北京大学中国经济研究中心
China Center for Economic Research
讨论稿系列
Working Paper Series

No. E2006013 2006-10

Statistical Matching for Longitudinal Data of Rural Households in China: Construction of MHTS Panel Data Set and Estimation of Attrition Bias

Hisatoshi Hoken[1], Tetsuji Senda[2], Yoshiro Matsuda[3],
Hiroshi Tsujii[4] and Cao Liqun[5]

NO. E2006013 October 18, 2006

[1] Visiting research fellow, China Center for Economic Research at Peking University. E-mail: hoken@ccer.edu.cn
[2] Associate professor, Kagawa University
[3] Professor, Aomori Public College
[4] Professor, Ishikawa Prefectural University
[5] Research fellow, Research Center for Rural Economy

Statistical Matching for Longitudinal Data of Rural Households in China: Construction of MHTS Panel Data Set and Estimation of Attrition Bias

Hisatoshi Hoken[1], Tetsuji Senda[2], Yoshiro Matsuda[3],
Hiroshi Tsujii[4] and Cao Liqun[5]

NO. E2006013 October 18, 2006

Abstract:

This paper has examined the results of data matching of RCFPO (*Rural China Fixed Point Observations*) and the structures and characteristics of a new panel database termed MHTS (*Minor sets of High-quality Time Series*). The reliability of original ID number of RCFPO ascribed to each household must be questioned since the ID number is often mismanaged. In order to check the accuracy of the original ID and the continuity of RCFPO, we have developed data matching methods and construct new panel databases. Our studies have also demonstrated that a large number of spurious continuities of panel survey appear to exist in the original ID with the advent of time. Moreover, in order to test sample attrition biases of MHTS panel data sets, we have conducted the estimations on attrition probit and the BGLW test both by utilizing the entire sample and by village. The results indicate that it is highly probable that the attrition of sample households on MHTS data sets may produce an estimation bias. Therefore, close inspections and econometric adjustments are strongly recommended in order to reduce the bias of estimations for MHTS data sets.

Keywords attrition, panel data, statistical matching, rural household

JEL classifications C42, C81, O12

[1] Visiting research fellow, China Center for Economic Research at Peking University. E-mail: hoken@ccer.edu.cn
[2] Associate professor, Kagawa University
[3] Professor, Aomori Public College
[4] Professor, Ishikawa Prefectural University
[5] Research fellow, Research Center for Rural Economy

I. Introduction

The Utilization of longitudinal data is widespread in the field of economic analysis. Since many types of panel data sets are collected and preserved by academic organizations for public use, there has been a rapid development of new econometric theories on panel analysis in recent years. Although the construction of databanks contributes to the development of microeconometrics and the understanding of economic behavior, it could also cause the quality and structure of data sets to be ignored. There is a tendency among recent economists to not pay attention to the sampling design and characteristics of the data sets that they are using, and to conduct econometric estimations without checking the representativeness of data sets.

As Deaton [1997] indicated, panel data have a number of specific problems. One of the most important problems is sample attrition, which implies that with the advent of time, a lesser number of the original observations remain in the survey—caused due to withdrawal from and refusal to participate in the survey. Attrition causes a selection bias and reduces the representativeness of survey data. Ignorance toward these problems could cause the people who utilize panel data for economic analysis to deduce inaccurate estimated values and erroneous policy implications. Therefore, it is extremely necessary to be well acquainted with the sampling design and characteristics of panel data for appropriate economic analysis.

The aims of this paper may be divided into two parts. First, we explain the results of data matching of a large-scale household longitudinal survey termed *Rural China Fixed Point Observations* (RCFPO, *zhongguo guding guanchadian diaocha*), which has been jointly conducted and compiled by the Chinese Communist Party (CCP) and the Ministry of Agriculture (MOA). This survey has been tracking the same household for over 20 years. The sample size of this compilation is approximately 300 villages and 20,000–25,000 households for each year. From the viewpoint of sample size and duration of survey year, RCFPO is one of the most valuable household panel data

sets pertaining to developing countries. However, because of some problems of data management, it is necessary to conduct procedures for data matching in order to construct a more reliable panel database.

From the interviews with statistical staff in charge of RCFPO and the enumerators in local areas, we confirmed that there are certain cases in which the ID numbers ascribed to individual households were changed year by year, and that when some households are withdrawn from the survey and consequently replaced by new households, the ID numbers of the former were ascribed to the latter. If we compile annual data sets based on original ID numbers, different households could be erroneously combined as the same household. For this reason, we have developed data matching methods and utilized these programs in order to construct panel data. Based on the results of data matching, we explain the structures and characteristics of the new panel database termed *Minor sets of High-quality Time Series* (MHTS)[1].

Second, comparisons among households that are surveyed continuously and those that have withdrawn from the survey have been conducted in order to measure the effects of attrition and non-sampling errors by using MHTS panel data set. At that time, we adopted the attrition probit model and BGLW test in order to detect attrition biases. These experiments contribute toward reminding us of the importance of data cleaning, understanding of sampling designs of survey data, and confirmation of a data set.

This paper is structured as follows: Section 2 describes the sampling design of RCFPO and 20 percent of the resampling panel data set that we have used for constructing the new panel data set. Section 3 presents the basic concept and the results of data matching; the characteristics of MHTS

[1] The MTHS panel dataset is constructed through the joint research among the Research Center for Rural Economy (RCRE) in China, Kyoto University, and Hitotsubashi University in Japan during 1999–2002 supported by the Grant-in-Aid for Scientific Research (GIASR) of the Ministry of Education, Culture, Sports, Science and Technology, and the Japan Society of the Promotion of Science (JSPS) [GIASR re. no. 11691074]. We also accepted GIASR in 1996 [GIASR re. no. 08209113] and in 1997 [GIASR re. no. 09206108] for conducting preliminary works for our joint research. In addition, the results of this working paper have been achieved through the financial support received from GIASR during 2003–2006 [GIASR re. no. 15402020].

panel data are also presented in this section. The theoretical frameworks for panel attrition are presented in Section 4. We estimate two types of attrition models to clarify the effects of attrition biases in the MHTS panel data set. A summary of major findings and concluding remarks are presented in Section 5.

II. Explanation of RCFPO and 20 percent of the resampling data set

(1) Sampling design of RCFPO

RCFPO is a longitudinal survey data set on rural households in China; it began in 1986 and is continuing until today with two interruptions in 1992 and 1994[1]. This survey has been jointly compiled by CCP and MOA in China, and the RCFPO database has been managed by the Research Center for Rural Economy (RCRE), which is a research organization under the jurisdiction of MOA. RCFPO encompasses all provinces except Tibet, and the sample size of this compilation is approximately 300 villages and 20,000–25,000 households in each year. The changes in the sample size of RCFPO are presented in Table 1. This table indicates that the sample size decreased during 1986–1991, and there was a large decline in 1993; the sample size reduced to 20,000 households ever since.

The sampling method of RCFPO is based on a "3-stage purposive selection." The sampling unit in the first stage was "county"; in other words, a county was sampled from every region (*diqu*) of each province. In the second stage, all villages of a county were classified into "upper," "middle," and "lower" areas according to per capita income and geographical features. The ratio of classes was fixed at 30, 40, and 30 percent, respectively. The villages were selected from each class in accordance with these ratios.

[1] Prior to the introduction of RCFPO, a rural household survey termed "the rural social and economic typical sampling survey" was conducted in 1984. At that time, nationwide survey organizations were established and contributed to the establishment of RCFPO.

In the third stage, the sampling and survey units were households. The systematic sampling method was adopted for selecting households; households were arranged in lists on the basis of per capita income, from which approximately 100 households were sampled. While the sampling unit was selected purposively in the first and second stages, the households within a village were selected randomly.

Three types of questionnaires exist in RCFPO: questionnaire for 1986–1991, 1993, and 1995–2001. RCFPO had two interruptions in 1992 and 1994; however, it was resumed in 1993 and 1995. During the revival, the contents and volumes of the questionnaire was changed substantially; hence, the panel data sets must be handled cautiously. The basic classifications of questions on the three questionnaires are presented in Table 2. It should be noted that the questions that are included in the same classifications are not necessarily identical among the three questionnaires, although most of the questions are the same. It is evident that the number of questions has increased since the survey in 1993, and the disparity between the questionnaires for the surveys in 1993 and 1995–2001 is not much larger than that between the surveys of 1986–1991 and 1993.

By conducting a detailed examination of the definitions of each question in the three questionnaires, we have compiled a table of the common questions during 1986–2001. The details of this table are summarized in Table 5. There are a total of 578 questions in the 3 questionnaires—of which 178 questions are identical among the 3 questionnaires and 210 questions are identical between the surveys in 1993 and 1995–2001. The definitions of other questions are different in all 3 questionnaires with the exception of 1 question. This result also indicates that the questionnaires were modified substantially in 1993; the questionnaires of the surveys in 1993 and 1995–2001 are highly similar, although the basic questions on households have remained the same in the questionnaires during all the survey periods.

(2) Joint research and 20 percent resampling from RCFPO

One of the purposes of the international joint research among RCRE, Kyoto University, and Hitotsubashi University in Japan is to rearrange RCFPO for a more reliable panel database. Our joint research during 1999–2002 has been conducted with financial support in the form of a Grant-in-Aid for Scientific Research (GIASR) from the Ministry of Education, Culture, Sports, Science and Technology and the Japan Society of the Promotion of Science (JSPS). A chronology of our joint research is presented in Table 3. As mentioned above, since the ID numbers ascribed to individual households are not rigorously managed in certain areas, reliability of RCFPO as a panel data set could be damaged to a certain extent. In order to improve the quality of RCFPO as a panel database, we have developed data-matching methods and utilized these programs to construct a new database termed MHTS panel database.

However, because of institutional reasons, our joint research can utilize 20 percent of the resampling data for 1986–2001 from RCFPO. In order to make the most of the advantages of the survey, we determine 3 criteria for selecting 20 percent of the resampling data.

1) Continuity of households

In certain areas, the sample size of household changes substantially every year; in other words, it is not fixed. For example, in Jiangxi province, the sample size of households is maintained at approximately 1,200 during 1986–1988. However, the household survey was interrupted abruptly in 1989—the survey was revived in 1990. This discontinuity of the panel survey caused critical losses in the construction of the panel database; hence, we give priority to the provinces that have been surveyed continuously. Owing to attrition of sample households and changes in the ID number ascribed to individual households, we determine the village as the sampling unit of 20 percent of the resampling data.

2) Reliability and quality

RCRE provided Japanese teams with the 4 sample village data sets of RCFPO pertaining to Hebei, Shanxi, Jiangsu, and Guangdong in advance, and also offered supplementary data (entire ID number and data on matching key variables of entire households) in order to select sample villages. We examined the non-sampling errors of these data and discovered that the data pertaining to the northern area—particularly Hebei and Shanxi—were superior. For this reason, more sampling weightage was placed on the sample pertaining to the northern area.

3) Coverage and distribution

One of the aims of the joint research is to compare household behavior and farm management of different geographical regions. For the purpose of conducting a comparison among regions, it is necessary to select the sample that would represent the characteristics of each agricultural area. In keeping with the studies of Rossing Buck, we divided rural China into 8 agricultural areas and selected sample villages from each agricultural area, although the sampling weights ascribed to each agricultural area were not the same due to the continuity and the reliability of household data.

Based on these criteria, we selected 54 villages from approximately 300. Twenty percent of the resampling data encompasses 14 provinces, and the sample size of households is approximately 5,000 each year. The results of village selections are presented in Table 4. It should be noted that the percentage of villages in the northern region is higher—9 villages were selected from Hebei province and 7 from Shanxi province. The percentage of villages from the aforementioned regions in the original RCFPO was 8 percent in 1986. Contrastingly, the proportion of Hebei and Shanxi villages in the 20-percent sample is approximately 30 percent. It is important to note that when

dealing with these data sets, attention must be paid to bias from this sampling selection.

III. Method of data matching and results

(1) Basic concept of data matching

In constructing panel data from 20 percent of the resampling database, it is natural to believe that the ID number is the most important information for matching data sets. However, there are certain cases in which the ID numbers ascribed to individual households are changed annually. Moreover, when a household withdraws, it is replaced with a new household that is ascribed the old ID number. Hence, if the annual data sets are compiled on the basis of the ID number, different households would be erroneously combined as the same household. For this reason, we developed data matching methods without utilizing the ID numbers for matching keys.

The basic concept of data matching and its algorithm are presented in Figures 1 and 2 respectively; both the concept and algorithm have been invented by Dr. Inaba, a member of our joint research team. The questionnaire pertaining to the 1986–1991 and 1995–2001 surveys contains questions regarding items of "year-end" and "year beginning." Using the data pertaining to 1986 and 1987 as an example, the item from the data set pertaining to the end of 1986 must equal that pertaining to the beginning of 1987. Even if the numbers were not exactly equal, they would probably be close. We use this basic idea for exact matching of data set[1]. In doing so, we rely on matching variables that are recorded both at the beginning and end of the year. To be precise, we use "*the balance of savings*" (*cunkuan yu'e*) and "*the balance of cash in hand*" (*shoucun xianjin*) as our matching key variables for data pertaining to 1986–1991.

Based on the results of data matching, we divide households into three groups. First, if the

[1] The concept of data-matching by using items of year-end and year beginning is also adopted in Chen and Ravallion [1996]. They constructed household panel datasets of 4 southern provinces (Guangdong, Jiangxi, Guizhou, and Yunnan) for 1985–1990 by using microdata pertaining to the Rural Household Survey (RHS) conducted by the National Statistics Bureau (NSB).

year-end number of the abovementioned variables is exactly equal to that in the beginning of the next year, we regard these households as (a) "*exactly matched households*."

Second, in the case that the number of these variables is close, and the reference variables are also the same, we define such households as (b) "*matched households by analogy*" (statistical matched households). In addition to key variables, we selected reference variables, which would be exactly matched or would not easily change between consecutive years, for data matching in order to check the results of exact matching and supplement matching by analogy.

When numeral variables of key variables and reference variables are completely different, and no matching observations between consecutive years could be detected, we regard such households as (c) "*unmatched households*." Details on matching key variables and reference variables are presented in Table 6. However, with regard to data pertaining to 1991–1993–1995, the methods of comparing the items of "year-end" and that of "year beginning" could not be used rigorously; therefore, according to the outcome of data matching for non-consecutive years using sample data, we have selected certain variables that would not be easy to change (land area, non-production assets, etc.) as key and reference variables in order to conduct data matching.

(2) Results of data matching and tables of panel patterns

The results of data matching between consecutive years are presented in Table 7. In the aggregate, 75.6 percent of household data are matched exactly, and along with the number of "matched households by analogy", 93.4 percent of household data is matched by this method. However, disparities of matching ratio existed among different years. The share of household data that was exactly matched is 87.5 percent for data pertaining to 1986–1991, 35.7 percent for data pertaining to 1991–1993–1995, and 77.3 percent for data pertaining to 1995–2001. The proportion of exactly matched households has decreased by approximately 10 percent between data pertaining to

1986–1991 and 1995–2001. It is natural that the percentage of exactly matched households is extremely low for data pertaining to 1991–1993–1995, since matching keys are not necessary coincidental between years.

On the other hand, a large gap in matched ratio is not observed between data pertaining to 1986–1991 and 1995–2001. The percentages of matched households are 95.9 percent for data pertaining to 1986–1991, 75.0 percent for data pertaining to 1991–1993–1995, and 97.2 percent for data pertaining to 1995–2001. Although data pertaining to 1986–1991 and 1995–2001 possess a high matched percentage, much more attention must be paid to the non-sampling errors in data pertaining to 1995–2001.

Since the proportion of matched observations for data pertaining to 1991–1993–1995 is also relatively low in certain villages, key variables are completely mismatched and data matching is impossible for these villages (16 out of 54 villages). Therefore, we construct 3 types of panel databases for 1986–1991, 1995–2001, and 1986–2001 in order to make the most of the data. We term these data sets as MHTS panel data sets. These data sets have been developed as a consequence of our joint international project among RCRE, Kyoto University, and Hitotsubashi University during the period 1999–2002.

Table 8 presents the results of data matching by provinces. While the percentages of matched households are almost 90 percent in all provinces, those of exactly matched households are dispersed among provinces. In Hunan province, the proportion of exactly matched households is 89.1 percent—the highest among all provinces—on the other hand that of Shangdong province is 60.4 percent—the lowest of all provinces. Regardless of a large sample size in Hebei and Shanxi provinces, the proportions of exactly matched households are 77.3 percent and 86.2 percent, respectively. These results indicate that the accuracy of the surveyed data set is varied among regions; thus, attention must be paid to the disparities of non-sampling errors among surveyed areas.

In order to illustrate the pattern of continuity of surveyed years on individual households—based on the result of matching—we have constructed the table of panel data patterns in Tables 9 to 11. In the table, “O” (circle) indicates that households have been surveyed in the concerned year and “×” (ekes) indicates that households have not been surveyed. Table 9 presents panel patterns for the 1986–1991data set. In total, 6,469 households have been included; the second row indicates that 3,632 households—over 50 percent of the households—have been surveyed 6 years in succession. On the other hand, 604 households were surveyed only in 1986 and 1987—merely for 2 years—and 459 households were surveyed only in 1986. In other words, a large number of households were withdrawn from the survey in the initial years. These results indicate that substantial losses of panel members had occurred in the initial few years. This is partially because the sample size of RCFPO itself has decreased since 1987.

Panel patterns for the 1995–2001 data sets are presented in Table 10. The households that have been surveyed 7 years in succession account for 67.1 percent of the total surveyed households. Except for this pattern, the share of other panel patterns is smaller than that of the 1986–1991 data sets. This indicates that the panel survey for 1995–2001 is more stable than that for 1986–1991.

Next, Table 11 presents panel patterns for the 1986–2001 data set. As mentioned above, since certain villages—in which household data are completely unmatched during 1991–1993–1995—have been excluded from this data set, the number of villages that have been excluded from the data set account for 16 out of 54 villages (30 percent of the entire 1986–2001 data set). Thus, caution must be exercised to ensure that the 1986–1991 and 1995–2001 panel data sets are not subsets of the 1986–2001 data sets.

The share of households that have been surveyed in succession from 1986–2001 is 31 percent, while large discontinuities of panel survey have been found around 1991–1995. To be precise, 7.4 percent of the households have withdrawn from the panel survey since 1993, and 5.6 percent of the

households were added to the panel survey since 1995 and have been surveyed in succession since then. Moreover, surveying for 4.3 percent of the households was discontinued in 1993, and surveys of 3.9 percent of the households have begun since 1993. These results indicate that the continuity of panel survey was considerably impaired by the interruptions of the survey in 1992 and 1994. It can be concluded that the attrition of RCFPO is caused not only by household factors (refusal, removal, migration, etc.), but also by institutional factors (change of sampling design and data management, etc.).

Finally, Table 12 presents the percentage of households that have been surveyed in succession by province during the wave. We have termed these households as "complete panel households." The composition of these households is strikingly different among provinces. In Yunan province, the percentages are approximately 90 percent for every period, while in Heilongjiang province the proportions are relatively low—30.0 percent for 1986–1991, 51.4 percent for 1995–2001, and 11.1 percent for 1986–2001.

Furthermore, we find that the percentages have changed strikingly among surveyed periods in certain provinces. In Hebei province, the composition of complete panel households is 72.4 percent for 1986–1991, while the percentage has declined drastically to 48.5 percent for 1995–2001 and 37.7 percent for 1986–2001. On the other hand, in Liaoning province the percentage has improved from 22.6 percent for 1986–1991 to 84.1 percent for 1995–2001, although the share of complete panel households is 8.6 percent because of high attrition of surveyed households during 1986–1991. For other provinces, striking differences of the percentages pertaining to 1986–1991 and 1995–2001 are not observed, except for Jiangsu province. From the results of matching by province, we can conclude that compositions and characteristics of panel data differ strikingly among provinces, and it is required that databases are handled taking into consideration these disparities.

(3) Consistency of panel patterns between new and original ID

The results of panel patterns presented above are arranged by use of new ID numbers that we ascribed to each household. For the purpose of comparing panel patterns based on original and new ID numbers, we select the top 5 panel patterns based on each ID number. Table 13-1 is a comparison table of panel patterns based on the 1986–1991 data sets. This table shows that there is a large disparity in the proportion of complete panel households: the percentage of complete panel households of original ID is 76.2%, which is much higher than that of new ID—20 percent higher. The proportion of complete panel households on original ID for other data sets is much greater than those on new ID. With regard to the 1995–2001 and 1986–2001 data sets, the proportions of complete panel household on original ID are 88.2 percent and 40.7 percent, respectively. These percentages are 10–20 percent higher than those of the new IDs.

In order to examine the discrepancies of panel patterns in detail, we summarize the (in)consistencies of panel patterns between the two types of ID numbers in Table 14. The inconsistencies of panel patterns account for 35.8 percent for 1986–1991, 29.9 percent for 1995–2001, and 54.9 percent for 1986–2001, respectively. The discrepancy of panel patterns between the two types of ID numbers is more acute on longer panel data sets; this finding is consistent with that of Table 10. These results reveal that the substantial number of spurious continuities of panel survey appear to exist in the original ID with the change of time. Hence, our trials for data matching and the construction of the MHTS panel data set may be highly valued in terms of rectifying spurious continuities of surveyed households and controlling grave errors caused by blind dependence on the original ID.

VI. Experiment on panel database

VI-1 Existing studies on panel data attritions of developing countries

Attritions of sample households are widely discussed using U.S. household panel data since large-scale longitudinal household surveys such as the Michigan Panel Survey of Income Dynamics (PSID) and the National Longitudinal Survey of Youth (NLSY) have been conducted in the U.S. In contrast to this, the amount of literature on attrition in developing countries is considerably limited because of the rareness of existing panel data sets on developing countries.

In the spring of 1998, a special issue of *The Journal of Human Resources* on "Attrition in Longitudinal Surveys," which concentrated on the problems of sample attrition, was published. This issue was widely noticed and attracted attention to sample attrition of panel data. In particular, Fitzgerald, Gottschalk and Moffitt [1998a, 1998b] have improved the statistical frameworks for testing the attrition bias[1]. Inspired by these researches, the amount of literature on attrition in developing countries has increased since the end of the 1990s in proportion to the increase in the number of panel surveys on developing countries, which have been conducted by the government of these countries and international organizations.

Based on the framework of Fitzgerald, Gottschalk and Moffitt [1998a, 1998b], certain statistical tests have been attempted for attrition bias in panel data of developing countries. Alderman, Behrman, Kohler, Maluccio and Watkins [2000] has considered the extent and implications of attrition for three longitudinal household surveys from Bolivia, Kenya, and South Africa. Maluccio [2000] has examined the attrition of the South African panel study KwaZulu-Natal Income Dynamics Study and assessed the extent of attrition bias for a specific empirical model[2].

These studies have indicated that the problem of attrition appears to be more serious in developing countries as compared with developed countries because availability of information and capability for tracking is superior in developed countries; moreover, the high level of mobility and

[1] The results of these studies indicate that the biases in estimated socioeconomic relations due to attrition are small, despite attrition rates of 50 percent and with significant differences between attritors and nonattritors.

[2] The results of these two studies are not necessarily consistent. The former insisted that attrition is not a general problem for obtaining a consistent estimate of the coefficients, while the latter emphasized attrition was indeed

long-distance migration associated with development would complicate survey work in developing countries. The refusal rate of developing countries is generally much lower that that of developed countries, thereby reflecting lower opportunity cost of time and possibly different cultural attitudes toward the interviewing process (Maluccio [2000]). For instance, PSID has experienced approximately a 50 percent sample loss from cumulative attrition in twenty years.

By using the summary table of attrition rates for longitudinal household survey data in developing countries by Alderman, Behrman, Kohler, Maluccio and Watkins [2000], we compare the level of attrition rate of MHTS panel data with those of other developing countries. As shown in Table 15, attrition rates among rounds of 1986–1991, 1995–2001, 1986–2001 MHTS panel data sets are 43.9 percent, 32.9 percent and 69.0 percent, respectively. These rates are placed at the highest level among panel surveys for developing countries. However, from the viewpoint of attrition rates per year, these proportions remain at 8.8 percent, 5.5 percent, and 4.8 percent, respectively. These results indicate that the attrition rates per year of MHTS are much lower that those of African countries and slightly higher than those of Asian countries.

ICRISAT Indian household survey is one of the most famous longitudinal surveys for developing countries, which encompassed 240 rural households during 1975–1976 to 1984–1985. Unlike other longitudinal surveys, ICRISAT and RCFPO (MHTS) household surveys are basically conducted every year during the survey period. Both surveys established a special survey system for collecting and managing household data, which contributed to a low attrition rate, although the sample size of RCFPO is much larger than that of ICRISAT. In this sense, RCFPO and MTHS possess unique characteristics as the panel databases pertaining to developing countries.

VI-1 Statistical framework of data attrition and method for testing attrition[1]

---

model-specific.

[1] The descriptions of this subsection can mainly be referred to Alderman, Behrman, Kohler, Maluccio and Watkins

The statistical frameworks for testing the attrition bias have been developed by Fitzgerald, Gottschalk and Moffitt [1998a, 1998b]. They distinguish the types of attrition bias from selection on observables and selection on unobservables. The differences between the two types of selection are based on the relationships between a conditional population density $f(y|x)$ and an attrition function. The conditional population density is defined as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad y \text{ is observed if } A = 0 \quad (1)$$

where A is an attrition indicator equal to 1 if an observation is missing its value of y because of attrition, and equal to zero if an observation is not missing its value of y. Therefore, we observe only the density $g(y \mid x, A = 0)$, and we require additional information or restriction to infer $f(\cdot)$ from $g(\cdot)$. This can be acquired from the probability of attrition $\Pr(A = 0 \mid y, x, z)$, where *z* is an auxiliary variable that is assumed to be observable for all units, but is not included in *x*.

The distinctions between selections on observables and selections on unobservables are presented in equations of the conditional probability function. In other words, selection on observables occurs if

$$\Pr(A = 0 \mid y, x, z) = \Pr(A = 0 \mid x, z) \quad (2)$$

Selection on unobservables occurs if (2) fails to hold; therefore, the attrition function cannot be reduced from $\Pr(A = 0 \mid y, x, z)$. The estimation form of selection on unobservables is formulated by the use of the attrition index function. However, a suitable instrument for the selection on unobservables is difficult in the case of nonresponse because there are few variables that affect nonresponse that can be credibly excluded from the main equation for y. Our database also possesses

---

[2000] and Fitzgerald, Gottschalk and Moffitt [1998a, 1998b].

few suitable instruments for estimating the attrition function.

If there is selection on observables, the variable *z* affects attrition propensities and also is related to the density of *y* conditional on *x*; in other words, *z* is endogenous to *y*. At this time, a lagged value of *y* can play the role of *z* if it is not in the structural relation being estimated and if it is related to attrition. Two sufficient conditions for the absence of attrition bias on observables are either that the weights equal to 1 (*z* dose not affect *A*) or that z is independent of *y* conditional on *x*.

Specification tests can be based on either of two conditions. One test is simply to determine whether candidate variables (lagged value of *y*) for *z* significantly affect *A*. In order to test this specification, we begin by estimating attrition probit on the sample of rural households. The dependent variable in this probit is whether attrition occurs between the survey round (1 = yes, 0 = no). We have included per capita income in the first year as well as predetermined family (household head) background as independent variables.

Another test—termed the BGLW test—is based on Becketti, Gould, Lillard and Welch [1988]. In this test, y_0 (per capita income in the first year) is regressed on x and on future *A* (in other words, whether or not the individuals attrite later) and other predetermined family (household head) background variables. The test for attrition is based on the significance of *A* in that equation. This test has strong relationships with the first attrition probit, and the two equations are simply inverses of one another.

VI-3 Estimated results on attrition bias

We utilize the abovementioned models to test the attrition bias of MHTS panel data. First, we have conducted two types of tests by using MHTS panel data as a whole. The results of attrition probit are presented in Table 16. The basic trends of estimated coefficients are almost the same for both databases, except for the results of the educational dummy; moreover, the initial economic

characteristics of households have significant influences on later attrition. The household whose per capita income was higher in the beginning of the survey wave have a significant tendency for withdrawing from the survey in the future. On the other hand, per capita cultivated area, number of labor, and floor area of housing at the beginning of the survey wave have a significant negative effect on attrition.

The attributes of household head (dummy variables on CCP members and cadre of village) have no significant impact on later attrition. In contrast to this, the effects of educational level are different between the 1986–1991 and 1995–2001 databases. With regard to the 1986–1991 database, all educational dummies have a significant negative effect on attrition, although these variables are not significant for 1995–2001. These results imply that predetermined economic characteristics of households are significantly different among attritors and nonattritors, and would induce sample attrition bias in estimations, although basic characteristics of household heads at the beginning of the survey wave are not necessarily significantly correlated with later attrition.

In addition, we have estimated an income function in order to conduct the BGLW test. Attrition dummy and interaction terms of attrition dummy with other predetermined family background variables are included in the estimation model. For the purpose of clarifying the effects of attrition on coefficients, two formations of F-statistics are estimated: one hypothesis is that all interaction terms are equal to zero, the other is that all interaction terms and the attrition dummy are equal to zero.

The result of the estimations is reported in Table 17. Both the F-tests—conducted to test whether the attrition variable is independent of other predetermined family background variables—are significantly rejected for 1986–1991 and 1995–2001. Thus, the slope coefficients and intercept of the income function are significantly affected by attrition, and predetermined family background variables are not independent from attrition. This result is consistent with that of attrition probit.

However, in constructing the MHTS panel database from 20 percent of the resampling data sets, we have adopted the method of purposive selection. Since the estimated results using the entire MHTS panel data inevitably involve severe sampling biases—with the exception of attrition bias—we have estimated attrition probit by village for the purpose of controlling the biases caused by non-randomness of village selection. The results of estimation by village are summarized in Table 18. For the purpose of saving space, the number of coefficients that are statistically significant at the 5 percent village level estimation is shown in this table.

For the 1986–1991 databases, 9 cases of the coefficient on per capita income in 1986 are significant, which accounts for approximately 23 percent of valid sample villages. The coefficients are significantly positive in 6 cases and negative in 3 cases. This result indicates that the households whose per capita income level was higher in 1986 tend to withdraw from the survey. The estimated results of per capita income for 1995–2001 are slightly different from those of 1986–1991. The cases in which the coefficients of per capita income are significant are decreased to 7 villages (21 percent), and the number of signs of significant coefficients is almost the same.

Except for per capita income, the educational level of household heads is likely to be strongly related with attrition of sample households, although the incidents of positive and negative signs are not the same between 1986–1991 and 1995–2001. For 1995–2001, attrition from the panel survey was much more likely for the households where the educational level of household heads is elementary and junior high school. On the other hand, the cases that the coefficients of number of labor are significant increased from 6 to 12 between 1986–1991 and 1995–2001, while the number of positive and negative signs is almost the same for 1995–2001. The same trend is observed with regard to the coefficient of area of housing. The number of significant coefficients increased from 5 to 9, and the coefficient is likely to be negatively correlated with attrition of sample households.

It should be noted that the likelihood function of attrition probit cannot be suitably converged for

approximately 10 to 20 percent of the villages because of the limited number of attrition, which would imply that attrition of these villages would happen randomly. Thus, we must be cautious with regard to the diversity of attrition among villages, and it would be suitable to distinguish the villages whose attrition rates are considerably high from those whose attrition would happen randomly.

Next, the results of the BGLW test by village are summarized in Table 19. This table presents the number of villages for which the joint F-test is significant at 5 percent level and the number of villages for which it is not. For 1986–1991 databases, F-tests that do not include the constant are significantly rejected at 5 percent level in 31 villages out of 46—or 67 percent of the total villages—and F-tests that include the constant are significantly rejected in 34 villages (74 percent). This trend of estimated results is the same as that pertaining to 1995–2001. In other words, the number of villages for which the F-test is conducted both excluding the constant and including it is 29 villages, or 71 percent of the valid sample. The results of the BGLW test imply that the economic conditions of sample households are significantly correlated with future attrition in more than half of the sample households. These results are consistent with the results of attrition probit.

From the results of attrition probit and the BGLW test both using the entire sample and by village, it can be concluded that it is highly probable that the attrition of sample households on MTHS data sets produces an estimation bias; thus, close inspections and econometric adjustments are strongly required in order to reduce the bias of estimations for MHTS panel database.

V. Conclusion

This paper has examined the results of data matching of RCFPO and the structures and characteristics of a new panel database termed MHTS. Although RCFPO is one of the most valuable panel databases pertaining to rural households in developing countries, the reliability of original ID number ascribed to each household must be questioned since the ID number is often mismanaged. In

order to check the accuracy of the original ID and the continuity of RCFPO, we have developed data matching methods and construct new panel databases by utilizing 20 percent of the resampling data of RCFPO.

The results of data matching indicate that the accuracy of the surveyed data set and the patterns of panel data are strikingly different among regions and durations. Our studies have also demonstrated that a large number of spurious continuities of panel survey appear to exist in the original ID with the advent of time. In particular, we must bear in mind that the continuity of the panel survey was considerably hampered by the interruptions in the survey in 1992 and 1994. Thus, it is strongly required that the RCFPO database be modified by checking the original ID numbers for conducting panel data analysis. MHTS panel data set is a result of the trials to construct a more reliable panel data set using RCFPO.

Second, in order to test sample attrition biases of MHTS panel data sets, we have conducted the estimations on attrition probit and the BGLW test both by utilizing the entire sample and by village. The results of the estimations indicate that it is highly probable that the attrition of sample households on MHTS data sets may produce an estimation bias; therefore, close inspections and econometric adjustments are strongly recommended in order to reduce the bias of estimations for MHTS data sets. Furthermore, it is desirable to distinguish sample villages based on attrition rate since the villages whose attrition rates are extremely low tend to be free from attrition bias.

References

Alderman, H., R. Behrman, H.-P. Kohler, J. A. Maluccio and S. C. Watkins [2000], "Attrition in Longitudinal Household Survey Data: Some Tests for Three Developing Country Samples," FCND Discussion Paper, No. 96, International Food Policy Research Institute.

Becketti, S., W. Gould, L. Lilliard and F. Welch [1988], "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," *Journal of Labor Economics*, vol. 6, no. 4, pp. 472–492.

Chen, Shaohua and Martin Ravallion [1996], "Data in Transition: Assessing Rural Living Standards in Southern China" *China Economic Review*, vol. 7, no. 1, pp. 25–56.

Deaton, Angus [1997], *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore: Johns Hopkins University Press.

Fitzgerald, J., P. Gottschalk and R. Moffitt [1998a], "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *The Journal of Human Resources*, vol. 33, no. 2, pp. 251–299.

Fitzgerald, J., P. Gottschalk and R. Moffitt [1998b], "An Analysis of the Impact of Sample Attrition on the Second Generation of Respondents in the Michigan Panel Study of Income Dynamics," *The Journal of Human Resources*, vol. 33, no. 2, pp. 300–344.

Kitamura, Yukinobu [2005], *Panel data bunseki (Analysis on Panel Data)*, Iwanami Shoten (in Japanese).

Maluccio, J. [2000], "Attrition in the Kwazulu Natal Income Dynamics Study, 1993-1998," FCND Discussion Paper, No. 95, International Food Policy Research Institute.

Matsuda, Yoshiro ed. [1999], *The Study on the Rural China Fixed Point Observations of Ministry of Agriculture, China*, The Discussion Paper of the Research Project "Exploring New Frontier in Statistical Analysis Using Micro Data Sets" (in Japanese).

Table 1 Sample size of original RCFPO by year and province

unit: household

	1986	1987	1988	1989	1990	1991	1993	1995	1996	1997
Beijing	110	109	110	110	110	110	110	110	110	110
Tianjin	396	400	320	320	271	200	200	195	195	200
Hebei	1,100	1,100	1,080	1,100	1,100	1,100	1,100	1,100	1,082	1,090
Shanxi	992	994	994	996	1,005	995	938	942	935	935
Neimenggu	221	222	222	222	222	205	164	122	90	136
Liaoning	2,615	2,237	1,200	1,200	1,193	1,193	600	1,200	1,195	1,200
Jilin	1,049	1,050	850	850	840	850	500	850	849	849
Heilongjiang	1,553	1,524	1,738	1,397	700	1,399	1,000	996	1,000	1,000
Shanghai	398	390	400	400	400	400	394	500	505	500
Jiangsu	1,515	1,000	1,000	990	1,000	1,000	920	810	810	805
Zhejiang	1,082	1,147	1,158	1,190	1,208	1,158	500	501	500	500
Anhui	810	1,338	1,412	1,310	1,398	1,409	1,402	1,402	1,402	1,401
Fujian	1,598	1,598	1,600	1,600	1,500	1,599	1,090	1,100	1,100	1,099
Jiangxi	1,191	1,199	1,198	0	980	1,184	710	898	950	950
Shandong	984	981	985	987	985	842	959	640	640	630
Henan	1,600	1,250	1,240	1,250	1,250	1,249	1,000	1,000	1,000	1,000
Hubei	1,587	1,666	1,689	1,728	1,500	1,648	899	900	900	895
Hunan	1,060	1,051	990	1,000	995	998	550	549	550	554
Guangdong	1,048	1,058	922	922	921	950	966	840	949	885
Guangxi	1,200	1,180	1,200	1,200	1,190	1,198	667	700	700	700
Hainan	0	0	149	148	300	300	300	360	360	300
Sichuan[1]	683	690	790	790	790	770	750	750	750	750
Guizhou	1,271	1,264	1,273	1,204	1,274	1,274	501	800	800	800
Yunnan	497	500	516	500	512	519	500	499	500	500
Shanxi	1,000	940	1,000	1,000	1,000	1,000	1,000	1,000	999	992
Gansu	283	283	230	303	303	303	303	303	303	303
Qinghai	347	346	251	251	249	250	185	250	250	251
Ningxia	698	484	703	697	350	350	700	350	350	350
Xinjiang	615	0	612	261	612	610	575	582	580	582
Total	27,503	26,001	25,832	23,926	24,158	25,063	19,483	20,249	20,354	20,267

Source: RCFPO database of RCRE

Note: Chongqing was independen from Sichuan Province in 1997. The sample size of Sichuan in 1997 includes that of Chongqing (150 households).

Table 2 Classification of questions by questionnaire

classification of questions	86–91 survey	93 survey	95–01 survey
(1) characteristic of household	17	15	15
(2) status of family number	25	33	33
(3) farmland condition	8	29	29
(4) asset holdings	27	33	33
(5) agricultural productions and sales	35	125	135
(6) allocation of labor	16	22	25
(7) income	56	45	48
(8) expenditure	49	48	50
(9) balance of savings and borrowings	57	17	17
(10) earnings and expenditures of grain	0	0	23
(11) consumption of main food	13	13	13
(12) holdings of durable consumer goods	9	14	18
total items	312	394	439

Table 3 Chronology of joint research

Date	Activity
1998 April	RCRE offered sample data of 4 villages for Japanese side
1998 October	Discussion on data-matching method at RCRE
1999 March	Conclusion of a protocol of joint research
1999 April	Beginning of our joint research (until 2002 March)
2000 October	Presentation of preliminary results at the Seventh Japan-China Symposium on Statistics
2002 March	International symposium on RCFPO data analysis in Beijing
2005 February	Publication of results of joint research in Japanese

Table 4 Sample size of 20 percent of the resampling data set of RCFPO by year and province

Province	villages	1986	1991	1993	1995	1998	2001
Hebei	9	850	850	850	850	850	807
Shanxi	7	654	655	667	669	670	663
Liaoning	4	899	400	200	400	400	400
Heilongjiang	5	660	500	385	377	385	380
Shanghai	2	198	200	198	200	200	200
Jiangsu	1	168	120	120	97	97	97
Anhui	6	473	473	470	471	473	471
Shandong	4	281	233	240	160	160	120
Hunan	2	135	125	100	100	99	100
Guangdong	3	300	313	310	307	323	307
Sichuan	5	250	250	260	250	250	245
Yunnan	2	199	200	200	200	201	200
Gansu	2	58	63	63	63	63	63
Ningxia	2	220	120	240	120	120	120
Total	54	5,345	4,502	4,303	4,264	4,291	4,173

Source: RCFPO database

Table 5 Structure of questionnaire (total 578 items)

items	percentage	86–91 survey	93 survey	95–01 survey
178	30.8	○	○	○
210	36.3	×	○	○
133	23.0	○	×	×
51	8.8	×	×	○
5	0.9	×	○	×
1	0.2	○	○	×

Note: Second row indicates that 178 questions are common among the three questionnaires.

Table 6 List of matching key variables and reference variables

(1)1986–91 dataset

	t year		t+1 year	
	item	item. No.	item	item No.
Key Var.	balance of savings (year end)	a275	balance of savings (year beginning)	a225
	balance of cash in hand (year end)	a281	balance of cash in hand (year beginning)	a231
Reference Var.	balance of government bonds (year end)	a276	balance of government bonds (year beginning)	a226
	balance of bank borrowings (year end)	a277	balance of bank borrowings (year end)	a227
	balance of lendings (year end)	a278	balance of lendings (year beginning)	a228
	balance of private borrowings (year end)	a279	balance of private borrowings (year beginning)	a229
	balance of investments (year end)	a280	balance of investments (year beginning)	a230

(2)1995–2001dataset

	t year		t+1 year	
	item	item. No.	item	item No.
Key Var.	area of cultivated land (year end)	nh057	area of cultivated land (year beginning)	nh049
	balance of food preserved (year end)	nh405	balance of food preserved (year beginning)	nh386
Reference Var.	amount of nonproductive assets (book price, year end)	nh104	amount of nonproductive assets (book price, year beginning)	nh104
	size of housing (year end)	nh109	size of housing (year beginning)	nh109
	price of housing (year end)	nh110	price of housing (year beginning)	nh110

(3)1991–93 dataset

	t year		t+1 year	
	item	item. No.	item	item No.
Key Var.	size of cultivated land (year end)	a043	size of cultivated land (year beginning)	t049
	size of wood land (year end)	a047	size of wood land (year end)	t073
	size of housing (year end)	a075	(size of housing (year end)) minus (size of newly-built housing within year)	(t109)－(t107)
Reference Var.	number of household	a018	number of household	t016
	number of household labor	a020	number of household labor	t018
	number of cultivated land plot	a045	number of cultivated land plot	t061
	amount of nonproductive assets (book price, year end)	a070	amount of nonproductive assets (book price, year beginning)	t104
	price of housing (year end)	a077	price of housing (year beginning)	t110

(4)1993–95 dataset

	t year		t+1 year	
	item	item. No.	item	item No.
Key Var.	size of cultivated land (year end)	t057	size of cultivated land (year beginning)	nh049
	size of wood land (year end)	t073	size of wood land (year end)	nh073
	size of housing (year end)	t109	(size of housing (year end)) minus (size of newly-built housing within year)	(nh109)－(nh107)
Reference Var.	number of household	t016	number of household	nh016
	number of household labor	t018	number of household labor	nh018
	number of cultivated land plot	t061	number of cultivated land plot	nh061
	amount of nonproductive assets (book price, year end)	t104	amount of nonproductive assets (book price, year beginning)	nh104
	price of housing (year end)	t110	price of housing (year beginning)	nh110

Table 7 Results of data matching by year

year	(1)sample size of matching				
		(a)exactly matched	(b) matched by analogy	(a)/(1)	{(a)+(b)}/(1)
1986–91 total	23,639	20,695	1,979	87.5%	95.9%
1986-87	5,240	4,369	439	83.4%	91.8%
1987-88	4,771	3,994	594	83.7%	96.2%
1988-89	4,616	4,139	395	89.7%	98.2%
1989-90	4,530	4,079	323	90.0%	97.2%
1990-91	4,482	4,114	228	91.8%	96.9%
1991–93–95 total	8,194	2,923	3,221	35.7%	75.0%
1991-93	4,146	1,269	1,791	30.6%	73.8%
1993-95	4,048	1,654	1,430	40.9%	76.2%
1995–2001 total	24,834	19,206	4,928	77.3%	97.2%
1995-96	4,194	3,006	1,082	71.7%	97.5%
1996-97	4,049	3,225	728	79.6%	97.6%
1997-98	4,031	3,193	694	79.2%	96.4%
1998-99	4,219	3,284	894	77.8%	99.0%
1999-00	4,189	3,221	851	76.9%	97.2%
2000-01	4,152	3,277	679	78.9%	95.3%
Total	56,667	42,824	10,128	75.6%	93.4%

Source: Estimations from 20 percent of the resampling data set of RCFPO.

Table 8 Results of data matching by province

Province	(1)sample size of matching				
		(a)exactly matched	(b)matched by analogy	(a)/(1)	{(a) + (b)}/(1)
Hebei	10,554	8,160	2,012	77.3%	96.4%
Shanxi	8,579	7,395	979	86.2%	97.6%
Liaoning	5,241	4,151	778	79.2%	94.0%
Heilongjiang	5,798	4,021	1,091	69.4%	88.2%
Shanghai	2,594	1,930	533	74.4%	94.9%
Jiangsu	6,110	4,538	901	74.3%	89.0%
Anhui	2,551	1,618	649	63.4%	88.9%
Shandong	1,379	833	275	60.4%	80.3%
Hunan	1,431	1,275	141	89.1%	99.0%
Guangdong	3,920	3,014	546	76.9%	90.8%
Sichuan	3,243	2,115	1,026	65.2%	96.9%
Yunnan	2,599	2,024	464	77.9%	95.7%
Gansu	808	603	159	74.6%	94.3%
Ningxia	1,860	1,147	574	61.7%	92.5%
Total	56,667	42,824	10,128	75.6%	93.4%

Source: Estimations from 20 percent of the resampling data set of RCFPO

Table 9 Panel patterns of MHTS panel data set (1986–1991)

Total: 6469 households

Nos.	%	Σ (%)	1986	1987	1988	1989	1990	1991
3632	56.1%	56.1%	○	○	○	○	○	○
604	9.3%	65.5%	○	○	×	×	×	×
459	7.1%	72.6%	○	×	×	×	×	×
338	5.2%	77.8%	×	○	○	○	○	○
227	3.5%	81.3%	○	○	○	○	×	×
226	3.5%	84.8%	○	○	○	×	×	×
172	2.7%	87.5%	×	×	×	×	×	○
167	2.6%	90.0%	×	×	○	○	○	○
132	2.0%	92.1%	○	○	○	○	○	×
117	1.8%	93.9%	×	×	×	×	○	○
90	1.4%	95.3%	×	×	×	○	○	○
89	1.4%	96.7%	×	○	×	×	×	×
65	1.0%	97.7%	×	×	×	×	○	×
46	0.7%	98.4%	×	×	○	×	×	×
29	0.4%	98.8%	×	×	×	○	×	×
28	0.4%	99.3%	×	○	○	×	×	×
16	0.2%	99.5%	×	×	×	○	○	×
13	0.2%	99.7%	×	×	○	○	×	×
7	0.1%	99.8%	×	×	○	○	○	×
7	0.1%	99.9%	×	○	○	○	×	×
5	0.1%	100.0%	×	○	○	○	○	×

Source: MHTS panel data set

Note: 1. "○" indicates that the households are surveyed in the concerned year and "×" implies that the households have not been surveyed.

2. Certain small number patterns have been omitted from the table.

3. The households whose data are completely unmatched have not been included in this table.

Table 10 Panel patterns of MHTS panel data set (1995–2001)

Total: 4838 households

Nos.	%	Σ (%)	1995	1996	1997	1998	1999	2000	2001
3246	67.1	67.1	○	○	○	○	○	○	○
261	5.4	72.5	○	○	×	×	×	×	×
230	4.8	77.2	○	○	○	○	○	○	×
161	3.3	80.6	○	○	○	×	×	×	×
155	3.2	83.8	×	×	×	○	○	○	○
126	2.6	86.4	×	○	○	○	○	○	○
124	2.6	88.9	○	○	○	○	○	×	×
101	2.1	91.0	○	×	×	×	×	×	×
88	1.8	92.9	×	×	○	○	○	○	○
79	1.6	94.5	×	×	×	×	×	×	○
60	1.2	95.7	○	○	○	○	×	×	×
59	1.2	96.9	×	×	×	×	×	○	○
39	0.8	97.8	×	×	×	×	○	○	○
34	0.7	98.5	×	○	×	×	×	×	×
26	0.5	99.0	×	×	○	×	×	×	×
13	0.3	99.3	×	×	×	×	○	×	×
9	0.2	99.4	×	×	×	○	○	×	×
7	0.1	99.6	×	×	×	×	×	○	×
6	0.1	99.7	×	×	×	×	○	○	×
3	0.1	99.8	×	○	○	○	○	○	×
2	0.0	99.8	×	×	×	○	○	○	×
2	0.0	99.9	×	×	○	○	×	×	×
2	0.0	99.9	×	×	○	○	○	×	×
2	0.0	99.9	×	×	○	○	○	○	×
2	0.0	100.0	×	○	○	×	×	×	×
1	0.0	100.0	×	○	○	○	×	×	×

Source: MHTS panel data set

Note: See notes pertaining to Table 9.

Table 11 Panel patterns of MHTS panel data set (1986–2001)

Total: 5212 households

Nos.	%	Σ (%)	86	87	88	89	90	91	93	95	96	97	98	99	00	01
1625	31.0	31.0	○	○	○	○	○	○	○	○	○	○	○	○	○	○
390	7.4	38.4	○	○	○	○	○	○	×	×	×	×	×	×	×	×
296	5.6	44.1	×	×	×	×	×	×	×	○	○	○	○	○	○	○
295	5.6	49.7	○	○	×	×	×	×	×	×	×	×	×	×	×	×
249	4.8	54.5	○	×	×	×	×	×	×	×	×	×	×	×	×	×
227	4.3	58.8	○	○	○	○	○	○	○	×	×	×	×	×	×	×
202	3.9	62.6	×	×	×	×	×	×	○	○	○	○	○	○	○	○
192	3.7	66.3	○	○	○	○	○	○	○	○	○	○	○	○	○	×
148	2.8	69.1	○	○	○	×	×	×	×	×	×	×	×	×	×	×
137	2.6	71.7	○	○	○	○	×	×	×	×	×	×	×	×	×	×
95	1.8	73.5	○	○	○	○	○	○	○	○	○	×	×	×	×	×
90	1.7	75.3	○	○	○	○	○	○	○	○	○	○	○	○	×	×
80	1.5	76.8	×	○	○	○	○	○	○	○	○	×	×	×	×	×
76	1.5	78.2	×	○	×	×	×	×	×	×	×	×	×	×	×	×
71	1.4	79.6	×	×	×	×	×	×	×	×	×	○	○	○	○	○
65	1.2	80.8	×	×	×	×	×	×	×	×	×	×	×	×	×	○
59	1.1	82.0	×	○	○	○	○	○	○	○	○	○	○	○	○	○
56	1.1	83.0	×	×	×	×	×	×	×	×	×	×	○	○	○	○
55	1.1	84.1	×	×	○	○	○	○	○	○	○	○	○	○	○	○
53	1.0	85.1	×	×	×	×	×	×	×	×	○	○	○	○	○	○
44	0.8	85.9	○	○	○	○	○	○	○	○	×	×	×	×	×	×
43	0.8	86.7	×	×	×	×	×	×	○	×	×	×	×	×	×	×
41	0.8	87.5	×	×	×	×	×	×	×	×	×	×	×	×	○	○
36	0.7	88.2	×	×	×	×	○	○	○	○	○	○	○	○	○	○
36	0.7	88.9	○	○	○	○	○	○	○	○	○	○	○	×	×	×
34	0.7	89.6	×	×	×	×	×	×	×	×	×	×	×	○	○	○
32	0.6	90.2	×	×	×	×	×	×	×	×	○	×	×	×	×	×

Source: MHTS panel data set

Note: See notes pertaining to Table 9.

Table 12 Composition of complete panel households by province

Province	1986–91 panel data			1995–2001 panel data			1986–2001 panel data		
	Nos. of household surveyed			Nos. of household surveyed			Nos. of household surveyed		
		complete panel household			complete panel household			complete panel household	
		Nos.	%		Nos.	%		Nos.	%
Hebei	942	682	72.4%	908	440	48.5%	1,033	389	37.7%
Shanxi	695	616	88.6%	709	626	88.3%	904	451	49.9%
Liaoning	1,168	264	22.6%	439	369	84.1%	671	58	8.6%
Heilongjiang	1,002	301	30.0%	512	263	51.4%	928	103	11.1%
Shanghai	231	173	74.9%	221	179	81.0%	338	94	27.8%
Jiangsu	203	68	33.5%	99	95	96.0%			
Anhui	660	321	48.6%	644	316	49.1%	544	123	22.6%
Shandong	313	213	68.1%	161	79	49.1%			
Hunan	142	118	83.1%	111	90	81.1%	153	85	55.6%
Guangdong	336	278	82.7%	354	263	74.3%	137	55	40.1%
Sichuan	258	242	93.8%	278	219	78.8%	254	155	61.0%
Yunnan	216	194	89.8%	202	198	98.0%	106	94	88.7%
Gansu	68	53	77.9%	76	51	67.1%	53	18	34.0%
Ningxia	235	109	46.4%	124	58	46.8%			
Total	6,469	3,632	56.1%	4,838	3,246	67.1%	5,121	1,625	31.7%

Source: Estimation from MHTS data set

Table 13 Comparison of panel patterns between new ID and original ID

(1)1986–1991 data set

New ID

Nos.	%	Σ (%)	1986	1987	1988	1989	1990	1991
3632	56.1%	56.1%	○	○	○	○	○	○
604	9.3%	65.5%	○	○	×	×	×	×
459	7.1%	72.6%	○	×	×	×	×	×
338	5.2%	77.8%	×	○	○	○	○	○
227	3.5%	81.3%	○	○	○	○	×	×

Original ID

Nos.	%	Σ (%)	1986	1987	1988	1989	1990	1991
4,127	76.2	76.2	○	○	○	○	○	○
352	6.5	82.7	○	○	×	×	×	×
257	4.7	87.4	○	○	○	×	×	×
128	2.4	89.8	○	○	×	○	○	○
110	2.0	91.8	×	○	○	○	○	○

(2)1995–2001 data set

New ID

Nos.	%	Σ (%)	1995	1996	1997	1998	1999	2000	2001
3,246	67.1	67.1	○	○	○	○	○	○	○
261	5.4	72.5	○	○	×	×	×	×	×
230	4.8	77.2	○	○	○	○	○	○	×
161	3.3	80.6	○	○	○	×	×	×	×
155	3.2	83.8	×	×	×	○	○	○	○

Original ID

Nos.	%	Σ (%)	1995	1996	1997	1998	1999	2000	2001
3,924	88.2	88.2	○	○	○	○	○	○	○
99	2.2	90.5	○	○	○	○	○	○	×
53	1.2	91.7	○	○	○	○	×	×	×
44	1.0	92.7	○	×	○	○	○	○	○
42	0.9	93.6	×	×	×	×	×	×	○

(3) 1986–2001 data set

New ID

Nos.	%	Σ (%)	86	87	88	89	90	91	93	95	96	97	98	99	00	01
1,625	31.0	31.0	○	○	○	○	○	○	○	○	○	○	○	○	○	○
390	7.4	38.4	○	○	○	○	○	○	×	×	×	×	×	×	×	×
296	5.6	44.1	×	×	×	×	×	×	×	○	○	○	○	○	○	○
295	5.6	49.7	○	○	×	×	×	×	×	×	×	×	×	×	×	×
249	4.8	54.4	○	×	×	×	×	×	×	×	×	×	×	×	×	×

Original ID

Nos.	%	Σ (%)	86	87	88	89	90	91	93	95	96	97	98	99	00	01
1,755	40.7	40.7	○	○	○	○	○	○	○	○	○	○	○	○	○	○
189	4.4	45.1	○	○	○	○	○	○	○	○	○	○	○	○	○	×
181	4.2	49.3	○	○	○	○	○	○	×	×	×	×	×	×	×	×
98	2.3	51.6	○	○	○	○	○	○	○	○	○	×	×	×	×	×
92	2.1	53.7	○	○	○	○	○	○	×	○	○	○	○	○	○	○

Source: Estimation from 20 percent of the resampling data set of RCFPO and MHTS panel data set

Table 14 Consistency of panel patterns between original ID and new ID

		Households	
		Nos.	%
1986–91 dataset	consistent	4,023	64.2%
	inconsistent	2,246	35.8%
	total	6,269	100.0%
1995–2001 dataset	consistent	3,391	70.1%
	inconsistent	1,448	29.9%
	total	4,839	100.0%
1986–2001 dataset	consistent	2,364	45.1%
	inconsistent	2,880	54.9%
	total	5,244	100.0%

Source: MHTS panel database

Table 15 Attrition rates for longitudinal household survey data in developing countries

Country	Time period	Interval	Attrition rate between rounds (%)	Attrition rate per year (%)
Bolivia (urban)	1995/96 to 1998	2 year	35	17.5
Kenya (rural, South Nyanza Province)	1994/95 to 1996/97	2 year		
couples			41	20.5
men			33	16.5
women			28	14.0
Nigeria		5 year	50	10.0
South Africa (KwaZulu-natal)	1993 to 1998	5 year		
households			16	3.2
preschool children			22	4.4
India (rural)	1970/71 to 1981/82		33	3.0
Malaysia		12 year	25	2.1
Indonesia	1993 to 1997	4 year	6	1.5
China (rural, MHTS)				
1986–2001	1986 to 2001	15 year	69	4.6
1986–1991	1986 to 91	5 year	44	8.8
1995–2001	1995 to 2001	6 year	33	5.5

Source: Alderman et al. [2000] table 1 and our estimation

Table 16 Estimation of attrition probit

	1986–1991		1995–2001	
	coefficient	*z* statistic	coefficient	*z* statistic
Income	0.000	3.57 ***	0.000	4.22 ***
Cultivated area	-0.042	-4.47 ***	-0.069	-4.02 ***
Number of labor	-0.048	-2.21 **	-0.131	-5.43 ***
Member of CCP	-0.029	-0.47	-0.060	-0.86
Cadre	0.035	0.29	-0.011	-0.13
Elementary school	-0.360	-5.83 ***	-0.084	-0.98
Junior high school	-0.455	-6.86 ***	-0.100	-1.15
High school	-0.484	-5.17 ***	-0.091	-0.83
Floor area of housing	-0.001	-1.78 *	0.000	-2.44 **
Constant	-0.734	-7.77 ***	0.542	3.51 ***
Log pseudo-likelihood	-2356		-1913	
Number of obs	5280		4227	
Wald χ^2 (22)	1519.11***		693.43***	
Pseudo R^2	0.28		0.18	

Source: MHTS panel database

*** significant at 1% level, ** at 5% level, and * at 10% level

Table 17 Estimation of income function

	1986–1991		1995–2001	
	coefficient	*t* statistic	coefficient	*t* statistic
Cultivated area	0.038	3.80 ***	0.070	15.42 ***
Number of labor	0.061	7.04 ***	0.051	5.08 ***
Member of CCP	0.008	0.29	-0.083	-2.75 ***
Cadre	0.215	4.34 ***	-0.032	-0.92
Elementary school	0.120	3.76 ***	0.200	4.28 ***
Junior high school	0.142	4.20 ***	0.257	5.48 ***
High school	0.164	3.66 ***	0.215	3.93 ***
Constant	5.510	124.06 ***	7.134	97.22 ***
Number of obs.	5270		4219	
R-squared	0.171		0.320	
F test that all attrition interactions equal to zero excluding the intercept				
F(7, 5241)	5.97***		3.59***	
F test that all attrition interactions equal to zero including the intercept				
F(8, 5241)	7.13***		10.45***	

Source: MHTS panel database

*** significant at 1% level, ** at 5% level, and * at 10% level

Table 18 Result of attrition probit by village

	1986–1991			1995–2001		
	number of significant variables			number of significant variables		
		positive	negative		positive	negative
per capita income	9	6	3	7	4	3
per capita land area	5	1	4	8	3	5
number of labor	6	1	5	12	5	7
CCP dummy	2	2	0	6	3	3
cadre dummy	1	0	1	4	0	4
elementary school dummy	8	3	5	12	8	4
junior high school dummy	12	6	6	8	5	3
high school dummy	5	3	2	4	2	2
area of housing	5	1	4	9	2	7
valid sample size	40			33		

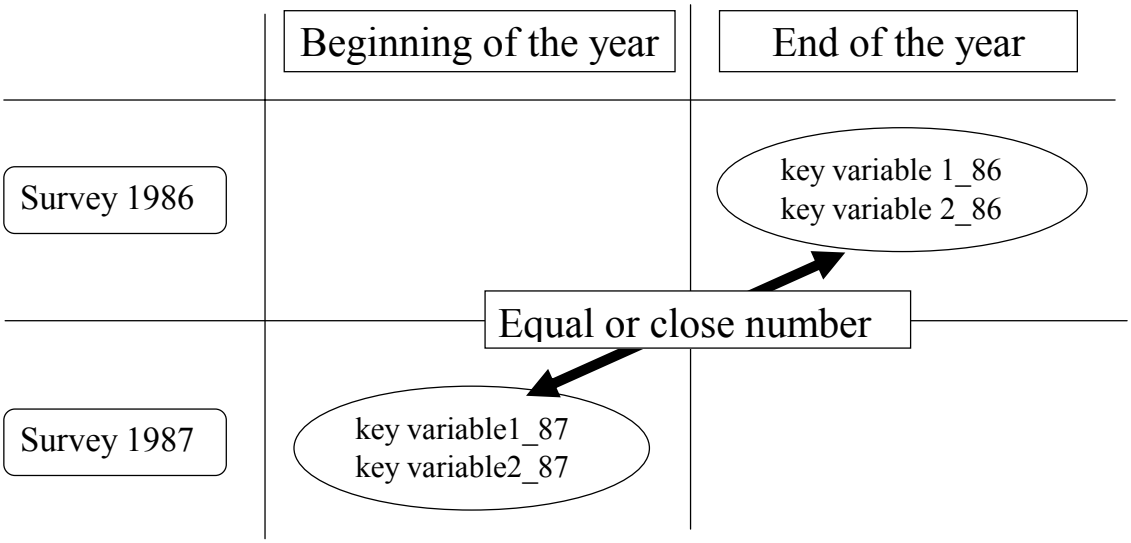
Source: MHTS panel database

Table 19 Summary table on F-test of joint effect of attrition on income by village

	1986–1991		1995–2001	
	F test not including constant	F test including constant	F test not including constant	F test including constant
Significantly rejected	31	34	29	29
Not rejected	15	12	12	12
Total	46	46	41	41

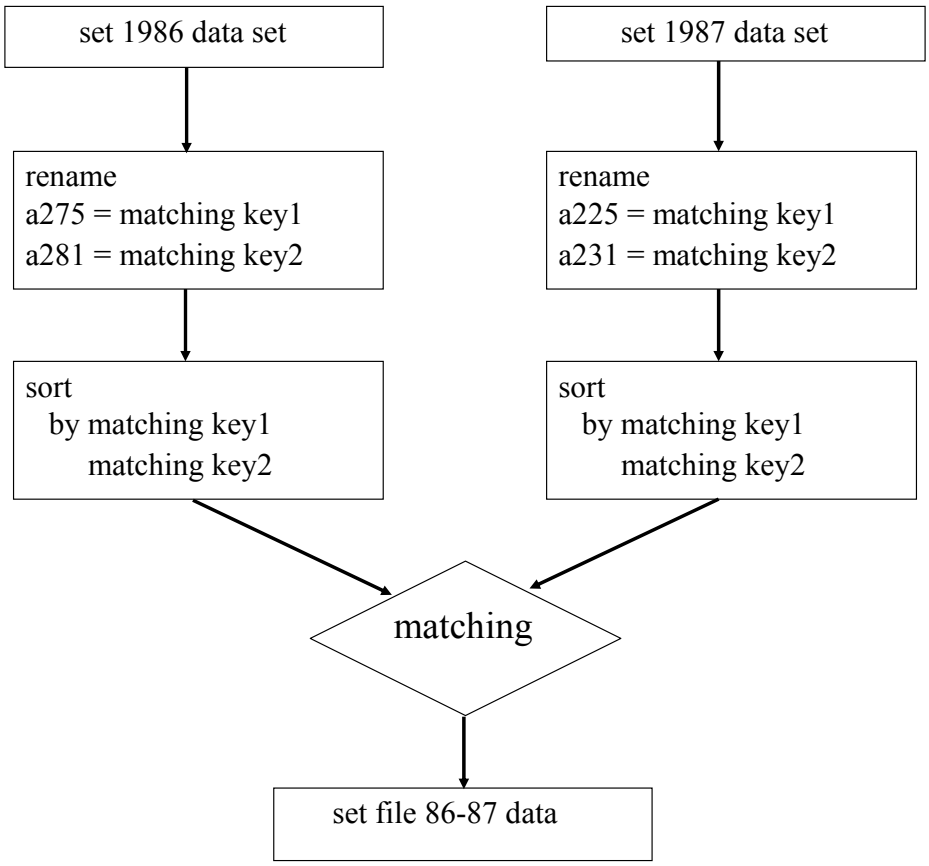
Source: MHTS panel database

Figure 1 Basic concepts of exact matching



Note: The methods of data matching are based on Matsuda ed. [1999].

Figure 2 Algorithm of data matching



Source: Matsuda ed. [1999] (revised by authors)