



Thailand Vietnam Socio Economic Panel

Data Cleaning Guidelines

2017 (7th wave)

Rattiya S. Lippe

Mai Thanh Tu

Somkid Naprom

Introductory Note (by HW):

This is an “open” document which was developed on the basis of initial experience with checking test questionnaires. The points listed here are no more than “entry points”. Each DCA will add his/her experience to it and we invite you to add your suggestions. For this we will open a DCA skype group where observations can be posted and ad hoc meetings can be called for.

Remember that you must only check the interview the day after it took place, and only if it was completed on the date of the interview.

Thank you for your time and effort in helping us to check the questionnaires of the 2017 TVSEP wave.

I. First steps:



Look at the task bar elements for a first overview.

- a) Enabled: This shows you how many questions were enabled during the interview (e.g. not counting any questions that were skipped using skip commands)
- b) Unanswered: This shows how many questions in total were left unanswered. Some of these cases may not be feasible. This you could check through.
- c) Invalid: Ideally, there should not be any entries here. Invalid answers are answers that go against plausibility checks in the tablet e.g. If someone went to school for 10 years started school at the age of 7 and left at the age of 15
- d) Commented: This shows for which variables the enumerator made comments. Usually this would signal that something is out of the order in this question or that something special applies and this should be checked.
- e) Flagged: When checking in detail through the questionnaire it may be a good idea to briefly skim through the questions first and mark any that you find strange (e.g. flag). Once you have checked through you can then click on this button and it will show you all of the cases that are perhaps errors.
- f) Hidden: not relevant.

II General Hints:

- 1) Is the order of IDs correct?
- 2) Are all rosters that are created answered
- 3) For monetary values where the answer is zero (e.g. Section 2.4 if no transfer was made) a value of 0 should be added. Entries of “00” or “000” are wrong.
- 4) The exception is when the dummy question that allows drop down menus to work at the end of some sections does not disappear.
- 5) Always consider if a missing answer or out of range value can be checked by the enumerator through phone call before you decide to make a comment, (e.g. this is not possible for most of the behavioural questions)

- 6) Please do not make comments because you find something especially interesting in the interview and you wanted to know more about it. Only comment on those cases that obviously reduces the quality of the interview (please avoid flooding the enumerators with comments).

For non-Thai/VN language DCAs: whenever an enumerator makes a comment (other than Code 90, specify!) which you judge to be important for drawing a conclusion you may first google translate the comment. If the translation does not make sense you may copy it to the DCA Skype group for the nearest native speaking DCA to answer and check if translation makes sense.

III. Detailed Hints:

Section 1:

- crosscheck if the QID matches the geographical location of the household,
- check if both GPS functions (tablet and GPS logger) were entered. At least one should be there (e.g. in most cases GPS logger should at least be there)

Section 2.1 - 2.3:

- Check that the IDs and Names of respondents match (e.g. compare with HH information e.g. number of IDs and Names must match!).
- Check information by HH member separately.
- In section 2.1 variables especially check if telephone number was entered (at least for adult household members, i.e. > 16).
- In section 2.2 check for the duration of schooling, year started/year finished to assess plausibility. Check if the member was ever absent for more than one year.
- In section 2.3 a quick plausibility check for height and weight is to use BMI calculator (weight in kg/height in cm²). Don't do for very young children.
- Check answers to the "disability questions" in section 2.3 especially carefully, e.g. only entries should exist for disabilities which are selected in the multi-select question.

Section 2.4:

- Check plausibility of number of events and sum of payments do not seem to be out of place.
- Check if any amount has not already been occurring in 2.1; in section 2.4. Its only payments should be for non-household members!

Section 2.5:

- Since these questions are subjective assessments by the respondent we cannot question them and enumerator cannot go back or call. Nevertheless make a comment in case of severe inconsistencies, e.g. "not a religious person" but visit temple every week.
- You may cross-check with section 2.1 for question about religion and comment

Section 8:

- Conduct rough per capita consumptions calculations to check plausibility
- If no rice (or other staples) is purchased compare with section 4.2, home consumption.

Section 3.1:

- Check loss figures for assets income and extra expenditures if these are plausible for the stated shock event. Compare also with assets section.
- For income loss you may cross-check with section 2.1,5 and 6 if it is feasible that this income was lost due to someone not being able to work
- You may also check with 4.2 if total production is low relative to land area or 4.3 loss of livestock
- **Section 3.2:**
- Check complete entries in 32002; must be answered in any case whether or not the risk is expected; every risk roster must be answered.
- Compare with shock section; if many shocks but no expected risks respondent/enumerator may be lazy...

Section 4.1:

- Homestead must always be in roster 1 and some answers are fixed e.g. v13 always = 0 and v14 = 99.

Section 4.2:

- Check if each crop was allocated to a plot of land listed for agricultural production in section 4.1.
- Check if the area planted is less than or equal the total area of land for the plot on which it is grown. If several crops are grown on the same plot in different periods, i.e. planting/harvesting time differs or plot is split for a second or third crop sum up the total area planted and compare with plot size.
- For multiple cropping, i.e. wet season and dry season rice do not sum up but compare each crop with plot size
- In the case of intercropping, e.g. two crops are grown on the same plot at the same or similar time e.g. maize and beans or maize and cassava both the main and the intercrop may be entered in different rosters if the second crop is considered to be important (see enumerator manual). That may also mean that all inputs were allocated to the main crop; except seeds and labour for planting and harvesting
- Check the units of production and its conversion to kg/t to assess plausibility. Respondents might state total production in local units and disposal of harvest could be in kg.
- Crosscheck total production by summing disposal (e.g. sales, consumption, processing, etc. ...).
- Check plausibility of person hours are plausible, compare with area planted, e.g. a 1000 labor hours for fertilizer application of 500 sqm is too high.
- For family labour comparison with section 2.1, can be made, e.g. days present in household; stated occupation etc.

No need to spend lots of time for detailed calculations, just do rough calculations. STL should sweep most of these errors because most of them know the production conditions

Section 4.3.1:

- Check opening and end value and compare with changes; generally end value less opening value equal changes. The exception is silkworm/silk production and other .
- Have look at labour inputs if they are very high. Labour for some livestock are hard to estimate by the respondent, e.g. feeding 10 chickens every day and chat with them is leisure not labour!

Section 4.3.2:

- Check if the labour in person hours seems plausible.

Section 5:

- Cross check member occupation in section 2.1.
- Crosscheck the years of job experience with the age of the household member in section 2.1.
- Variables 28-29 should be plausible in terms of values and can be compared with the salary received.
- Check whether the number of times the HH member returned home matches the information filled in Q16. Check whether the number of months worked per job in the reference period matches with the answer for Q16.

Section 6:

- Similar to section 5.

Section 7.1C:

-

Section 7.1D:

- Check that the payment was indeed paid late (e.g. v14a should not be 0).

Section 7.1A:

- Check monetary values and interest rates. If loan was made for an investment check with section 6.2.
- If loan was used on agriculture check size of crop land, type of crops (e.g. rubber plantations etc., large number of livestock
- Many other similar checks....."you will find out"

Section 7.1B:

-

Section 7.1E:

- Check if all sub-rosters are completed even if no bank accounts there must be an entry in V13.

Section 7.2:

- Check that the recipient seems to be valid in terms of characteristics e.g. age in section 2.1.

Section 10:

- check only completeness, do not question the answers.

Section 6.2:

- Check if investments made in prior sections or as revealed from asset section (section 9.1) are found here if made in the reference period.
- Check also plausibility of investment volume against the size of household's resource endowments, income and amount borrowed, (e.g. a 1 ha Thai rice farmer can afford to buy a Mercedes....unless there was some (may be foreigner who has send money in section 2.4).

Section 9.1:

- Check completeness, e.g. everything that was selected should also be answered in the rosters that are generated.

Section 9.2:

- Check with the homestead in section 4.1 (e.g. check if size house > land).
- House improvements that are investments should be listed in section 6.2 if amount is above 5000 THB/1,5 mill VND threshold.

Section 11:

- For quality of interview check total answers missing
- Check for high frequency of 98 "no answer"
- Look at enumerator notes (V110018) where enumerator indicate problematic parts of the interviews and then re-visit corresponding section.