmatrix}$$

By identical reasoning, matrix E is the product of the census-I nonresponse ratios and the responding continuers matrix H:

$$\begin{bmatrix} .1290 & 0 \\ 0 & .0578 \end{bmatrix} \begin{bmatrix} 47{,}411 & 3{,}322 \\ 2{,}099 & 10{,}905 \end{bmatrix} = \begin{bmatrix} 6{,}117 & 429 \\ 121 & 630 \end{bmatrix}$$

Thus, at step 2 the expanded transition matrix looks like this:

0	A		B	
C	D		6,117	429
			121	630
F	6,605	115	47,411	3,322
	292	337	2,099	10,905

Step 3: Continuing Farms That Responded in Neither Census

The farms that failed to respond to either census (matrix D) are calculated analogously to the farms in matrices E and G. Each entry in D will be proportional to the corresponding entry in H by a factor of proportionality that equals the product of two nonresponse rates. Specifically, d_{ij} equals h_{ij} times the class-i nonresponse ratio from census I times the class-j nonresponse ratio from census II. In matrix terms, D equals the product of census-I nonresponse ratios multiplied by H which is in turn multiplied by census-II nonresponse ratios:

$$\begin{bmatrix} .1290 & 0 \\ 0 & .0578 \end{bmatrix} \begin{bmatrix} 47{,}411 & 3{,}322 \\ 2{,}099 & 10{,}905 \end{bmatrix} \begin{bmatrix} .1393 & 0 \\ 0 & .0346 \end{bmatrix} = \begin{bmatrix} 852 & 17 \\ 15 & 22 \end{bmatrix}$$

The expanded transition matrix upon completion of step 3 is:

0	A		B	
C	852	15	6,117	429
	17	22	121	630
F	6,605	115	47,411	3,322
	292	337	2,099	10,905

Step 4: Entrants and Exiters That Responded

Matrices B and F represent the responding farms that exited and entered farming, and are simply computed as residuals from the responding farms totals in each census year class (table 1). That is, the number of responding class-1 exiters in census I (the top entry in matrix F) is the difference of the class-1 missing farms and the class-1 continuing farms (the sum of the top row entries in matrix G). The bottom entry in F is the difference of the class-2 missing farms and the sum of the bottom row entries in G. That is:

26,708 - 6,605 - 115 = 19,988
3,311 - 292 - 377 = 2,642.

Likewise, the responding entrants (matrix B) are computed as residuals:

$$21{,}164 - 6{,}117 - 121 = 14{,}926$$
$$3{,}136 - 429 - 630 = 2{,}077.$$

Step 4 results in the following expanded transition matrix:

0	A		14,926	2,077
C	852	15	6,117	429
	17	22	121	630
19,988	6,605	115	47,411	3,322
2,642	292	337	2,099	10,905

Step 5: Entrants and Exiters That Did Not Respond

Matrices A and C represent entrants and exiters that failed to answer the census questionnaire, and are computed from B and F using the nonresponse ratios of table 2. For example, the number of nonresponding class-1 entrants in census I is proportional to the number of responding class-1 entrants by a factor of 0.1290, the nonresponse ratio for class 1 farms in census I. In matrix notation, the nonresponding exiters of matrix F equal:

$$\begin{bmatrix} .1290 & 0 \\ 0 & .0578 \end{bmatrix} \begin{bmatrix} 19{,}988 \\ 2{,}642 \end{bmatrix} = \begin{bmatrix} 2{,}579 \\ 153 \end{bmatrix}$$

Likewise, the nonresponding entrants of matrix B equal:

$$\begin{bmatrix} 14{,}776 & 2{,}314 \end{bmatrix} \begin{bmatrix} .1393 & 0 \\ 0 & .0346 \end{bmatrix} = \begin{bmatrix} 2{,}079 & 72 \end{bmatrix}$$

The complete expanded transition matrix equals:

The complete expanded transition matrix equals:

0	2,079	72	14,926	2,077
2,579	852	15	6,117	429
153	17	22	121	630
19,988	6,605	115	47,411	3,322
2,642	292	337	2,099	10,905

Step 6: The Corrected Transition Matrix

Step 5 completes the calculation of the expanded transition matrix, with the exception of the number of potential farmers. (No empirical estimate for this group exists. We will posit a population of potential farmers equalling 50,000.) The total number of continuing farms is simply the sum of matrices D, E, G, and H. The total number of exiting farms is the sum of matrices C and F. And the total number of entering farms are the sum of matrices A and B. The corrected transition matrix is shown in table 3 and reproduces the totals of table 2.

Table 3--Example: The corrected transition matrix

Item	Census II farms: Missing	Census II farms: Class 1	Census II farms: Class 2	Row totals
		Number		
Census I farms:				
Missing	50,000	17,005	2,149	69,154
Class 1	22,567	60,985	3,880	87,432
Class 2	2,794	2,530	11,934	17,258
Column totals	75,361	80,520	17,963	

The Procedure: A Matrix Formulation

While the above example used the simplest classification scheme consisting of only two classes, the six steps can be applied to classifications of any number of classes. The above procedure is easily implemented as a set of formulas in standard spreadsheet packages. Unfortunately, the computation becomes cumbersome and prone to typographic and arithmetic error as the number of classes increases. This problem can be minimized, however, by a different formulation of the procedure. The six steps can be condensed into a single matrix equation that may be easily computed with the matrix operation features available in most standard spreadsheets:

$$\mathbf{A}_{I} \quad \mathbf{T} \quad \mathbf{A}_{II} \quad = \quad \mathbf{T}^{*}$$

where $\mathbf{T}$ is the uncorrected transition matrix, $\mathbf{T}^{*}$ is the corrected transition matrix, $\mathbf{A}_{I}$ is the nonresponse adjustment matrix for census I, and $\mathbf{A}_{II}$ is the nonresponse adjustment matrix for census II. These matrices are all of rank n+1.

$$\mathbf{T} = \begin{bmatrix} z & a_1 & a_2 & a_3 & \cdots & a_n \\ b_1 & c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ b_2 & c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ b_n & c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix}$$

$$\mathbf{A}_{I} = \begin{bmatrix} 1 & -r_1 & -r_2 & -r_3 & \cdots & -r_n \\ 0 & 1+r_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1+r_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1+r_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1+r_n \end{bmatrix}$$

$$
\mathbf{A}_{II} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -s_1 & 1+s_1 & 0 & 0 & \cdots & 0 \\ -s_2 & 0 & 1+s_2 & 0 & \cdots & 0 \\ -s_3 & 0 & 0 & 1+s_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ -s_n & 0 & 0 & 0 & \cdots & 1+s_n \end{bmatrix}
$$

In matrix **T**, z is the number of potential farmers, a_i is the number of entrants in the ith class, b_j is the number of exiters in the jth class, and c_{ij} are the number of continuing farms that originated in the ith class and moved to the jth class by the next period. $\mathbf{T}^*$ is identical in form to **T**. In the matrices $\mathbf{A}_I$ and $\mathbf{A}_{II}$, r_i and s_j are the nonresponse:response ratios in censuses I and II respectively. The forms of these matrices are derived from the six-step procedure by using symbols for the partitioned matrices, collecting terms, and simplifying the resultant expressions.

The adjustment matrix for census I, $\mathbf{A}_I$, contains a topmost row of $-r_i$ entries which deflates all the exiter figures (which we know to be overestimated) by an amount proportional to each class's nonresponse ratio. $\mathbf{A}_I$ has a diagonal of $1+r_i$ which inflates all the continuing figures (known to be underestimated) by an amount proportional to each class's nonresponse ratio. Similarly the adjustment matrix for census II, $\mathbf{A}_{II}$, has a column of $-s_j$ which deflates the entrant figures (known to be overestimated) by an amount proportional to each class's nonresponse ratio. $\mathbf{A}_{II}$ has a diagonal of $1+s_j$ which has the effect of inflating all the continuing figures (known to be underestimated) by an amount proportional to each class's nonresponse ratio. Applying this matrix adjustment equation to our example of tables 1 and 2 gives the equation shown on the next page:

$$\begin{bmatrix} 1 & -.1290 & -.0578 \\ 0 & 1.1290 & 0 \\ 0 & 0 & 1.0578 \end{bmatrix} \begin{bmatrix} 50,000 & 21,164 & 3,136 \\ 26,708 & 47,411 & 3,322 \\ 3,311 & 2,099 & 10,905 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -.1393 & 1.1393 & 0 \\ -.0346 & 0 & 1.0346 \end{bmatrix}$$

$$= \begin{bmatrix} 44,212.1 & 17,005.9 & 2,149.0 \\ 22,567.2 & 60,983.3 & 3,880.3 \\ 2,793.9 & 2,529.6 & 11,934.4 \end{bmatrix}$$

This result agrees with the corrected transition matrix in table 3 (except for rounding differences). The only anomaly is in the row-1 column-1 entry representing potential farmers--a quantity that cannot be empirically established. Fortunately, the magnitude of potential farmers does not affect the computation of the other entries.

Given the form of the equation, in the limit as all the nonresponse ratios tend to zero, the two adjustment factors converge to identity matrices, and the adjusted matrix converges to the unadjusted matrix. This accords with common sense as to what effect diminishing nonresponse rates would have on census estimates of transition matrices.

Summary and Conclusions

A method has been presented which corrects for the failure of individuals who qualify as farm operators under census definitions to respond to the census questionnaires. This correction is necessary to remove bias from the estimates of flows into and out of agriculture by class and from one class to another. The correction is also necessary for the proper estimation of the transition probability matrix (which is central to Markov analysis) because the transition probability matrix is a linear transformation of the transition matrix. The method developed here is easy to implement because the method is merely a matrix product of three factors: a premultiplication of the raw unadjusted matrix by a matrix of nonresponse rates in the previous census period, and a postmultiplication of the raw unadjusted matrix by a matrix of nonresponse rates in the subsequent census period.

☆ U.S. Government Printing Office : 1991 - 282-958/40172