The Australian National University

Centre for Economic Policy Research

DISCUSSION PAPER

The Reliability of Matches in the 2002-2004 Vietnam Household Living Standards Survey Panel

Brian McCaig

DISCUSSION PAPER NO. 622

September 2009

ISSN: 1442-8636
ISBN: 978-1-921693-03-8

Brian McCaig, School of Economics and RSSS Economics Program, HC Coombs Bldg [9], The Australian National University, ACT 0200 Australia. Email: brian.mccaig@anu.edu.au

Acknowledgements:
I would like to acknowledge useful feedback on the suggested revisions to the panel by Le Dang Trung and Hung Pham.

Abstract

Tracking households and individuals over time is important for a variety of research and policy questions. Exploring the validity of matched household and individuals within a dataset is a necessary step for ensuring the reliability of analysis designed to address such questions. This paper examines the quality of matching in the household panel between the 2002 and 2004 Vietnam Household Living Standards Surveys (VHLSS). Of the original 4,476 matches suggested in the household datasets, 429 matches are found to be incorrect, almost ten percent of total matches. Revised matches are suggested for 402 of these mismatches. Two simple applications illustrate the potential problems associated with analysis conducted using a poorly matched panel. The original panel overestimates the frequency of changes in household size. It also leads to biased estimates of per capita consumption growth, as it overestimates growth for poor households and underestimates growth for rich households.

JEL codes: C8, O12

Keywords: Vietnam Household Living Standards Survey, household panel

1. Introduction

In 2002, the General Statistics Office (GSO) of Vietnam began a series of household surveys, the Vietnam Household Living Standards Surveys (VHLSS), to be conducted every two years between 2002 and 2010. The surveys are very broad in nature, collecting data on education, health, employment, income generation activities, consumption, housing, and participation in poverty alleviation programs. Also included is a particularly useful household panel between the 2002 and 2004 VHLSS.[1] Given the breadth of information available in the surveys and the existence of the household panel, these datasets have proven very popular with researchers in a wide variety of disciplines. For example, a search on the academic article search engine Google Scholar for the term “VHLSS” returned 269 hits. A similar search for “Vietnam Household Living Standards Survey” returned 138 hits. Focusing on the panel, searches for “VHLSS panel” and “Vietnam Household Living Standards Survey panel” returned 113 and 59 hits respectively. No doubt some of these are repeated hits, but the already substantial number of researchers using these datasets and the panel is expected to grow. The VHLSS’s predecessor, the 1993 and 1998 Vietnam Living Standards Surveys (VLSS), has also been extremely popular with researchers. A Google Scholar search returned almost 800 hits for “Vietnam Living Standards Survey” and close to 300 hits when the term “panel” was added. Thus, given the large number of researchers that are presently using, and are likely to use in the future, the 2002 and 2004 VHLSSs, along with the associated household panel, it is important to verify the accuracy of the household matches. Mismatched households are likely to introduce bias into parameter estimates, thereby misleading researchers and policymakers.

The goal of the current paper is to document the results from a review of the 2002-2004 VHLSS household panel. Of the 4,476 households interviewed in the 2004 VHLSS that should have a matching household in the 2002 VHLSS, 429 appear to be mismatched (9.6 percent). This

[1] The household panel actually extends to the 2006 VHLSS as approximately half of the households that form the 2002-2004 household panel were interviewed a third time in 2006.

is clearly a non-trivial fraction of the total panel size. Revised matches are suggested for 402 of these mismatches, leading to a revised panel of 4,449 households.[2]

The remainder of the paper proceeds as follows. Section 2 describes the 2002 and 2004 VHLSS and the verification of the household panel. Section 3 provides two short applications to demonstrate the potential biases introduced by using the original panel. Finally, Section 4 provides some closing remarks.

2. The 2002 and 2004 VHLSSs and the household panel

2.1 Overview of the 2002 and 2004 VHLSS datasets

The 2002 VHLSS had a target sample of 75,000 households (Phung and Nguyen, unknown). Of the total target sample size, 45,000 households were to be asked a short version of the survey (all modules of the questionnaire except for the expenditure module). These households were all to be interviewed in the first two quarters of 2002. The remaining 30,000 households were to be asked the complete questionnaire and to be interviewed in equal instalments throughout each quarter of 2002. The difference in questionnaire, and in timing of the interviews, introduced two significant problems for correctly matching households between the 2002 and 2004 VHLSS. These problems will be expanded on below. In comparison, the 2004 VHLSS had a smaller target sample size of 45,000 households. Of these, 9,000 households would be asked the complete questionnaire, and the remaining 36,000 households would be asked all modules of the questionnaire, except the expenditure module.

To match households between the two surveys the researcher must be able to properly construct a unique household identifier in both surveys and subsequently know how to match these household identifiers. In the 2004 VHLSS, a unique household identifier can be constructed using five pieces of geographic information: province, district, commune, cluster, and a two-digit household number within each cluster. The household identifier for the 2002 VHLSS requires the same five pieces of geographic information plus information on the quarter

[2] A dataset containing the suggested revised panel is available from the author upon request.

of the year in which the household was interviewed in 2002. This is important because the same cluster in 2002 could, for example, have two households with the 2-digit household number “01”. The only way to differentiate these two households is to determine whether they were interviewed in the first or second half of 2002. This turns out to have introduced a substantial number of errors that come into play when trying to match households between the two surveys. Many of the errors in matching are due to one of 6 pieces of identifying information being incorrectly recorded. These mismatches can often be corrected by looking at the identifiers provided for other households in the same geographic area.

A second complication derives from the fact that the GSO only publicly releases the data for households that were asked the full questionnaire in each year. Thus, the publicly released data for the 2002 VHLSS contains 29,530 households and for the 2004 VHLSS contains 9,188 households.[3] Unfortunately, some of the panel households that are part of the publicly released 2004 data are properly matched with 2002 households found in the non-publicly released data. I have had access to the complete dataset for 2002 allowing these matches to be discovered.

2.2 Verification and corrections

The household panel is constructed using the dataset “m1b.dta” from the 2004 data. It contains the information necessary to construct a household identifier for 2004 and the suggested household identifier for 2002. However, as suggested above, the matches created using the original dataset appear to be flawed for a significant number of the matches. In general, there are three types of problems associated with the 2002-2004 panel identifiers as included in the original 2004 VHLSS datasets: (1) more than one household in 2004 survey is matched with the same household in 2002[4], (2) the household identifier in 2002 pointed to by the household in the 2004 survey does not exist, and (3) the suggested match is simply incorrect. In the remainder of

[3] It is unknown to the author why the actual sample size in 2002 was only 29,530 households instead of the intended 30,000 households.

[4] In principle, this could be due to a 2002 household splitting into two smaller households by 2004. However, the VHLSS does not track the division of households in this manner and thus there should not be any 2002 households matched with more than one 2004 household.

this section I discuss the prevalence of these three types of errors, and how often a correct match is found.

Table 1 summarizes the makeup of the original and revised versions of the 2002-2004 VHLSS household panel. According to the original dataset there are a total of 4,476 households from the 2004 survey that were also interviewed in 2002. The original panel suggested a match for all of these households, whereas the revised version of the panel that I suggest only has matches for 4,449 households, meaning that there are 27 potential panel households without a valid corresponding match in 2002. It is unfortunate to be unable to find identifiable matches for these households, but leaving these households unmatched is no doubt preferable than incorrectly matching them. Table 1 also identifies one of the biggest problems for a researcher who does not have access to the complete set of interviewed households in 2002: a large number of the 2004 households are properly matched with 2002 households that are not in the publicly released data. This is likely confusing to many researchers and could lead to incorrect matches as researchers try to use the best available match in the publicly released data.

Table 1: Summary of the number of households in the original and revised 2002-2004 VHLSS household panels

	Original Panel	**Revised Panel**
Number of 2004 households	4476	4476
Number of 2002 household identifiers provided	4476	4449
Number of unique 2002 household identifiers provided	4392	4449
Number of unique and valid 2002 household identifiers provided	4305	4449
Number of unique 2002 households identifiers that appear in the expenditure data	3926	4091

Table 2 summarizes the 429 households (9.6 percent) for which the original and revised versions of the panel disagree. Of these differences, 43 are due to the same 2002 household being matched to more than one 2004 household. I find a valid match for 41 of these households. Secondly, there are 87 instances in which the originally suggested 2002 household does not exist in the 2002 datasets. I am able to find a correct match for 80 of these households. Finally, I identify 299 mismatches. These mismatches are indicated by gender or age mismatches across individuals within a matched household. In principle, the gender or age of a household member may not match for a variety of reasons: there may have been a recording error in one of the surveys, the household may be a correct match but the matching of individuals within the household is incorrect, or the mismatch is indicative of a bad household match. I visually inspect all such instances to try to determine the reason for the mismatched gender or age, and thus the corrective action needed. I find valid matches for 281 of these households, but am forced to leave 18 of these households unmatched.

Table 2: Summary of differences between the original and revised versions of the 2002-2004 VHLSS household panels

	Differ	Fixed	Missing
All differences	429	402	27
Duplicated 2002 household identifiers	43	41	2
Invalid 2002 household identifiers	87	80	7
Bad matches	299	281	18

A quick analysis indicates that the probability of finding a valid match is not influenced by a dummy variable for being urban in 2004, the household size in 2004, or total household expenditures in 2004. Although this is not definitive proof of random attrition, it is at least consistent with it, and thus should not introduce attrition bias.

Of course, there is no definite way to conclude that a particular original match is "incorrect" and that the new "correct" match has been found. However, supporting evidence is readily available. One simple way is to examine outcomes for three different groups of households: (1) households with original matches that are deemed to be correct, (2) households with original matches that are deemed to be incorrect, and (3) households with revised matches. In particular, I graph the rank of per capita expenditure in 2004 versus the rank of real per capita expenditure in 2002 for each group of households. The associated plots are shown in Figures 1 through 3. Two stylized facts quickly emerge from the three plots. First, for the households with matches I deem incorrect in the original panel, there is much greater dispersion in the plot relative to the correct households in the original panel. Given the short time period, only 2 years, one would naturally expect a fair amount of persistence in a household's rank in the per capita expenditure distribution. The corollary of this is that incorrectly matched households should show greater movement within the per capita expenditure distribution. This is exactly what is seen. The second key fact to emerge from the figures is that the dispersion shown for corrected matches is much more similar to the dispersion for correct households in the original panel. Moreover, the correlations in the plots are very telling: 0.79 for correct matches in the original panel, only 0.53 for the incorrect matches in the original panel, and 0.80 for the corrected matches in the revised panel. Thus, the corrected household matches exhibit extremely similar dynamics to the correct household matches in the original panel, whereas this is noticeably less true for the incorrect household matches.

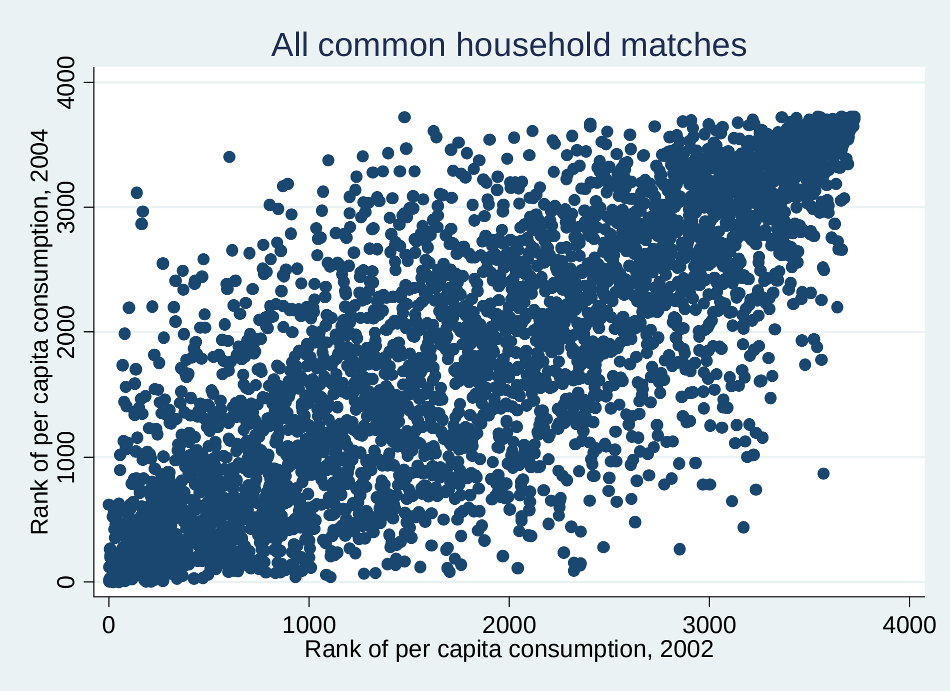

Figure 1: Plot of rank in per capita consumption distribution in 2004 versus 2002 for all household matches that are common to the original and revised versions of the 2002-2004 VHLSS household panel

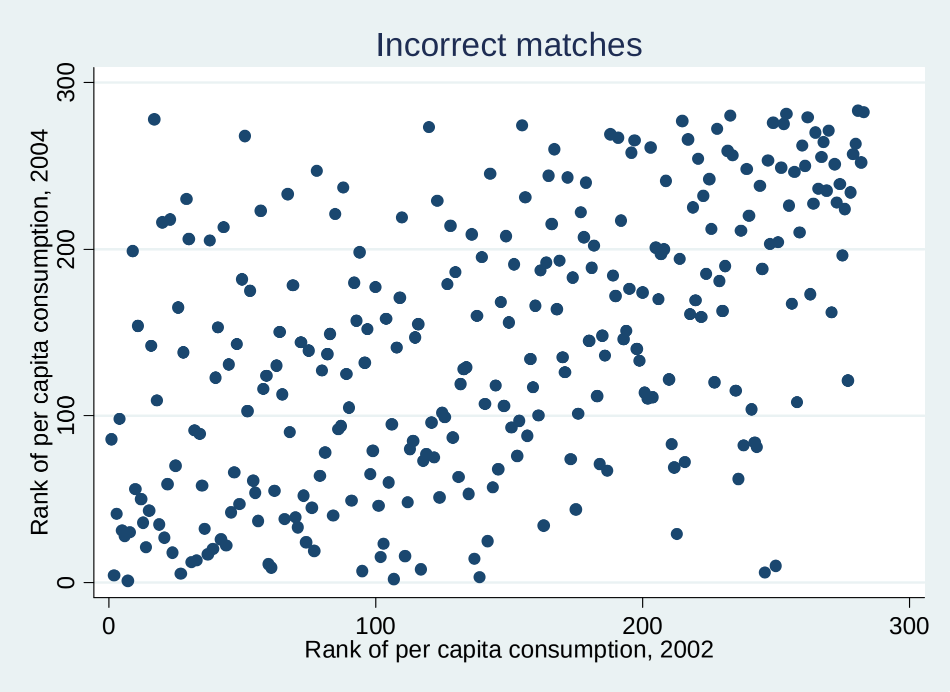

Figure 2: Plot of rank in per capita consumption distribution in 2004 versus 2002 for original household matches deemed incorrect

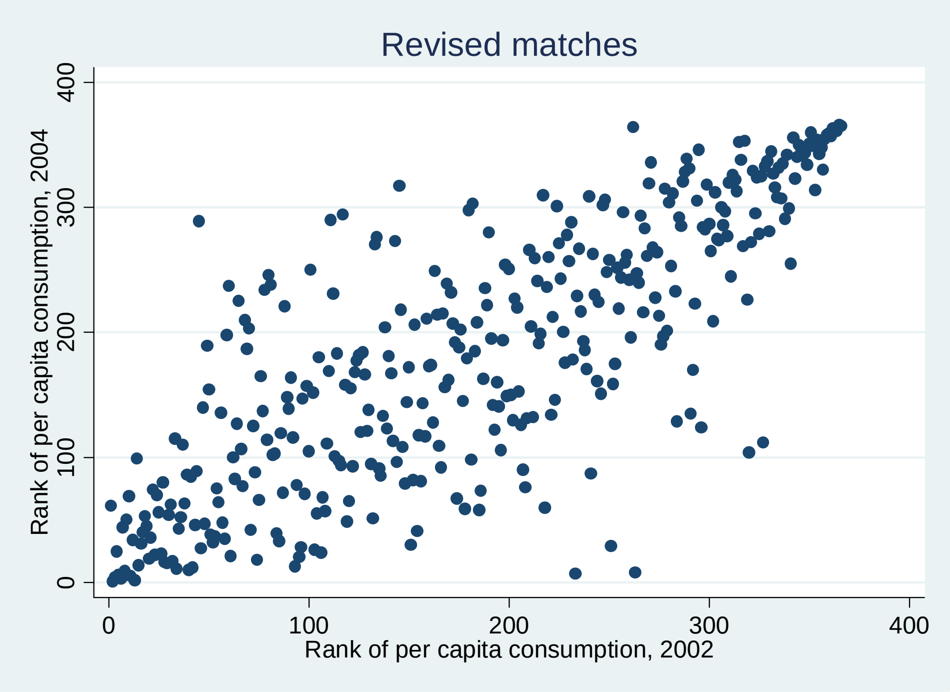

Figure 3: Plot of rank in per capita consumption distribution in 2004 versus 2002 for revised household matches

3. Implications of mismatched households for various household dynamics

In this section, I explore the impacts of incorrect matching of households on the analysis of changes in household size and growth of per capita expenditures. For each dynamic outcome, I show results based on the original panel, the revised panel, and the subset of households for which the two versions differ.

Table 3 shows the evolution of household size between 2002 and 2004. The first row shows the changes in household size between 2002 and 2004 for all households in the original panel for which the suggested 2002 household identifier exists (i.e., information on household size exists in both 2002 and 2004). A total of 2,602 out of these 4,389 households did not change size between 2002 and 2004. This represents 59.3 percent of the households. By comparison, using the revised version of the panel suggests that a larger share of households, 2,791 out of 4,449, did not change household size between 2002 and 2004. The direction of the difference is what one would expect, as some changes in household size are due to mismatching households of different sizes. The third row reports on the household matches that are common to both the original and revised panel. The extent of the problem is more clearly seen by focusing on the subset of households for which the original and revised versions of the panel differ, shown in the fourth and fifth rows. The fourth row, based on the 430 households for which the two versions of the panel disagree, and for which there is information on the household size in 2002 for the originally suggested match, shows that a much smaller share of households remained the same size between 2002 and 2004; only 70 of 342 households, or 20.5 percent. By comparison, using the revised version of these matches suggests that 259 out of 402 households remained the same size, or 64.4 percent. Hence, for the subset of households for which the original and revised versions of the panel differ, there is a substantial upward bias in the share of households that changed size. Any research on household formation and migration using the original household panel would clearly suffer from the bias induced by incorrect panel matches. A final noteworthy result from Table 3 is the consistency of the share of households with no change in size using the revised panel. When using all households in the revised panel, 62.7 percent of households did not change size as compared to 64.4 percent when using only the households in the revised panel that differ from those suggested in the original panel. In contrast, the same comparison for the original panel shows that 59.3 percent of all original panel households did not change size against only 20.5 percent of original panel households that differ from the revised panel. This is strong evidence of measurement error in the original panel.

Table 3: Summary of changes in household size between 2002 and 2004 using the original and revised versions of the 2002-2004 VHLSS household panels

		Change in household size between 2002 and 2004		
Sample	**Number of households**	**Decrease**	**No change**	**Increase**
Original panel	4389	1003	2602	784
Revised panel	4449	937	2791	721
All common household matches	4047	858	2532	657
Original panel - only panel differences	342	145	70	127
Revised panel - only panel differences	402	79	259	64
		Shares		
Original panel	1.000	0.229	0.593	0.179
Revised panel	1.000	0.211	0.627	0.162
All common household matches	1.000	0.212	0.626	0.162
Original panel - only panel differences	1.000	0.424	0.205	0.371
Revised panel - only panel differences	1.000	0.197	0.644	0.159

A very common use for the VHLSS datasets is to explore changes in and determinants of a household's standard of living, usually measured using per capita consumption. Thus, I now focus on the impact of mismatched panel households on changes in per capita consumption.

Table 4 explores the implications of mismatched households for growth in per capita consumption between 2002 and 2004. It displays the same five groups of households as in Table 3. The first row displays mean per capita consumption in 2002 and 2004, as well as the per annum growth rate, for all households in the original panel for which there is information on per capita consumption in 2002. Similarly, the second row displays the same information for all households in the revised panel for which there is information on per capita consumption. Both the levels and the growth rates are very similar. Rows four and five show the same information, but this time only for the subset of households for which the two versions of the panel differ.

Here substantial differences are exposed. First, the rate of growth for mismatched households is much lower, at 7.69 percent per annum, than for correctly matched households, at 11.10 percent per annum. Second, the level of per capita consumption among mismatched households in 2002, 3495 thousand dong, is noticeably higher than that for the entire original panel, 3347 thousand dong. This is largely responsible for the observed difference in growth rates between these two sets of households as the level of per capita consumption is almost identical in 2004. Hence, the measurement error introduced by mismatched households would lead to a substantial underestimate of the rate of growth for the mismatched households.

Table 4: Summary of changes in per capita consumption between 2002 and 2004 using the original and revised versions of the 2002-2004 VHLSS household panels

		Per capita consumption (Jan 02 prices, 000 VND)		
Sample	**Number of households**	**2002**	**2004**	**Per annum growth rate**
Original panel	4008	3347	4053	10.04%
Revised panel	4091	3338	4061	10.29%
All common household matches	3725	3336	4053	10.22%
Original panel - only panel differences	283	3495	4053	7.69%
Revised panel - only panel differences	366	3355	4141	11.10%

Note: All observations are weighted by 2004 household size.

On average, using the original version of the panel yields a very similar estimate of per capita consumption growth as the revised panel. Given the differences in household size documented in Table 3, this is perhaps a bit unexpected. However, mean growth of per capita consumption is liable to hide important ramifications of mismatched households, due to matching errors cancelling out on average. I turn to this next. Table 5 shows growth in per capita consumption by quartile using the original and revised versions of the household panel. The quartiles are defined based on per capita consumption in 2002 for each sample. In comparison to the results presented in Table 4, the impact of mismatched households on estimates of per capita consumption growth is more pronounced. For households in the first quartile in 2002, the original panel overestimates growth in per capita expenditures by 0.85%. For the second and third quartiles the impact of incorrectly matched households is minor. For households in the fourth quartile in 2002, the original panel underestimates growth in per capita consumption by

0.94%. This pattern can be explained based on the following intuition. Suppose a mismatched 2002 household has been randomly matched with a household in 2004. If this is the case, then on average, the magnitude of the difference between the true and incorrectly matched value of 2004 per capita consumption should be greatest for 2002 households that are farthest from the mean – the households in the first and fourth quartiles – and least for 2002 households that are relatively close to the mean – the households in the second and third quartiles. On average, the measurement error for households in the first and fourth quartiles cancel out, leading to the minor differences in mean growth of per capita consumption shown in Table 4 between the two versions of the household panel.

Table 5: Summary of changes in per capita consumption between 2002 and 2004 using the original and revised versions of the 2002-2004 VHLSS household panel based on 2002 quartiles

		Per capita consumption (Jan 02 prices, 000 VND)		
Sample	**Number of households**	**2002**	**2004**	**Per annum growth rate**
		Quartile 1		
Original panel	1002	1527	2106	17.43%
Revised panel	1023	1528	2077	16.58%
		Quartile 2		
Original panel	1002	2332	3040	14.18%
Revised panel	1023	2334	3042	14.17%
		Quartile 3		
Original panel	1002	3329	4100	10.98%
Revised panel	1023	3321	4073	10.75%
		Quartile 4		
Original panel	1002	6887	7684	5.63%
Revised panel	1022	6889	7824	6.57%

4. Conclusion

In this paper I document errors in the matching of households for the 2002-2004 VHLSS household panel. I find three types of errors in the matching: (1) the same 2002 household being matched to two households in 2004, (2) the suggested 2002 household identifier does not exist in the 2002 datasets, and (3) mismatched households detected on the basis of gender and age discrepancies. Based on visual inspection, I make changes to matches for 429 households out of 4,476. I find new, valid matches for 402 of these households, but am forced to declare no valid match for the other 27 households.

I demonstrate that the measurement error introduced by the incorrect matches biases estimates of changes in household size and estimates of per capita consumption growth. The bias in estimates of per capita consumption growth is larger for 2002 households at the bottom and top of the per capita consumption distribution. The matching errors are likely to influence a variety of dynamic issues that could be addressed using the 2002-2004 VHLSS panel. These include, but are not limited to, changes in labour force participation, changes in health status, analysis of migration decisions, and changes in cropping patterns. It is my hope that by highlighting concerns with the original panel that the community of researchers using the VHLSS panels can benefit from this work.

The matching errors in the 2002-2004 VHLSS household panel will also contribute to mismatches in the 2002-2004-2006 VHLSS household panel. A similar verification process for the 2004-2006 VHLSS household panel suggests that the match quality is much higher. Out of 4,298 possible household matches only 112 incorrect matches are identified and 101 correct matches are found.[5] It is clear that most of the matching error in the 2002-2004-2006 VHLSS household panel is found in the 2002-2004 component of the panel dataset.

It is my hope that this short paper can contribute to improved analysis using the 2002-2004 and 2002-2004-2006 VHLSS household panels. Moreover, I anticipate that this article can also serve as a reminder of the importance of checking the reliability of data before proceeding with analysis.

5. References

Phung Duc Tung and Nguyen Phong. (unknown). “Vietnam Household Living Standards Survey (VHLSS), 2002 and 2004: Basic Information.” Vietnam General Statistics Office.

[5] A revised version of the 2004-2006 VHLSS household panel is also available from the author upon request.