# Users' Manual for Handling Resampled Micro Data of Vietnamese Household Living Standard Survey (VHLSS) 

## VHLSS - Overall and Survey Process

2015
The Institute of Statistical Mathematics (ISM) and

History of revision of the manualVersion 1.1 in February 2015

- Revised based on the discussion during the Sixth International Workshop on Analysis of Micro Data of Official Statistics in December 2014
- Added chapter 6.8 EducationFirst draft version 1.0 in September 2014


## CONTENTS

## VHLSS - Overall and survey process

1. About this Manual ..... Page 3
2. Outline of VHLSS
2.1 Objective of the survey ..... 4
2.2 Topics covered by the survey ..... 4
2.3 Coverage of the survey ..... 5
2.4 Sample design ..... 6
2.5 Data collection and editing ..... 10
2.6 Data processing ..... 11
3. Data and metadata provided
3.1 VHLSS 2002/2004/2006 ..... 12
3.2 Other source of metadata ..... 33
4. Map of provinces and regions ..... 38
5. Panel data (Trial) ..... 40
6. Population and household
6.1 Ethnicity ..... 66
6.2 Religion ..... 68
6.3 Age heaping ..... 69
6.4 Literacy ..... 71
6.5 Polygamy ..... 74
6.6 Relationship to the household head ..... 78
6.7 Family type
6.7.1 Family type based on the number of persons by generation ..... 84
6.7.2 Family type based on the generation and number of married couples ..... 94
6.8 Education ..... 100
7. Agricultural land ..... 103

## 1. About this Manual

1. This manual was prepared for users to use the next $80 \%$ resampled micro data sets of Vietnamese Household Living Standard Survey (VHLSS). It described mainly overall and survey process of VHLSS 2002, 2004 and 2006.

| Survey year | Data files | Contents |
| :--- | :--- | :--- |
| VHLSS 2002 | 31 files | Household questionnaire for Income and Expenditure Survey |
| VHLSS 2004 | 47 files | Household questionnaire for Income and Expenditure Survey |
| VHLSS 2006 | 49 files | Household questionnaire for Income and Expenditure Survey |

2. The questionnaires, file layout and data dictionary of each VHLSS are described in the manual for each year's VHLSS, respectively.
3. The original micro data sets composed of all the samples were provided by NSO, Viet Nam based on the Charter for Experimental Laboratory for Research Purpose Statistical Use of Micro Data, and resampled at the rate of $80 \%$ by Sinfonica.
4. This manual was first compiled in September 2014 by;

Hiroshige Furuta
Visiting Senior Research Fellow, Sinfonica

## Acknowledgements

Special thanks to Ms. Van Nghiem, GSO, Vietnam, who assisted my work of compiling the manuals by properly answering to my queries via email.

## 2. Outline of VHLSS

This chapter mainly describes on VHLSS 2002, 2004 and 2006.

### 2.1 Objective of the survey

In order to evaluate the living standards for policy and socio-economic plan making, the General Statistics Office (GSO) has conducted many household living standards surveys. From 2002 to 2010 particularly, the household living standards survey (VHLSS) is conducted every 2 years by the GSO in the years end with even numbers in order to monitor systematically living standards of Vietnam population's groups; monitor and assess the implementation of the Comprehensive Poverty Reduction and Growth Strategy; making contribution to evaluating results of realization of the Millennium Development Goals (MDGs) and Vietnam Development Goals (VDGs).
The GSO publishes the official results of the VHLSS. Data of other statistical areas compiled from the VHLSS are not for replacing data of these areas which were published before rather than for clarifying and further analyzing factors affecting to living standards.

VHLSS collects information to be used as basis for assessment of living standard, including poverty and the gap between the rich and the poor serving for policy making, planning and national targeted programs of the party and the State in order to continuously improve the living standard of population across the country, in all regions and localities.
In addition to that, information is collected to serve for research, analysis of some topics on health, education, employment and to calculate weight to compile consumer price index and national account.

### 2.2 Topics covered by the survey

VHLSS includes main content reflecting living standard of households in the entire country, and main socio-economic conditions of communes in the rural areas which affect living standard of population in their particular area.

VHLSS consists of two surveys; for household and for commune.

### 2.2.1 For households

- Some demographic characteristics of household's members including age, sex, ethnicity, marital status.
- Household’s income includes: income level; income by different sources (salary, wage; selfagricultural, forestry, fishery production; self-household’s business production/service; others);
income by economic sector and industry.
- Household's expenditure: expenditure level, expenditure by purpose and item (expenditure on food, clothes, accommodation, travel, education, health, culture, etc... and others.
- Education level, professional level of each household's member.
- Illness and use of type of health clinic.
- Employment status, working hours.
- Assets, housing and facilities, such as appliances, electricity, water, sanitation condition.
- Participation in hunger elimination and poverty reduction, credit status.

If budget is available, extended content on education and health will be surveyed.
2.2.2 For communes which have surveyed households

- Some information on demography, ethnicity.
- Basic socio-economic infrastructure including: situation of electricity, roads, schools, medical stations, markets, post offices, water sources.
- Economic situation, including: Agriculture production (land, production increase/decrease tendency of some main crops and causes, assistances for production development such as irrigation, agricultural encouragement); non-agricultural job opportunities.
- Some main information on social order and safety.

Note: Micro data of the survey for commune was not provided.

### 2.3 Coverage of the survey

2.3.1 VHLSS covers the whole country. Scope of the survey includes all selected enumeration areas and communes in 64 provinces and cities (in VHLSS 2004 and 2006) under central management (hereafter called province/city).

### 2.3.2 Target population

The target population comprises the civilian, non-institutional population.
Only persons considered as permanent residents are eligible for inclusion in the survey.
The following persons are included;
-persons on vacation
-temporarily in a hospital
-students living away from home during the school year
The following households are ineligible;
-housing unit containing only students living away from home during the school year -military compounds
-hospitals
-prisons

### 2.4 Sample design: three-stage stratified cluster design

Sample of the VHLSS is selected in the way to represent the entire country (in which: urban/rural areas), 8 regions (in which: urban/rural areas), and provinces/cities.

## Master sample

Two-stage area sample from enumeration areas of the 1999 Population and Housing
Census.
Strata: province and urban/rural

## PSU (primary sampling unit): communes

Out of 10,476 communes as of 1999 Population Census, 3,063 communes/words were selected.
Communes contain on average about 1,600 households while EAs were set up during the censuses to contain about 100 households.

## SSU (Secondary sampling unit): EAs

Three EAs were selected per commune.
Only one of EA was used for each year of VHLSS survey.
This is technically a three-stage design counting the selection of households, but it is operationally equivalent to a two-stage design.

## Sample allocation

The sample was allocated over strata proportional to the square root of the number of households.

Both communes and EAs were selected with probability proportionate to size (PPS), the size being the number of households according to 1999 Population Census.

The master sample is designed specifically for VHLSSs in the period 2002-2010. It contains 3,063 communes/wards, and 3 EAs are selected in each commune/ward from EAs of the Population and Housing Census 1999.

Figure 1 Structure of VHLSS 2006

| Type of survey | Target and size (2006) | Questionnaire/Questions | Objective |
| :--- | :--- | :--- | :--- |
| Commune survey | 3,063 commune/wards <br> which have surveyed <br> households | Basic socio-economic <br> information |  |
| Income survey | 36,756 households | (Short) Questionnaire 1A; <br> Questions on income | To assess living <br> standards at <br> provincial/city level |
| Income and <br> expenditure <br> survey | 9,189 households | (Long) Questionnaire 1B; <br> Questions on income and <br> expenditure | To assess living <br> standards at national <br> and regional level |
| In total | 45,945 households |  |  |

## Sample size

Sample size of the VHLSS 2006 includes 45,945 households selected from 3,063 areas of the master sample frame, and is divided into 2 types: Sample for income survey includes 36,756 households to collect information as mentioned above, excluding household's expenditure, to assess living standard at national level, regional and provincial/city level; Sample for income-expenditure survey includes 9,189 to collect sufficient information for further assessment and analysis of living standard at national and regional level.

Sample for income survey and sample for income-expenditure survey was divided for two rounds are as follows:

| Time for <br> Data collection | Survey on income <br> \& expenditure | Survey on <br> income | Total |
| :--- | :--- | :--- | :--- |
| Total <br> Of which: | 9.189 <br> (households) | 36.756 | 45.945 |
| May 2006 | 4.593 | 18.372 | 22.965 |
| September 2006 | 4.596 | 18.384 | 22.980 |

## Sample selection for each year of VHLSS: 50\% rotation of households

The Socio-Environmental Statistics Department coordinates with provinces/cities to select sample as follows:

## Step 1: Select EAs.

EAs of the VHLSS 2006 will be selected rotationally, specifically: re-select 50\% enumeration areas of the VHLSS 2004 (in which half of the areas were surveyed in the 2002 and 2004 VHLSSs and another half of the areas were only surveyed in the 2004 VHLSS) and the other $50 \%$ areas will be newly selected from the master sample, which were not selected in the 2002 and 2004 VHLSSs.

Figure 2 Selection method of EAs


Note:
in VHLSS 2006 were surveyd in 2002 and 2004. in VHLSS 2006 were surveyd only in 2004.

## Panel data

The sample design enables to make use of panel data. The questionnaire of VHLSS 2004 and 2006 included questions about the identification code of the previous VHLSS, which were recorded by the enumeration team leader.

Remarks:
Identifiers of VHLSS 2002 differ from the later VHLSS, as follows.

| VHLSS 2004/2006 |  | VHLSS 2002 |  |
| :--- | :--- | :--- | :--- |
| tinh | 3 | tinh |  |
| huyen | 2 |  | ха |
| xa | 2 |  |  |
| diaban | 3 |  |  |
| hoso | 2 |  |  |

## Updating of sample frame of households

The Socio- Environmental Statistics Department is responsible for selecting areas and sending the list of selected areas to Provincial Statistics Offices for reviewing and updating attached with the map and list of areas of the 1999 Population and Housing Census of the new areas.
Provincial Statistics Offices review and propose to change some areas for more suitability with geographical, socio-economic characteristics of provinces with less than the change of $5 \%$ of total number of provinces' areas with an agreement of the GSO (the SocioEnvironmental Statistics Department) during the year prior to implementing the survey.

## Step 2: Select households.

Provincial Statistics Offices select households, specifically:

- For areas which are re-selected from the 2004 VHLSS, select all 15 households in which 12 households were already surveyed with income (income households) in 2004 to be surveyed with income for the 2006 VHLSS and 3 households were already surveyed with income-expenditure in 2004 to be surveyed with income-expenditure for the 2006 VHLSS. In case of households which were surveyed in 2002 or 2004 moved from the area, find alternate households to be assured of 12 households for income and 3 households for income-expenditure in each enumeration area.
- For new areas, select 20 households from the updated household's list. From these 20 households, select 15 households (12 official households, 3 spare households) for income survey, the 5 remaining households ( 3 official and 2 spare households) for income-expenditure survey.

Household selection follows the methods in the 2006 VHLSS manual.
Provincial Statistics Office will equally divide EAs by urban/rural areas and geographical region for the two rounds of the surveys conducted in May and September.

Communes which have selected EAs for the household interview will simultaneously interview commune questionnaire.

The list of selected households will be kept in two places: PSOs and Socio-environmental Statistics Department for the implementation, monitoring, and supervision.

## Remarks:

Sample design was changed based on the results of analysis of VHLSS 2002. In VHLSS 2002, the cluster size (the number of sample households within EA ) was 25 households. The cluster sizes for the expenditure module in VHLSS 2004 and 2006 were smaller, 3 households per EA as compared to 5 or 20 households per EA in VHLSS 2002.

## Definition of urban/rural

The definition of urban/rural is not clear.
There are three types of local administrative unit; ward, district town and commune. They are classified into urban and rural.

Urban: ward and district town
Rural: commune

Note: The area of urban/rural might vary year by year.

### 2.5 Data collection method

## Field work

The field work was done about one month for each round of the survey.
In VHLSS 2002, it was implemented in 4 quarters (starting at first month of each quarter). Each long questionnaire was completed within one and a half day, and short questionnaire was completed within one day.

In VHLSS 2004, the survey was implemented in May and September. The long questionnaire (including two special modules) was often completed within 2.5 days and the short is completed in one and a half.

## Questionnaire

This survey uses two types of questionnaire: questionnaire for commune and questionnaire for household, as in Figure 1.

The questionnaire for household includes: income-expenditure questionnaire (applied for the sample of the income-expenditure survey) which contains all information of the survey; and income questionnaire (applied for sample of the income survey) which contains all information of the survey excluding information on household's expenditure. Questionnaires are designed in details so that it will facilitate the recording of interviewers, help avoid the missing of items and increase the consistency among interviewers. Thus the survey data will be improved.

Direct interview is applied in this survey. Interviewers visit households, meet with heads of households and related households' members to interview and fill information in questionnaires for households.

### 2.6 Data processing

1) The data is first checked by team leader before sending to PSO.
2) The PSO checks the data again and send it to data entry staffs.
3) The data entry staffs uses CSPro for entering data twice and data cleaning. The program gives a warning for inconsistent or out of range for each suspicious numbers in each question.
4) After the data entry is completed at PSO, it is sent to Center of Information and Statistics as well as to Social and Environment Statistics Department. The second data cleaning process is done at the Social and Environment Statistics Department. Total time for data cleaning often takes from 8 to 12 months.

## [VHLSS 2006]

After finishing data entering and cleaning up, Provincial Statistics Offices have to preliminarily compile the survey data. Leaders from Provincial Statistics Offices examine and send the preliminarily compiled data and data testing report to the Socio- Environmental Statistics Department for assessment. The compilation and assessment of the preliminary results should be finished in 1 month, of them 20 days are for preliminary results and 10 days are for assessment of the preliminary results.

After the preliminary results have been assessed, Provincial Statistics Offices proceed to officially compilation, as well as send the entered raw data to Hanoi Statistical Informatics Center. The Statistical Computing Centre of Hanoi chairs the cooperation with Socio- Environmental Statistics Department to compile the nationwide data.

The General Statistics Office publishes the preliminary results of the VHLSS 2006 in June 2007 and official results in December 2007.

## 3 Data and metadata provided by NSO

### 3.1 Viet Nam Household Living Standard Survey 2002/2004/2006

The data and metadata were provided to Sinfonica in February 2012 by Ms. Ngheim Thi Van, GSO official cum TIU student at that time, with the permission of NSO.

In the above data set, the weight data had not been included. The weight data files were submitted to Sinfonica by Ms. Van, GSO official in July 2014 upon request.

Data and metadata provided

Organization of data files

| Summary sheet.xlsx | Description of list of data files including the next items; <br> Year, STATA filenames, Section in the questionnaire, Main contents, <br> Number of variables |  |
| :--- | :--- | :--- |
|  | Sheet '2002' | Number of data files is 31. <br> Number of sections is 9. |
|  | Sheet '2004' | Number of data files is 47. <br> Number of sections is 9. |
|  | Sheet '2006' | Number of data files is 49. <br> Number of sections is 8. |

Example of Sheet ‘2006’

| 4 | A | B | C | D | E | F |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | List of data files |  |  |  |  |  |
| 2 | Year | No. | STATA filenames | Section | Main contents | Number of variables |
| 3 | 2006 | 1 | muc1a.dta | Section 1: List of household members | List of household members | 17 |
| 4 | 2006 | 2 | muc1b.dta | Section 1: List of household members | List of household members | 16 |
| 5 | 2006 | 3 | muc2a.dta | Section 2: Education, training and vocational training | General information | 34 |
| 6 | 2006 |  | muc2b.dta | Section 2: Education, training and vocational training | Detail information on general and tertiary education | 21 |


| VHLSS 2002 - List of data files |  |  |  |  | Number of variables |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Year | No. | STATA <br> filenames | Section | Main contents |  |
| 2002 | 1 | muc1 | Section 1: List of household members | List of household members | 13 |
| 2002 | 2 | muc2 | Section 2: Education | Education level and expense of household members | 25 |
| 2002 | 3 | muc2ho | Section 2: Education | Total expense for education of household | 9 |
| 2002 | 4 | muc3 | Section 3: Employment | Employment | 27 |
| 2002 | 5 | muc4 | Section 4: Health | Health situation and expense of household members | 11 |
| 2002 | 6 | muc4ho | Section 4: Health | Total expense for health of household | 14 |
| 2002 | 7 | muc5a | Section 5: Income and other inflows of money | income from salary, wage of household members | 18 |
| 2002 | 8 | muc5aho | Section 5: Income and other inflows of money | Total income from salary, wage of household | 10 |
| 2002 | 9 | muc5b1 | Section 5: Income and other inflows of money | Agricultural, forestry and fishery activities | 68 |
| 2002 | 10 | muc5b2ho | Section 5: Income and other inflows of money | Agriculture production, Income from Annual and perennial industrial crops, Income from Fruit crops, Income from Crop by-production, Agricultural production expenditure | 26 |
| 2002 | 11 | muc5b4 | Section 5: Income and other inflows of money | Farm services, Income from Farm services, Expenses for Farm services of household | 88 |
| 2002 | 12 | muc5b5 | Section 5: Income and other inflows of money | Sylviculture, Income from sylvicuture, Sylvicultural expenditure | 62 |
| 2002 | 13 | muc5b21_23 | Section 5: Income and other inflows of money | Income from rice and other food crops, Income from Annual and perennial industrial crops, Income from Fruit crops of household | 13 |
| 2002 | 14 | muc5b25 | Section 5: Income and other inflows of money | Agricultural production expenditure | 11 |


| 2002 | 15 | muc5b31 | Section 5: Income and other inflows of money | Income from livestock | 108 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 2002 | 16 | muc5b32 | Section 5: Income and other inflows of money | Expenditure on livestock | 72 |
| 2002 | 17 | muc5b61 | Section 5: Income and other inflows of money | income from aquaculture | 73 |
| 2002 | 18 | muc5b62 | Section 5: Income and other inflows of money | Expenditure on planting and growing aquacultural products | 49 |
| 2002 | 19 | muc5c1 | Section 5: Income and other inflows of money | Income from non-farm, non-forestry and non-aquaculture businesses | 43 |
| 2002 | 20 | muc5c2 | Section 5: Income and other inflows of money | EXPENDITURE FOR non-farm, non-forestry and non-aquaculture businesses | 55 |
| 2002 | 21 | muc5d | Section 5: Income and other inflows of money | Other sources of income | 27 |
| 2002 | 22 | muc6a1 | Secton 6: Expenditure | expenditure on food and drinks during holidays | 10 |
| 2002 | 23 | muc6a2 | Secton 6: Expenditure | Daily Expenditure on food and drink | 15 |
| 2002 | 24 | muc6b1 | Secton 6: Expenditure | Expenditure on daily consumption | 10 |
| 2002 | 25 | muc6b2 | Secton 6: Expenditure | annual expenditure | 104 |
| 2002 | 26 | muc6b34 | Secton 6: Expenditure | other expenses included and not included in expenditure | 30 |
| 2002 | 27 | muc7 | Section 7: Fixed assets and durable things | fixed assets and durable things | 13 |
| 2002 | 28 | muc7ho | Section 7: Fixed assets and durable things | fixed assets and durable things | 9 |
| 2002 | 29 | muc8 | Section 8: Housing | housing | 120 |
| 2002 | 30 | muc9 | Section 9: Participation in poverty alleviation programs | Participation in poverty alleviation programs | 49 |
| 2002 | 31 | tongcong | Summary | summary | 58 |


| VHLSS 2004 - List of data files |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Year | No | STATA <br> filenames | Section | Main contents | Number of variables |
| 2004 | 1 | M1_2_3A | Section 1,2,3: List of household members, Education and Health | list of household members, Education, training and vocational training, health and health care | 49 |
| 2004 | 2 | M1B | Section 1: List of household members | list of household members | 14 |
| 2004 | 3 | M3B | Section 3: Health and health insurance | Health and health care and health insurance | 17 |
| 2004 | 4 | M4A | Section 4: Income | Employment | 45 |
| 2004 | 5 | M4B11 | Section 4: Income | Rice | 20 |
| 2004 | 6 | M4B12_14 | Section 4: Income | other starchy, vegetable, annual and perennial industrial crops, Fruit Crops | 14 |
| 2004 | 7 | M4B15 | Section 4: Income | income from crop by-products | 11 |
| 2004 | 8 | M4B16 | Section 4: Income | Crop Planting Expenditure | 12 |
| 2004 | 9 | M4B21 | Section 4: Income | Income from Livestock breeding | 15 |
| 2004 | 10 | M4B22 | Section 4: Income | Livestock breeding expenditure | 19 |
| 2004 | 11 | M4B31 | Section 4: Income | Income from Agricultural Services | 10 |
| 2004 | 12 | M4B32 | Section 4: Income | Expenses on Agricultural Services | 18 |
| 2004 | 13 | M4B41 | Section 4: Income | Income from forestry and hunting, trapping and domesticating forest animals and birds | 9 |
| 2004 | 14 | M4B42 | Section 4: Income | expenditure for forestry and hunting, trapping and domesticating forest animals and birds | 21 |
| 2004 | 15 | M4B51 | Section 4: Income | income from aquaculture | 15 |


| 2004 | 16 | M4B52 | Section 4: Income | expenditure for aquaculture | 20 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 2004 | 17 | M4C1 | Section 4: Income | Non-agriculture activity | 25 |
| 2004 | 18 | M4C2 | Section 4: Income | expenditure for from non-agriculture, non-forestry, non-aquaculture business and production trades; the process of agricultural, forestry, and aquacultural products | 19 |
| 2004 | 19 | M4D | Section 4: Income | Other income | 20 |
| 2004 | 20 | M5A1 | Section 5: Expenditure | Expenditure on foods and drinks during holidays | 11 |
| 2004 | 21 | M5A2 | Section 5: Expenditure | Daily expenditure on foods and drinks | 16 |
| 2004 | 22 | M5B1 | Section 5: Expenditure | Expenditure on daily consumption | 11 |
| 2004 | 23 | M5B2 | Section 5: Expenditure | Annual consumption expenditure | 9 |
| 2004 | 24 | M5B3_4 | Section 5: Expenditure | Other spending that is considered and is not considered as household expenditure, | 22 |
| 2004 | 25 | M6 | Secton 6: Fixed assets and consumer durables | Fixed assets and durable appliances | 68 |
| 2004 | 26 | M6A | Secton 6: Fixed assets and consumer durables | Fixed assets and durable appliances | 13 |
| 2004 | 27 | M6B | Secton 6: Fixed assets and consumer durables | Fixed assets and durable appliances | 12 |
| 2004 | 28 | M7 | Section 7: Housing, water and sanitation | Accommodation | 52 |
| 2004 | 29 | M8 | Section 8: Participation in the poverty alleviation and hunger eradication programme and credit | Participation in the poverty alleviation and hunger eradication program | 42 |
| 2004 | 30 | M8_XDGN | Section 8: Participation in the poverty alleviation and hunger eradication programme and credit | Participation in the poverty alleviation and hunger eradication program | 19 |


| 2004 | 31 | M10A_E3 | Section 10: Business other than agriculture, forestry and aquaculture (expanded) | Information on time, location, and labor, Business history, involvement in business associations and clubs, Contact with relevant agencies at different levels and industries, other characteristics | 89 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 2004 | 32 | M10F | Section 10: Business other than agriculture, forestry and aquaculture (expanded) | Occupations, industries which existed during the past 10 years but ceded operations over the 12 months | 15 |
| 2004 | 33 | M91 | Section 9: Agriculture, forestry and aquaculture (expanded) | FARMING, FORESTRY LAND AND WATER SURFACE FOR AQUACULTURE, residential LAND AND GARDEN, POND NEXT TO HOUSING LAND | 23 |
| 2004 | 34 | M92 | Section 9: Agriculture, forestry and aquaculture (expanded) | land that is rented/borrowed/temporarily exchanged | 14 |
| 2004 | 35 | M93 | Section 9: Agriculture, forestry and aquaculture (expanded) | Land that is temporaily rented out/ lent out | 14 |
| 2004 | 36 | M941 | Section 9: Agriculture, forestry and aquaculture (expanded) | rice | 11 |
| 2004 | 37 | M942_44 | Section 9: Agriculture, forestry and aquaculture (expanded) | staple food, food crops and other annual crops, annual and perennial industrial crops, fruit crops | 10 |
| 2004 | 38 | M946 | Section 9: Agriculture, forestry and aquaculture (expanded) | access to extension services | 11 |
| 2004 | 39 | M948 | Section 9: Agriculture, forestry and aquaculture (expanded) | Bought, bid, inherited land or use right-transferred land over the last 10 years | 19 |
| 2004 | 40 | M949 | Section 9: Agriculture, forestry and aquaculture (expanded) | Sold, tender-expired, acquired land or inheritance right-transfered land in the last 10 years | 20 |


| 2004 | 41 | M9451 | Section 9: Agriculture, forestry and aquaculture <br> (expanded) | Cropping structure | 10 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 2004 | 42 | M9452 | Section 9: Agriculture, forestry and aquaculture <br> (expanded) | Cropping structure | 15 |
| 2004 | 43 | M9453 | Section 9: Agriculture, forestry and aquaculture <br> (expanded) | Cropping structure | 13 |
| 2004 | 44 | M9471 | Section 9: Agriculture, forestry and aquaculture <br> (expanded) | conversions in agricultural, forestry and aquacultural land and over <br> the past 10 years | 17 |
| 2004 | 45 | M9472 | Section 9: Agriculture, forestry and aquaculture <br> (expanded) | conversions in agricultural, forestry and aquacultural land and <br> the past 10 years | 15 |
| 2004 | 46 | ho1 | Summary | Summary | 128 |
| 2004 | 47 | ho2 | Summary | Summary | 39 |


| VHLSS 2006 - List of data files |  |  |  |  | Number of variables |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Year | No. | STATA <br> filenames | Section | Main contents |  |
| 2006 | 1 | muc1a.dta | Section 1: List of household members | List of household members | 17 |
| 2006 | 2 | muc1b.dta | Section 1: List of household members | List of household members | 16 |
| 2006 | 3 | muc2a.dta | Section 2: Education, training and vocational training | General information | 34 |
| 2006 | 4 | muc2b.dta | Section 2: Education, training and vocational training | Detail information on general and tertiary education | 21 |
| 2006 | 5 | muc2c.dta | Section 2: Education, training and vocational training | Extra classes | 31 |
| 2006 | 6 | muc2d.dta | Section 2: Education, training and vocational training | Repeated the school year | 19 |
| 2006 | 7 | muc2e.dta | Section 2: Education, training and vocational training | Vocational training | 16 |
| 2006 | 8 | muc3a1.dta | Section 3: Health and health insurance | Health and health care | 12 |
| 2006 | 9 | muc3a2.dta | Section 3: Health and health insurance | Health and health care | 18 |
| 2006 | 10 | muc3b.dta | Section 3: Health and health insurance | Disability | 93 |
| 2006 | 11 | muc3c.dta | Section 3: Health and health insurance | Health | 17 |
| 2006 | 12 | muc3d.dta | Section 3: Health and health insurance | Fertility | 15 |
| 2006 | 13 | muc3e.dta | Section 3: Health and health insurance | Behavoriours that have impacts on health | 11 |
| 2006 | 14 | muc3f.dta | Section 3: Health and health insurance | Health insurance | 28 |
| 2006 | 15 | muc3g.dta | Section 3: Health and health insurance | Out-patient diagnosis treatment | 28 |
| 2006 | 16 | muc3h.dta | Section 3: Health and health insurance | In-patient diagnosis treatment | 31 |
| 2006 | 17 | muc3i.dta | Section 3: Health and health insurance | Self-treatment | 16 |
| 2006 | 18 | muc4a.dta | Section 4: Income | Employment | 52 |
| 2006 | 19 | muc4b0.dta | Section 4: Income | Land for agriculture, forestry and water surface for | 16 |


|  |  |  |  | aquaculture |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 2006 | 20 | muc4b11.dta | Section 4: Income | Rice | 19 |
| 2006 | 21 | muc4b12.dta | Section 4: Income | Other starchy, vegetable | 12 |
| 2006 | 22 | muc4b13.dta | Section 4: Income | Annual and perennial industrial crops | 13 |
| 2006 | 23 | muc4b14.dta | Section 4: Income | Fruit crops | 13 |
| 2006 | 24 | muc4b15.dta | Section 4: Income | Income from crops by products | 10 |
| 2006 | 25 | muc4b16.dta | Section 4: Income | Crop planning expenditure | 11 |
| 2006 | 26 | muc4b161.dta | Section 4: Income | Table of quantity of chemical fertilizers used for types of trees | 11 |
| 2006 | 27 | muc4b21.dta | Section 4: Income | Income from livestock breeding | 14 |
| 2006 | 28 | muc4b22.dta | Section 4: Income | Livestock breeding expenditure | 18 |
| 2006 | 29 | muc4b31.dta | Section 4: Income | Income from agricultural services | 9 |
| 2006 | 30 | muc4b32.dta | Section 4: Income | Expenses on agricultural services | 17 |
| 2006 | 31 | muc4b41.dta | Section 4: Income | Income from forestry and hunting, trapping and domesticating forest animals and birds | 13 |
| 2006 | 32 | muc4b42.dta | Section 4: Income | Expenditure from forestry and hunting, trapping and domesticating forest animals and birds | 20 |
| 2006 | 33 | muc4b51.dta | Section 4: Income | Income from aquaculture | 14 |
| 2006 | 34 | muc4b52.dta | Section 4: Income | Expenditure from aquaculture | 19 |
| 2006 | 35 | muc4c.dta | Section 4: Income | Non-agriculture, non-forestry, non-aquaculture business and production trades | 33 |
| 2006 | 36 | muc4c2.dta | Section 4: Income | Expenditure on -nNon-agriculture, non-forestry, non-aquaculture business and production trades | 9 |


| 2006 | 37 | muc4d.dta | Section 4: Income | Other income | 22 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 2006 | 38 | muc5a1.dta | Section 5: Expenditure | Expenditure on foods and drinks during holidays | 10 |
| 2006 | 39 | muc5a2.dta | Section 5: Expenditure | Daily expenditure on food and drinks | 15 |
| 2006 | 40 | muc5b1.dta | Section 5: Expenditure | Daily non-foods expenditure and other expenditures | 10 |
| 2006 | 41 | muc5b2.dta | Section 5: Expenditure | Annual consumption expenditure | 8 |
| 2006 | 42 | muc5b3_4.dta | Section 5: Expenditure | Other spending that is considered and not considered as household expenditure | 23 |
| 2006 | 43 | muc6.dta | Section 6: Fixed assets and durable appliances | Fixed assets and durable appliances | 67 |
| 2006 | 44 | muc6a.dta | Section 6: Fixed assets and durable appliances | Fixed assets and durable appliances | 12 |
| 2006 | 45 | muc6b.dta | Section 6: Fixed assets and durable appliances | Fixed assets and durable appliances | 11 |
| 2006 | 46 | muc7.dta | Section 7: Accommodation | Accommodation | 51 |
| 2006 | 47 | muc8.dta | Section 8: Participation in the poverty alleviation and hunger eradication programme | Participation in the poverty alleviation and hunger eradication programme | 37 |
| 2006 | 48 | muc8_vayvon.dta | Section 8: Participation in the poverty alleviation and hunger eradication programme | Participation in the poverty alleviation and hunger eradication programme | 20 |
| 2006 | 49 | ttchung.dta | Summary | Summary | 141 |

## [VHLSS 2002]

Outline of the survey;

| General Instruction of <br> VHLSS2002.doc | Description of survey objectives, survey contents, survey methodology, <br> training and field survey, main contents of VHLSS2002 |
| :--- | :--- |

Data in STATA and SAS format

| [1] "muc1.dta" | "muc2.dta" | "muc2ho.dta" | "muc3.dta" |
| :---: | :---: | :---: | :---: |
| [5] "muc4.dta" | "muc4ho.dta" | "muc5a.dta" | "muc5aho.dta" |
| [9] "muc5b1.dta" | "muc5b21_23.d | "muc5b25.dta" | "muc5b2ho.dta" |
| [13] "muc5b31.dta" | "muc5b32.dta" | "muc5b4.dta" | "muc5b5.dta" |
| [17] "muc5b61.dta" | "muc5b62.dta" | "muc5c1.dta" | "muc5c2.dta" |
| [21] "muc5d.dta" | "muc6a1.dta" | "muc6a2.dta" | "muc6b1.dta" |
| [25] "muc6b2.dta" | "muc6b34.dta" | "muc7.dta" | "muc7ho.dta" |
| [29] "muc8.dta" | "muc9.dta" | "tongcong.dta" |  |
| Number of records and variables in each data file; |  |  |  |
| muc1.dta : 13238413 |  |  |  |
| muc2.dta : 13237425 |  |  |  |
| muc2ho.dta : 295329 |  |  |  |
| muc3.dta : 10971527 |  |  |  |
| muc4.dta : 2952211 |  |  |  |
| muc4ho.dta : 2953214 |  |  |  |
| muc5a.dta : 2722018 |  |  |  |
| muc5aho.dta : 2953210 |  |  |  |
| muc5b1.dta : 2953068 |  |  |  |
| muc5b21_23.dta : 10986513 |  |  |  |
| muc5b25.dta : 15356511 |  |  |  |
| muc5b2ho.dta : 2953226 |  |  |  |
| muc5b31.dta : 29530108 |  |  |  |
| muc5b32.dta : 2953072 |  |  |  |
| muc5b4.dta : 2953088 |  |  |  |
| muc5b5.dta : 2953062 |  |  |  |
| muc5b61.dta : 2953073 |  |  |  |
| muc5b62.dta : 2953049 |  |  |  |
| muc5c1.dta : 2953043 |  |  |  |
| muc5c2.dta : 2953 |  |  |  |

```
muc5d.dta : 29530 27
muc6a1.dta : 436795 10
muc6a2.dta : 904003 15
muc6b1.dta: }3508231
muc6b2.dta : 29530 104
muc6b34.dta : 29530 30
muc7.dta: 250928 13
muc7ho.dta : 295329
muc8.dta : 29532120
muc9.dta : }295324
tongcong.dta : 2953251
```

Questionnaire

| Household Questionnaire for Income \& Expenditure Survey in English includes 11 Excel files. |  |
| :--- | :--- |
| File name |  |
| Cover.xls |  |
| Sec01xls to Sec09.xls |  |
| Balance.xls |  |
| Note: Vietnamese version was also provided. |  |

## Data dictionary



Codebook
Codebook_2002.xls Describes variable name, topics, scope, length, code and description in each sheet of dataset


## Province code

| Code of provinces and regions 2002.xls | Includes the next two list; |
| :--- | :--- |
|  | List of provinces in VHLSS 2002 (61 provinces) |
|  | List of regions in VHLSS 2002 (8 regions) |
|  | Note: Definition of regions was not included here. |

## [VHLSS 2004]

Outline of the survey;

| Survey plan.pdf | Description of objectives, contents, methodology, direction and <br> implementation of VHLSS2004 |
| :--- | :--- |
| Guidance to using | 1. Contents of the VHLSS2004 <br> database of <br> VHLSS2004.doc Survey sample of income and expenditure and representative level <br> 3. Content and structure of VHLSS2004 database <br> 4. Some items to pay attention to when using the VHLSS2004 database |

Data in STATA and SAS format
File names in STATA format;

| [1] "ho1.dta" | "ho2.dta" | "m1_2_3a.dta" | "m10a_e3.dta" |
| :--- | :--- | :--- | :--- |
| [5] "m10f.dta" | "m1b.dta" | "m3b.dta" | "m4a.dta" |
| [9] "m4b11.dta" | "m4b12_14.dta" "m4b15.dta" | "m4b16.dta" |  |
| [13] "m4b21.dta" | "m4b22.dta" | "m4b31.dta" | "m4b32.dta" |
| [17] "m4b41.dta" | "m4b42.dta" | "m4b51.dta" | "m4b52.dta" |
| [21] "m4c1.dta" | "m4c2.dta" | "m4d.dta" | "m5a1.dta" |
| [25] "m5a2.dta" | "m5b1.dta" | "m5b2.dta" | "m5b3_4.dta" |
| [29] "m6.dta" | "m6a.dta" | "m6b.dta" | "m7.dta" |
| [33] "m8.dta" | "m8_xdgn.dta" | "m91.dta" | "m92.dta" |
| [37] "m93.dta" | "m941.dta" | "m942_44.dta" | "m9451.dta" |
| [41] "m9452.dta" | "m9453.dta" | "m946.dta" | "m9471.dta" |
| [45] "m9472.dta" | "m948.dta" | "m949.dta" |  |

Number of records and variables in each data file;
ho1.dta : 9188128
ho2.dta : 918839
m1_2_3a.dta : 4043849
m10a_e3.dta : 437689
m10f.dta : 41615
m1b.dta : 2020914
m3b.dta : 1677717
m4a.dta : 3736845
m4b11.dta: 1110220
m4b12_14.dta : 3144414
m4b15.dta : 849311
m4b16.dta : 5416112

```
m4b21.dta : 1822515
m4b22.dta : }113111
m4b31.dta : }3011
m4b32.dta : }2991
m4b41.dta : 4593 9
m4b42.dta : }25332
m4b51.dta : 424315
m4b52.dta : 304620
m4c1.dta : 4544 25
m4c2.dta : }291521
m4d.dta : 918820
m5a1.dta : }1299931
m5a2.dta : }2901431
m5b1.dta : 10873511
m5b2.dta: 116575 9
m5b3_4.dta: 9188 22
m6.dta : 918868
m6a.dta : }165121
m6b.dta : 7923512
m7.dta : 918852
m8.dta : 918842
m8_xdgn.dta : 523319
m91.dta : 3531723
m92.dta : }12191
m93.dta : }8051
m941.dta : }110951
m942_44.dta : 31444 10
m9451.dta : 25119 10
m9452.dta : 16003 15
m9453.dta : }33471
m946.dta : 1222911
m9471.dta : 737 17
m9472.dta : }2621
m948.dta : 6506 19
m949.dta : }10612
muc6b34.dta : 29530 30
```

```
muc7.dta : 250928 13
muc7ho.dta: 295329
muc8.dta : 29532120
muc9.dta: 2953249
tongcong.dta : 2953251
```

Questionnaire

| Household Questionnaire for Income \& Expenditure Survey in English includes 12 Excel files. |  |
| :--- | :--- |
| File name |  |
| Cover.xls |  |
| Section01xls to Section10.xls |  |
| Indicators.xls |  |
| Note: Vietnamese version was also provided. |  |

## Codebook



Province code

| Province code.xls | Includes the next list; <br> List of provinces in VHLSS 2004 (64 provinces) |
| :--- | :--- |
| Remarks: Three provinces increased. |  |

## [VHLSS 2006]

Outline of the survey;

| Survey plan.pdf | Description of objectives, contents, methodology, direction and <br> implementation of VHLSS2006 |
| :--- | :--- |

Data in STATA and SAS format

| File names in STATA format; |  |  |  |
| :---: | :---: | :---: | :---: |
| [1] "muc1a.dta" | "muc1b.dta" | "muc2a.dta" | "muc2b.dta" |
| [5] "muc2c.dta" | "muc2d.dta" | "muc2e.dta" | "muc3a1.dta" |
| [9] "muc3a2.dta" | "muc3b.dta" | "muc3c.dta" | "muc3d.dta" |
| [13] "muc3e.dta" | "muc3f.dta" | "muc3g.dta" | "muc3h.dta" |
| [17] "muc3i.dta" | "muc4a.dta" | "muc4b0.dta" | "muc4b11.dta" |
| [21] "muc4b12.dta" | "muc4b13.dta" | "muc4b14.dta" | "muc4b15.dta" |
| [25] "muc4b16.dta" | "muc4b161.dta" | "muc4b21.dta" | "muc4b22.dta" |
| [29] "muc4b31.dta" | "muc4b32.dta" | "muc4b41.dta" | "muc4b42.dta" |
| [33] "muc4b51.dta" | "muc4b52.dta" | "muc4c.dta" | "muc4c2.dta" |
| [37] "muc4d.dta" | "muc5a1.dta" | "muc5a2.dta" | "muc5b1.dta" |
| [41] "muc5b2.dta" | "muc5b3_4.dta" | "muc6.dta" | "muc6a.dta" |
| [45] "muc6b.dta" | "muc7.dta" | "muc8.dta" | "muc8_vayvon.dta" |
| [49] "ttchung.dta" |  |  |  |
| Number of records and variables in each data file; |  |  |  |
| muc1a.dta: 3907117 |  |  |  |
| muc1b.dta : 1879216 |  |  |  |
| muc2a.dta : 3907134 |  |  |  |
| muc2b.dta : 3907121 |  |  |  |
| muc2c.dta : 3907131 |  |  |  |
| muc2d.dta : 3907119 |  |  |  |
| muc2e.dta: 3907116 |  |  |  |
| muc3a1.dta: 3907112 |  |  |  |
| muc3a2.dta : 1852418 |  |  |  |
| muc3b.dta : 3907193 |  |  |  |
| muc3c.dta : 3907117 |  |  |  |
| muc3d.dta : 664715 |  |  |  |
| muc3e.dta : 3382611 |  |  |  |
| muc3f.dta: 39071 |  |  |  |
| muc3g.dta : 11058 |  |  |  |

[^0]Questionnaire

| Household Questionnaire for Income \& Expenditure Survey in English |  |
| :--- | :--- |
| File name | VHLSS06_questionnaire.pdf (85 pages) |

Codebook

| Codebook_2006.xls | Describes variable name, topics, scope, length, code and description in each sheet of dataset |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Example |  | D | E | 6 |
|  |  |  | vaiables | 17 |
|  |  | $\underset{\text { Alsope }}{\text { Sal }}$ |  |  |
|  |  | ${ }^{\text {AIII }}$ | $\underset{\substack{20153 \\ 20195}}{ }$ | $\underset{\substack{\text { chanater } \\ \text { Chasater }}}{ }$ |
|  | (tan | ${ }_{\text {AII }}$ | ${ }_{\substack{\text { a }}}^{2010095}$ | Chasater |
|  |  | ${ }_{\text {AIII }}$ | ${ }_{\text {c }}^{813,25}$ | Numeic |
|  | ${ }_{13}^{12} 7$ MAAC2 Sexo fith member |  | ${ }_{8}$ | Yes |
|  | ${ }_{15}^{14} 8$ maC3 Reationstip wint HH member |  | ${ }^{2}$ |  |
|  | ${ }_{17}^{16}$ |  | 4 |  |
|  | ${ }_{18}^{18}$ |  | ${ }_{3}$ | coicle |
|  | ${ }_{2}^{19}$ |  | 5 |  |
|  |  |  | ${ }^{6}$ |  |
|  | $1{ }_{23}^{22}$ |  | 9 | Mssing |

Province code

| Province code.xls | Includes the next list; <br> List of provinces in VHLSS 2006 (64 provinces) |
| :--- | :--- |
| Note: Not confirmed whether this is the same as VHLSS 2004. |  |

Survey results

| Survey report of <br> VHLSS 2006 | Includes the results of VHLSS 2002, 2004 and 2006 as time series |
| :---: | :---: |
| File name | Introduction.pdf and Part 01.pdf to Part 11.pdf |
| Content of the report | Result of the Vietnam household living standards survey 2006 |
| Note | The survey report is also available at NSO's website. http://www.gso.gov.vn/default_en.aspx?tabid=515\&ItemID=8183 (Accessed on 11 July 2014) |

## ［Weight data］

Weight is unique within xa（commune），which is psu．The following weight data by xa was provided；

| File | Description | No of records |
| :--- | :--- | :--- |
| weight＿by＿xa02．dta | Weight for VHLSS 2002 by psu | 2,901 |
| weight＿by＿xa04．dta | Weight for VHLSS 2004 by psu | 3,061 |
| weight＿by＿xa06．dta | Weight for VHLSS 2006 by psu | 3,063 |

Example of VHLSS 2006：

| $\pi$ | データエティタ |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | tinh | huyen | xa | wt9 | var5 |
| 1 | 101 | 1 | 3 | 3107．318 |  |
| 2 | 101 | 1 | 9 | 3092.521 |  |
| 3 | 101 | 1 | 15 | 3085.123 |  |
| 4 | 101 | 1 | 17 | 3099.92 |  |
| 5 | 101 | 1 | 21 | 3085.123 |  |
| 6 | 101 | 1 | 23 | 3099.92 |  |
| 7 | 101 | 3 | 3 | 2988．944 |  |
| 8 | 101 | 3 | 11 | 3003.741 |  |

### 3.2 Other sources of metadata

$\square \quad$ IHSN (International Household Survey Network)

VHLSS 2002 (http://catalog.ihsn.org/index.php/catalog/3284/related_materials)
(Accessed on 11 July 2014)

The following documents are available;

## Documentation

Download the questionnaires, technical documents and reports that describe the survey process and the key results for this study.

| [ Household Living Standards Survey 2001-2002: Household Questionnaire | A 575.73 KB |
| :---: | :---: |
| [ Vietnam Household Living Standards Survey 2002: Commune Questionnaire | A 133.07 KB |
| - Technical Documents |  |
| [ Vietnam Household Living Standards Survey (VHLSS), 2002 and 2004: Basic Information | A 202.93 KB |
| [0] General Introduction of Vietnam Household Living Standards Survey (VHLSS) | - 121.6 KB |
| T Main difference between VLSS 1992/1993, VLSS 1997/1998 and VHLSS 2001/2002 | A 77.39 KB |
| [ Codes of Provinces and Regions | A 52.18 KB |


| Document | Description |
| :--- | :--- |
| VHLSS 2002 and 2004 <br> Basic Information <br> (47 pages) | This document was prepared by the officials from Social \& Environment <br> Statistics Department, Vietnam General Statistics Office. It describes on <br> sample design of VHLSS 2002 and 2004 in detail. <br> Main difference <br> between VLSS <br> 1992/1993, VLSS <br> $1997 / 1998$ and VHLSS <br> $2001 / 2002 \quad$ (5 pages) |

VHLSS 2004 (http://catalog.ihsn.org/index.php/catalog/3285/related_materials)

The following documents are available;

## Documentation

Download the questionnaires, technical documents and reports that describe the survey process and the key results for this study.

| - Questionnaires |  |
| :---: | :---: |
| 4 Household Living Standards Survey 2004: Questionnaire on Household Survey | A 3.18 MB |
| \% Household Living Standards Survey 2004: Questionnaire on Household Survey (Vietnamese) | A 3.32 MB |
| 4 Vietnam Household Living Standards Survey 2004: Commune Questionnaire | - 71.1 KB |


Other Materials
Stata programs

| Document | Description |
| :--- | :--- |
| VHLSS 2004 | It is the operational handbook in the field. It describes on the tasks and |
| Operational Handbook | responsibilities of team leader, enumerator and supervisor, as well as how |
| (124 pages) | to fill in questionnaire in detail. |

$\square \quad$ World bank: Living Standards Measurement Study
(http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,c ontentMDK:21709814~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html

Accessed on 07 August 2014)

$\square$ "Survey design and sample design in household budget surveys" by Hans Pettersson,
Statistics Sweden (14 pages)
(http://unstats.un.org/unsd/hhsurveys/pdf/Chapter_24.pdf) (Accessed on 11 July 2014)

This paper addresses some issues on survey design and sample design for household budget surveys. It explained the square root allocation method as one compromise among conflicting demands from National Accounts, and government planners and policy analysts. Square root allocation has been used for VHLSS. He had designed the sampling for VHLSS.

## $\square \quad$ Enumerators’ manual of VHLSS 2004

"Operational Handbook: VHLSS 2004" by GSO, April 2004, 124 pages
(Available at IHSN website. Contents copying and changing documents are restricted.)

Below is the contents of the manual. Especially, "Section 1. Part A. List of household members" is very important.

## CONTENTS

Part I: General principles ..... 3
Part II: Questionnaire for interviewing households ..... 10
Section 0: Information on the cover of the household questionnaire. ..... 16
Section 1. Part A. List of household members ..... 18
Section 1. Part B. Household members in VHLSS 2002 ..... 24
Section 2 . Education, training and vocational training ..... 25
Section 3. Health care. ..... 29
Section 4. Income ..... 33
Section 5. Expenditure ..... 74
Section 6. Fixed assets and consumer durables ..... 80
Section 7. Housing. ..... 82
Section 8. Participation in poverty reduction programmes. ..... 86
Section 9. Agriculture, forestry and fisheries (expanded) ..... 89
Section 10 . Business and production trades, non-agriculture, forestry and fisheries services (expanded) ..... 105
Part III: Questionnaire for interviewing communes ..... 109

## 4. Map of Provinces and Regions (as of 2006)

## Map of Provinces

The below map of Vietnam exhibiting its 59 provinces and 5 centrally controlled municipalities at the time of VHLSS 2004 and 2006. Ha Tay province, located in the south west of " 2 Ha Noi" was merged with Ha Noi municipality in 2008. The number of provinces is 63 in 2014.


## Map of regions

Provinces are grouped into eight regions.
Note: From VHLSS 2010, Vietnam is divided into 6 regions in GSO's publication.

| Code | Region | Province codes | Note: Regions from 2001 |
| :--- | :--- | :--- | :--- |
| 1 | Red River Delta | $101-117$ | Red River Delta |
| 2 | North East | $201-225$ | Northern Midland and <br> Mountains |
| 3 | North West | $301-305$ | North Central and Central |
| 4 | North Central | $401-411$ | Coast |
| 5 | South Central Coast | $501-511$ | Central Highlands |
| 6 | Central Highlands | $601-607$ | South East |
| 7 | South East | $701-717$ | Mekong River Delta |
| 8 | Mekong River Delta | $801-823$ |  |



## 5. Panel Data (Trial)

It is an advantage of survey design of VHLSS to make use of panel data.

## What is panel data?

The term panel data refers to observations obtained over multiple time periods for the same households or individuals.
\{Xijt\} i: household or individual, j: variable, t: year

In statistical surveys, panel data is essentially a set of pairs of household identifiers (ID06, ID04) and pairs of individual identifiers (PID06, PID04). Each identifier is uniquely linked with household or individual data in VHLSS 2006 and 2004.

pair of (PID06, PID04)

| ID06 | PID06 | Individual data in 2006 |
| :--- | :--- | :--- |


| ID04 | PID04 | Individual data in 2004 |
| :--- | :--- | :--- |

Where,
ID: Household identifier
PID: Individual identifier
06: VHLSS 2006
04: VHLSS 2004

## How can the design of VHLSS make panel data?

VHLSS 2006 has the question on whether the household was surveyed in VHLSS 2004 in Section 1b. And if yes, household identification of VHHLSS 2004 was recorded by the survey team leader based on the sample list of VHLSS 2004. Finally the data set ttchung has the variable of "ghepho" on whether matched with VHLSS 2004.

As for matching household members, the data set MUC1B, list of household members at the time of

VHLSS 2004 was prepared. Basic information was recorded by the survey team leader based on the list of household members of VHLSS 2004. It has the variables of person number in 2004, sex in 2004 and age in 2004 in addition to name in 2004, as well as person number in VHLSS 2006 if a person is the household member at the time of VHLSS 2006. And if a person is no longer living in the household at the time of VHLSS 2006, the reason was asked. Finally, the data set has the variable of ghep on whether matched with VHLSS 2004.


## What is available from panel data?

- Household panel data show the transition of household-level variables over time. It also provides information on the household which was selected for VHLSS 2006 but had removed to other EA.

Transition of households
VHLSS 2004
VHLSS 2006

| A half of sample <br> households in 2004 | Re-selected <br> in 2006 | Surveyed in 2006 |
| :--- | :--- | :--- |
| Removed out of EA <br> Another half of <br> sample households <br> in 2004 |  |  |

- For instance, as described in the survey report, GSO makes use of panel data to compare directly household income per capita at household level.
- Individual panel data provide information on changes of individual-level variables of persons who were surveyed both in 2004 and 2006, that is, remained in the same households during 2004 and 2006.

In addition, it reveals persons who were surveyed in 2004 but have moved out of the household or died, as well as persons who moved in the household or were born during 2004 to 2006.

Transition of household members
VHLSS 2004 During 2004-06 VHLSS 2006


## Possible errors during conducting surveys

If the pairs of (ID06, ID04) and (PID06, PID04) are perfect, making panel data is a task of merging data sets between 2006 and 2004. However, it is natural to consider errors in the pairs of (ID06, ID04) and (PID06, PID04) during the survey process.
Possible errors are as the next;

- Recording errors when a survey team leader filled in questionnaire from the lists of VHLSS 2004.
- Data entry errors by an operator.
- Response errors by a sample household

Therefore, we will treat pairs of (ID06, ID04) and (PID06, PID04) available in data sets as candidates of panel data.

## Strategy for creating panel data

1) To prepare data sets and variables for matching
2) To verify the information for matching
3) To drop the unqualified records from candidates of panel
4) To create panel at household level and individual level

## Preparation for making panel data

Goal: The below data sets with household and individual identifiers will be prepared.

| Data set | Original | Description | Household <br> identifier |  | Individual <br> identifier |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| HLD06 | TTCHUNG | All households surveyed in <br> 2004 and link data | ID06 | ID04 |  |  |
| HLD04 | HO1 | All households |  | ID04 |  |  |
| LIN06 | MUC1B | Individual-level link data | ID06 | ID04 | PID06 | PID04 |
| MEM06 | MUC1A | All household members <br> belonging to households <br> surveyed in 2004 | ID06 |  | PID06 |  |
| MEM04 | M1_2_3A | All household members |  | ID04 |  | PID04 |

## VHLSS 2006

List of data sets and variables related to panel (VHLSS 2006)

| Data set | Variable | Description |
| :---: | :---: | :---: |
| TTCHUNG |  | Summary file for all households ( 9,189 ) |
|  | m1c1 | Surveyed in 2004? (Yes: 4,298 No: 4,891) |
|  | ghepho | Matched with VHLSS 2004? <br> (0: 4,891 Not surveyed in 2004 1: 4,267 Data in 2004 9: 31) |
|  | tinh04 |  |
|  | huyen04 |  |
|  | xa04 |  |
|  | diaban04 |  |


|  | ttnt04 |  |
| :--- | :--- | :--- |
|  | hoso04 |  |
|  |  | All household members (39,071) |
| MUC1A |  |  |
|  |  | Household members surveyed in VHLSS 2004 (18,792) |
|  | m1bc3 | Person number in 2004 |
| MUC1B | m1bc4 | Sex in 2004 |
|  | m1bc5 | Age in 2004 |
|  | ghep | Matched with VHLSS 2004? (0: 671 Not surveyed in 2004 <br> 17,039 Data in 2004 9: 37 NA: 1,045) |
|  | m1bc6 | Household member in 2006? (1: 17,076 Yes 2: 1,716 No) |
|  | m1bc8 | Person number in 2006 |
|  | m1bc9 | Reason for moving out of the household |
|  | m1bc10 | Province code work in |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Generating household-level data frame HLD06 from TTCHUNG

```
> d<-subset(ttchung,m1c1==1)
> dim(d)
```

[1] $4298 \quad 144$
\# Renamed the household identifier ID as ID06
> d["ID06"]<-d\$ID
$\checkmark \quad$ Generated the household identifier ID04.
> d["ID04"]<-as.character(d\$hoso04+(10^2)*d\$diaban04+(10^5)*d\$xa04+
$+\left(10^{\wedge} 7\right) * \mathrm{~d} \$ h u y e n 04+\left(10^{\wedge} 9\right) * \mathrm{~d} \$$ tinh04)
$>$ HLD06<-d
\# Remarks: Out of 4,298 households, 3,667 households have the same ID both in 2004 and 2006.

```
> table(HLD06$ID==HLD06$ID04,useNA="ifany")
```

FALSE TRUE <NA>
63136674891

## Generating individual-level data frame MEM06 from MUC1A

```
> d<-muc1a
> str(d[1:6])
'data.frame': }39071\mathrm{ obs. of 6 variables:
$ tinh : chr "101" "101" "101" "101" ...
$ huyen : chr "01" "01" "01" "01" ...
$ xa : chr "03" "03" "03" "03" ...
$ diaban: chr "014" "014" "014" "014" ...
$ hoso : int 15151515191919192424 ..
$ matv : int 1234123412 ...
```

$\checkmark \quad$ Generated household identifier ID06 in MUC1A.
> d["ID06"]<-as.character(d\$hoso+(10^2)*as.integer(d\$diaban)+(10^5)*as.integer(d\$xa)+
$+(10 \wedge 7)^{*}$ as.integer(d\$huyen)+(10^9)*as.integer(d\$tinh))
$\checkmark \quad$ Generated individual identifier PID06 as a combination of ID06 and person number (matv).
> d["PID06"]<-paste(d\$ID06,formatC(d\$matv,width=2,flag="0"),sep="")
> MEM06<-subset(d,is.element(d\$ID06,HLD06\$ID06))
$>\operatorname{dim}($ MEM06 $)$
[1] 1841923

## Generating individual-level data frame LIN06 from MUC1B

```
> d<-muc1b
```

\# Renamed the household identifier ID as ID06
> d["ID06"]<-d\$ID
$\checkmark$ Generated individual identifier PID06 as a combination of ID06 and person number in 2006 (m1bc7) if the person was a household member in 2006.
> t<-table(d\$m1bc6)
> names(t)<-c("Member in 2006","No")
$>\mathrm{t}$
Member in 2006 No
17076
1716
> d["PID06"]<-ifelse(d\$m1bc6==1,paste(d\$ID06,formatC(d\$m1bc7,width=2,flag="0"),sep=""),NA)
$\checkmark$ Appended ID04 by merging with HLD06, and generated PID04 as a combination of ID04 and person number in 2004 (m1bc3).
> d<-merge(d,HLD06[c("ID06","ID04")],key="ID06",all.x=T)
> d["PID04"]<-paste(d\$ID04,formatC(d\$m1bc3,width=2,flag="0"),sep="")
$>$ LIN06<-d
$\checkmark$ LIN06 consists of 18,876 household members at the time of VHLSS 2004 in sample households surveyed in VHLSS 2004, of which 17,076 persons are household members in VHLSS 2006, that is, they have PID06.
> table(is.na(LIN06\$PID04),useNA="ifany")
FALSE
18792
> table(is.na(LIN06\$PID06),useNA="ifany")
FALSE TRUE
170761716

## VHLSS 2004

List of data sets and variables related to panel (VHLSS 2004)

| Data set | Variable | Description |
| :--- | :--- | :--- |
| HO1 |  | Summary file for all households (9,188) |
|  | m1c1 | Surveyed in 2002? (1 Yes: 4,476 2 No: 4,712) |
|  | tinh02 |  |
|  | xa02 |  |
|  | ttnt02 |  |
|  | hoso02 |  |
|  | quy02 | Quarter code in 2002 |


|  |  |  |
| :--- | :--- | :--- |
|  |  |  |
| M1_2_3A |  | All household members (40,438) |
|  |  |  |
|  |  | Household members surveyed in VHLSS 2002 (20,209) |
|  |  | Person number in 2002 |
| M1B | m1bc3 | m1bc4 |
|  | m1bc5 | Age in 2002 |
|  | m1bc6 | Household member in 2004? (1: 18,353 Yes 2: 1,856 No) |
|  | m1bc7 | Person number in 2004 |
|  | m1bc8 | Reason for no longer living in the household |
|  |  | Reason for moving out of the household |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Generating household-level data frame HLD04 from HO1

> d<-ho1
\# Generated household identifier ID04.
> d["ID04"]<-as.character(d\$hoso+(10^2)*d\$diaban+(10^5)*d\$xa+(10^7)*d\$huyen+(10^9)*d\$tinh)

## Generating individual-level data frame MEM04 from M1_2_3A

> d<-m1_2_3a
$\checkmark$ Generated household identifier ID04 in M1_2_3A.
$>d[" I D 04 "]<-$ as.character(d\$hoso+(10^2)*d\$diaban+(10^5)*d\$xa+(10^7)*d\$huyen+(10^9)*d\$tinh $)$
$\checkmark \quad$ Generated individual identifier PID04 as a combination of ID04 and person number (matv).
> d["PID04"]<-paste(d\$ID04,formatC(d\$matv,width=2,flag="0"),sep="")
$>$ MEM04<-d

## Verifying data sets and variables for matching

$\checkmark \quad$ To generate data set hp1 as pairs (ID06, ID04) from HLD06, which is a candidate for household-level panel.

```
> hp1<- HLD06[c("ID06", "ID04")]
> dim(hp1)
[1] 4298 2
>head (hp1)
    ID06 ID04
101010301415101010301415
2 101010301419101010301419
16101012301813101012301813
31101051100213101051100213
32101051100215101051100215
33101051100219101051100219
```

$\checkmark \quad$ To state two conditions for hp1 and seven conditions for LIN06 essential for panel, and prepare flags on whether conditions are satisfied or not for each record in the datasets.

## Tow conditions for hp1

1) ID04 in hp1 should be in HLD04.
2) This linkage should be one to one relationship. There should be no duplicated ID04 in hp1.

Condition 1: ID04 in hp1 should be in HLD04.
$\checkmark$ Out of 4,298 households surveyed in VHLSS 2004, the number of households linked with HLD04 is 4,219.
\# Appended flag1 for unmatched ID04 in hp1
>hp1["flag1"]<-ifelse(is. element (hp1\$ID04, HLD04\$ID04), 1, 0)
$>$ table (hp1\$flag1)
01
794219

| \# Example of errors |  |  |  |
| :---: | :---: | :---: | :---: |
| > head (subset (hp1, flag1==0)) |  |  |  |
|  | ID06 | ID04 | flag1 |
| 105 | 101082500115 | 101082500115 | 0 |
| 727 | 105213700813 | 105213700813 | 0 |
| 742 | 105231900313 | 105231900313 | 0 |
| 743 | 105231900314 | 105231900314 | 0 |
| 744 | 105231900315 | 105231800318 | 0 |
|  | 109051100413 | 109051100413 | 0 |

Condition 2: This linkage should be one to one relationship. There should be no duplicated ID04 in hp1.
$\checkmark \quad$ There are 13 duplicated ID04 in hp1 as follows.

```
> sum(duplicated(subset(hp1, flag1==1) $ID04))
```

[1] 13

```
# List of duplicated ID04
>dup. ID04<-hp1[dupl icated(hp1$ID04), "ID04"]
dup. ID04
```

    [1] "101093301019" "209214100214" "211033700214" "211073300814"
    [5] "211095700513" "211114700114" "217012100813" "221072300615"
    [9] "305101500215" "715130300514" "813070501613" "813191301219"
    [13] "819070700115"
\# Appended flag2 for duplicated records in hp1.
> hp1["flag2"]<-ifelse(! is. el ement (hp1\$ID04, dup. ID04), 1, 0)
$>\mathrm{t}<-$-addmargins (table(hp1\$flag1, hp1\$flag2)) $[\mathrm{c}(3,2,1), \mathrm{c}(3,2,1)]$
> rownames ( t )<-c ("Surveyed in 2004", "Linked with HLDO4", "No")
> colnames ( t )<-c ("Total", "No duplication", "Duplicated")
> t

Total No duplication Duplicated

| Surveyed in 2004 | 4298 | 4272 | 26 |
| :--- | ---: | ---: | ---: |
| Linked with HLD04 | 4219 | 4193 | 26 |
| No | 79 | 79 | 0 |

```
# List of 26 duplicated records
> hp1[hp1$flag2==0, c ("ID06", "ID04")]
    ID06 ID04
130 101093301015101093301019
131 101093301019101093301019
2417209214100214209214100214
2418209214100215 209214100214
2438211033700214211033700214
2439211033700215 211033700214
2471211073300814211073300814
2472 211073300815 211073300814
2491211095700513211095700513
2492211095700514211095700513
2521211114700114211114700114
2522 211114700115 211114700114
2800217012100813217012100813
2801217012100815 217012100813
2992221072300613221072300615
2994221072300615 221072300615
3 6 2 6 3 0 5 1 0 1 5 0 0 2 1 4 3 0 5 1 0 1 5 0 0 2 1 5
3 6 2 7 3 0 5 1 0 1 5 0 0 2 1 5 3 0 5 1 0 1 5 0 0 2 1 5
7 1 7 8 7 1 5 1 3 0 3 0 0 5 1 4 7 1 5 1 3 0 3 0 0 5 1 4
7 1 7 9 7 1 5 1 3 0 3 0 0 5 1 5 7 1 5 1 3 0 3 0 0 5 1 4
8362813070501613813070501613
8363813070501614813070501613
8456813191301214813191301219
8457813191301215813191301219
8 8 6 6 8 1 9 0 7 0 7 0 0 1 1 3 8 1 9 0 7 0 7 0 0 1 1 5
8 8 6 8 8 1 9 0 7 0 7 0 0 1 1 5 8 1 9 0 7 0 7 0 0 1 1 5
```

$\checkmark \quad$ There are two different ID06 for each duplicated ID04. While we assume that one of the two same ID04 might be right and the other might be a result of data entry miss, removed the above 26 records tentatively because we cannot discuss on which is the right linkage at this moment.
$\checkmark \quad$ Created household-level panel hp2 consisted of 4,193 pairs of ID06 and ID04.
$>$ hp2<-subset (hp1, flag1==1\&flag2==1) [c ("ID06", "ID04")]

```
> dim(hp2)
[1] 4193 2
> head (hp2)
    ID06 ID04
101010301415101010301415
2 101010301419101010301419
16101012301813101012301813
31101051100213101051100213
32101051100215101051100215
33101051100219101051100219
```


## Seven conditions for LIN06 in order to link the identical households

3) A person with PID04 should be in MEM04.
4) There should be no duplicated PID04 in LIN06.
5) The person's sex and age in LIN06 and MEM04 should be identical.
6) ID06 in LIN06 should be the same as ID06 in hp.
7) If a person in LIN06 was a household member in VHLSS 2006, that is, PID06 was defined in LIN06, the person with the PID06 should be in MEM06.
8) There should be no duplicated PID06 in LIN06.
9) The person's sex and age in LIN06 and MEM06 should be identical.

Condition 3: A person with PID04 in LIN06 should be in MEMO4.
$\checkmark$ Out of 18,792 persons in LIN06, the number of those linked with records in MEM04 is 18,348.

```
# Appended flag3 for unmatched PID04 in LIN06
> LIN06["flag3"]<-ifelse(is. element(LIN06$PID04, MEMO4$PID04), 1, 0)
>t<-addmargins(table(LIN06$m1bc6, LIN06$flag3)) [c (3, 1, 2), c (3, 2, 1)]
> rownames (t)<-c ("Person in LIN04", "Member in 2006", "No (move-out/died)")
> colnames(t)<-c ("Total", "Linked with MEMO4", "No")
t
```

|  | Total Linked with MEMO4 | No |  |
| :--- | :---: | ---: | ---: |
| Person in LIN04 | 18792 | 18348 | 444 |
| Member in 2006 | 17076 | 16685 | 391 |
| No (move-out/died) | 1716 | 1663 | 53 |

## Condition 4: There should be no duplicated PIDO4 in LIN06.

```
\checkmark Among 18,348 PID04 in LIN06 which linked with MEM04, there are 42 duplicated PID04 as
    follows. Which record is right one will be discussed in the next condition 5.
> sum(dupl icated (subset(LIN06, flag3==1) $PID04)==T)
[1] }4
# List of duplicated ID04
> dup.PID04<- subset(LIN06, flag3==1) [duplicated(subset(LIN06, flag3==1) $PID04) ==T, "PID04"]
> length (dup. PIDO4)
[1] 42
# Appended flag4 for duplicated records in LIN06
> LIN06["flag4"]<-ifelse(!is. element(LIN06$PID04, dup. PID04), 1, 0)
>t<-addmargins(table(LIN06$flag3, LIN06$flag4)) [c (3, 2, 1), c (3, 2, 1)]
> rownames(t)<- c ("Person in LIN06","Linked with MEMO4", "No")
>colnames (t)<-c ("Total", "No duplication", "Duplicated")
t
Total No duplication Duplicated
\begin{tabular}{llll} 
Person in LIN06 & 18792 & 18708 & 84
\end{tabular}
    Linked with MEMO4 18348 18264 84
    No 444 444 0
# Examples of duplicated records
> for(j in dup.PID04) print(LIN06[LIN06$PID04==j, c("PID06","PID04")])
    PID06 PID04
17810109330101502 10109330101902
1821010933010190210109330101902
    PID06 PID04
```

1791010933010150110109330101901
1831010933010190110109330101901

PID06 PID04
180118190707001130381907070011503
180278190707001150381907070011503

## Condition 5: The person' s sex and age in LIN06 and MEMO4 should be identical.

```
>d<-subset(LIN06, flag3==1)
>df1<-d[c ("PID06", "ID06", "PID04", "ID04", "m1bc4", "m1bc5", "flag3", "flag4")]
dim(df1)
[1] 18348 8
> colnames (df1)[5:6]<-c ("Isex", "Iage")
> head (df1)
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline & PID06 & ID06 & PID04 & ID04 & & & & flag4 \\
\hline 1 & 10101030141501 & 101010301415 & 10101030141502 & 101010301415 & 2 & 46 & 1 & \\
\hline 2 & 10101030141503 & 101010301415 & 10101030141504 & 101010301415 & 1 & 24 & 1 & \\
\hline 3 & 10101030141502 & 101010301415 & 10101030141503 & 101010301415 & 1 & 26 & 1 & \\
\hline & 10101030141504 & 101010301415 & 10101030141505 & 101010301415 & 1 & 18 & 1 & \\
\hline 5 & <NA> & 101010301415 & 10101030141501 & 101010301415 & 1 & 57 & 1 & \\
\hline & 10101030141902 & 101010301419 & 10101030141902 & 101010301419 & 1 & 24 & 1 & 1 \\
\hline
\end{tabular}
```

```
>df2<-MEM04[c("PID04", "ID04", "m1ac2", "m1ac5")]
>dim(df2)
[1] 40438 4
>colnames(df2)<-c ("PID04", "ID04", "sex04", "age04")
> head (df2)
    PID04 ID04 sex04 age04
110101030141301 101010301413 2 43
210101030141302101010301413 1 50
310101030141303101010301413 1 20
410101030141304101010301413 1 17
510101030141501 101010301415 1 50"
610101030141502 101010301415 2 46
```

```
> cond5<-merge(df1[c ("PID06", "PID04", "Isex", "lage", "flag3", "flag4")],
+ df2[c ("PID04", "sex04", "age04")],by="PID04", alI. x=T)
dim(cond5)
[1] 18348 8
> head (cond5)
\begin{tabular}{lllllllll} 
& \multicolumn{10}{c}{ PID04 } & PID06 & Isex & lage flag3 & flag4 & sex04 & age04 \\
1 & 10101030141501 & 〈NA〉 & 1 & 57 & 1 & 1 & 1 & 50 \\
2 & 10101030141502 & 10101030141501 & 2 & 46 & 1 & 1 & 2 & 46 \\
3 & 10101030141503 & 10101030141502 & 1 & 26 & 1 & 1 & 1 & 26 \\
4 & 10101030141504 & 10101030141503 & 1 & 24 & 1 & 1 & 1 & 24 \\
5 & 10101030141505 & 10101030141504 & 1 & 18 & 1 & 1 & 1 & 18 \\
6 & 10101030141901 & 10101030141901 & 2 & 52 & 1 & 1 & 2 & 52
\end{tabular}
```

$\checkmark$ Out of 18,348 ，the number of records satisfying the condition 5 is 17,723 ．

```
# Appended flag5 for inconsistent sex and age
> cond5["flag5"]<-ifelse (cond5$l sex==cond5$sex04 &
+ abs (cond5$ lage-cond5$age04) <=1, 1, 0)
> table(cond5$flag5)
        0 1
    6 2 5 1 7 7 2 3
```

```
# Renamed cond5 as ip1 and generated ID06 and ID04 from PID06 and PID04
> ip1<-cond5
> ip1["ID06"]<-substr(ip1$PID06, 1, 12)
> ip1["ID04"]<-substr(ip1$PID04, 1, 12)
# Example of records with inconsistent sex and age between LIN06 and MEMO4
head(ip1[ip1$flag5==0, c (2, 1, 3, 4, 7, 8,5,6,9)])
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline & PID06 & PID04 & & age & 04 & & & & \\
\hline 1 & ＜NA〉 & 10101030141501 & 1 & 57 & 1 & 50 & 1 & 1 & 0 \\
\hline 59 & ＜NA〉 & 10106070171305 & 2 & 21 & 2 & 23 & 1 & 1 & 0 \\
\hline 125 & 11401 & 10108030011401 & 2 & 44 & 2 & 42 & 1 & 1 & 0 \\
\hline
\end{tabular}
```

| 129 | 10108030011501 | 10108030011501 | 1 | 46 | 1 | 44 | 1 | 1 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 174 | 10109330101501 | 10109330101901 | 1 | 42 | 2 | 49 | 1 | 0 |
| 176 | 10109330101502 | 10109330101902 | 2 | 41 | 2 | 18 | 1 | 0 |

## Interpretation of errors:

- The pair of PID06=10109330101502 and PID04=10109330101902 is not proper.
- For PID04=10109330101902, there are two candidates, PID06=10109330101502 and PID06= 10109330101902 (condition 4).
- The pair of PIDO4=10109330101902 and PIDO6=10109330101902 is proper.
$\checkmark$ Out of 84 records regarded as duplicated under the condition 4, that is, given flag4=0, a half of the 84 records are identified with the records with the same PID04 and the other half are not identified.

Therefore, data set ip2, a subset of records with flag5=1 in ip1 is candidates for panel at the moment.

```
>t<-addmargins(table(ip1$flag4, ip1$flag5)) [c(3, 1, 2), c(3, 2, 1)]
> rownames (t)<-c ("Linked with MEMO4","Duplicated","Not duplicated")
> colnames(t)<-c ("Total","Identified", "Not identified")
t
    Total Identified Not identified
    Linked with MEMO4 18348 17723 625
    Duplicated 
    Not duplicated 18264 17681 583
> ip2<-subset(ip1, flag5==1)
dim(ip2)
[1] 17723 11
```

$\checkmark$ Again, applied the condition 4 for ip2 and found one duplication. We cannot distinguish from data of sex and age because they both satisfy the condition 5 . We have to check household composition.

```
> sum(duplicated(ip2$PID04))
```

[1] 1
>ip2[ip2\$PID04==ip2[duplicated(ip2\$PID04), "PID04"], c ("PID06", "ID06", "PID04", "ID04")]
$\begin{array}{llll}\text { PID06 } & \text { ID06 } & \text { PID04 } & \text { ID04 }\end{array}$
175988190707001130281907070011381907070011502819070700115 175998190707001150281907070011581907070011502819070700115

```
# Household composition of ID04=="819070700115"
> MEM04[MEM04$ID04=="819070700115", c ("PID04", "ID04", "m1ac2", "m1ac5")]
    PID04 ID04 mlac2 mlac5
3886081907070011504819070700115 2 9
3886 81907070011503 819070700115 1 11
3886281907070011502819070700115 2 30
3886381907070011501819070700115 1 30
# Household composition of ID06=="819070700113"
> MEM06[MEM06$ ID06=="819070700113", c ("PID06", "ID06", "m1ac2", "m1ac5")]
    PID06 ID06 m1ac2 m1ac5
3 7 5 6 1 8 1 9 0 7 0 7 0 0 1 1 3 0 1 8 1 9 0 7 0 7 0 0 1 1 3 ~ 1 ~ 3 4 ,
3756281907070011302819070700113 2 32
3756381907070011303819070700113 2 2
# Household composition of ID06=="819070700115"
> MEM06 [MEM06$ ID06=="819070700115", c ("PID06", "ID06", "m1ac2", "m1ac5")]
    PID06 ID06 m1ac2 m1ac5
3757181907070011501819070700115 1 32
3757281907070011502819070700115 2 32
3757381907070011503819070700115 1 13
3 7 5 7 4 8 1 9 0 7 0 7 0 0 1 1 5 0 4 8 1 9 0 7 0 7 0 0 1 1 5 ~ 2 ~ 1 1
```

$\checkmark$ As a result of comparing household composition, the pair of PID06="81907070011502" and PID04="81907070011502" is right. The pair of PID06="81907070011302" and PID04="81907070011502" should be dropped from ip2.

```
> ip2[ip2$PID04==ip2[duplicated(ip2$PID04),"PID04"], c("PID06", "ID06", "PID04", "ID04")]
    PID06 ID06 PID04 ID04
17598 8190707001130281907070011381907070011502819070700115 \leftarrowShould be dropped!
17599 81907070011502819070700115 81907070011502 819070700115 \leftarrowRight!
```

```
> ip2[!is.na(ip2$PID06)&ip2$PID06=="81907070011302", c("PID06", "ID06", "PID04", "ID04")]
    PID06 ID06 PID04 ID04
1 7 5 9 8 8 1 9 0 7 0 7 0 0 1 1 3 0 2 8 1 9 0 7 0 7 0 0 1 1 3 8 1 9 0 7 0 7 0 0 1 1 5 0 2 8 1 9 0 7 0 7 0 0 1 1 5
> ip3<-ip2[is.na(ip2$PID06)|ip2$PID06!="81907070011302",]
> dim(ip3)
[1] 17722 11
> sum(duplicated(ip3$PID04))
[1] 0
```

Condition 6: ID06 in LIN06 should be the same as ID06 in hp.
$\checkmark \quad$ As for LIN06, a set of ID06 is the same as a set of ID06 in HLIN.
$>$ table (unique (LIN06\$ID06) $==$ HLIN\$ ID06)
TRUE
4298
$\checkmark \quad$ As for ip3, the number of unique ID06 is 4,167 .
$\checkmark$ Appended flag6 to hp2.
>ip3. ID06<-unique (ip3\$ID06)
> length(ip3. ID06)
[1] 4167
> hp2["flag6"]<-ifelse(is. element (hp2\$ID06, ip3. ID06), 1, 0)
$>$ head (hp2)
ID06 ID04 flag6
$1 \quad 101010301415101010301415 \quad 1$
$2101010301419101010301419 \quad 1$
161010123018131010123018131
311010511002131010511002131
321010511002151010511002151
$33101051100219101051100219 \quad 1$
> addmargins (table (hp2\$flag6))
01 Sum
4041534193
$\checkmark \quad$ Defined hp3 as a subset of hp2 satisfying the condition 6.

```
> hp3<-subset(hp2, flag6==1)
dim(hp3)
```

[1] 41533

Condition 7: If a person in LIN06 was a household member in VHLSS 2006, that is, PID06 was defined in LIN06, the person with the PID06 should be in MEMO6. (Satisfied!)
$\checkmark \quad$ If a person in LIN06 is a household member of VHLSS 2006, the person belongs to MEM06.

```
>t<-table(is.na(LIN06$PID06), !is. element(LIN06$PID06, MEM06$PID06))
> rownames(t)<-c ("Member in LIN06", "No")
> colnames (t)<-c ("Belong to MEMO6", "No")
> t
    Belong to MEMO6 No
    Member in LIN06 17076 0
    No 0 1716
```


## Condition 8: There should be no duplicated PID06 in LIN06. (Satisfied!)

```
> sum(duplicated(subset(LIN06, m1bc6==1) $PID06))
```

[1] 0

Condition 9: The person' s sex and age in LIN06 and MEMO6 should be identical.

```
>df1<-ip3[c ("PID04", "PID06","Isex","lage")]
> dim(df1)
[1] 17722 4
> head(df1)
\begin{tabular}{lrrrr} 
& \multicolumn{2}{c}{ PID04 } & PID06 & Isex \\
& lage \\
2 & 10101030141502 & 10101030141501 & 2 & 46 \\
3 & 10101030141503 & 10101030141502 & 1 & 26 \\
4 & 10101030141504 & 10101030141503 & 1 & 24
\end{tabular}
```

|  | 510101030141505 | 10101030141504 | 1 | 18 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 610101030141901 | 10101030141901 | 2 | 52 |  |  |
|  | 710101030141902 | 10101030141902 | 1 | 24 |  |  |
| >df2<-MEM06[c ("PID06", "m1ac2", "m1ac5")] |  |  |  |  |  |  |
| > colnames (df2) <-c ("PID06", "sex06", "age06") |  |  |  |  |  |  |
| $>\operatorname{dim}(\mathrm{df} 2)$ |  |  |  |  |  |  |
| [1] 184193 |  |  |  |  |  |  |
| > head (df2) |  |  |  |  |  |  |
| PID06 sex06 age06 |  |  |  |  |  |  |
| 110101030141501248 |  |  |  |  |  |  |
| 210101030141502128 |  |  |  |  |  |  |
| 310101030141503126 |  |  |  |  |  |  |
| 410101030141504120 |  |  |  |  |  |  |
| 510101030141901254 |  |  |  |  |  |  |
| 610101030141902126 |  |  |  |  |  |  |
| > cond9<-merge (df1, df2, by="PID06", all. $\mathrm{x}=\mathrm{T}$ ) |  |  |  |  |  |  |
| $>\operatorname{dim}$ (cond9) |  |  |  |  |  |  |
| [1] 177226 |  |  |  |  |  |  |
| $>$ head (cond9) |  |  |  |  |  |  |
|  | PID06 | PID04 | sex | lage | 06 |  |
|  | 110101030141501 | 10101030141502 | 2 | 46 | 2 | 48 |
| 2 | 210101030141502 | 10101030141503 | 1 | 26 | 1 | 28 |
| 3 | 310101030141503 | 10101030141504 | 1 | 24 | 1 | 26 |
|  | 410101030141504 | 10101030141505 | 1 | 18 | 1 | 20 |
| 5 | 510101030141901 | 10101030141901 | 2 | 52 | 2 | 54 |
|  | 610101030141902 | 10101030141902 | 1 | 24 | 1 | 26 |

$\checkmark$ Out of 17,722 , the number of records satisfying the condition 9 is 16,103 .
\# Appended flag9 for inconsistent sex and age
> cond9["flag9"]<-ifelse (cond9\$ Isex==cond9\$sex06 \&

+ abs (cond9\$lage+2-cond9\$age06) <=2, 1, 0)
$>$ table (cond9\$flag9)

01
1816102

$\checkmark$ Generated data frame ip4 consisted of pairs of PID06 and PID04 as a subset of cond9 with flag9=1. Each pair of PID06 and PID04 identifies uniquely records in MEM06 and MEM04. > ip4<-subset (cond9, flag9==1) [c ("PID06", "PID04")]
> ip4["ID06"]<-substr (ip4\$PID06, 1, 12)
>ip4["ID04"]<-substr (ip4\$PID04, 1, 12)
$>\operatorname{dim}(i p 4)$
[1] 161024
$>$ head (ip4)
$\begin{array}{llll}\text { PID06 PID04 ID06 } & \text { ID04 }\end{array}$
11010103014150110101030141502101010301415101010301415

```
210101030141502 10101030141503101010301415 101010301415
31010103014150310101030141504101010301415 101010301415
4 1 0 1 0 1 0 3 0 1 4 1 5 0 4 1 0 1 0 1 0 3 0 1 4 1 5 0 5 1 0 1 0 1 0 3 0 1 4 1 5 1 0 1 0 1 0 3 0 1 4 1 5
510101030141901 10101030141901 101010301419101010301419
61010103014190210101030141902 101010301419101010301419
```

$\checkmark$ Comparing household size of individual panel data in 2004 with HLD04, 2,821 households are inconsistent but in 1,344 households the number of records of panel data within the household is smaller than the household size of the same household surveyed in 2004.

```
> t<-tapply(ip4$PID04, ip4$ID04, length)
> ip4. hhsz<-data. frame(ID04=names (t), ip. hhsz=as.vector (t), row. names=NULL)
> ip4. hhsz2<-merge(ip4. hhsz, HLD04[c ("ID04", "tsnguoi")], by="ID04", alI. x=T)
> head(ip4. hhsz2)
    ID04 ip.hhsz tsnguoi
1101010301415 4 5
2 101010301419 3 3
3101012301813 4 4
4101051100213 2 6
5101051100215 6 6
6 101051100219 4 4
> table(ip4. hhsz2$ip.hhsz==ip4. hhsz2$tsnguoi, useNA="ifany")
FALSE TRUE
1344 2821
```

$\checkmark \quad$ Finally, the individual-level panel data consisted of 11,411 pairs of (PID06, PID04) were generated, which is consistent with the household size in 2004.

```
> ip4. hhsz["flag"]<-ifelse(ip4. hhsz2$ip. hhsz==ip4. hhsz2$tsnguoi, 1, 0)
> table(ip4.hhsz$flag)
    0 1
13442821
> ip5.ID04<-subset(ip4. hhsz, flag==1) $ID04
> length(ip5.ID04)
[1] }282
```

```
> ip5<-subset(ip4, is.element(ip4$ID04, ip5. ID04))
dim(ip5)
[1] 11411 4
> head(ip5)
    PID06 PID04 ID06 ID04
5 10101030141901 10101030141901 101010301419101010301419
6 10101030141902 10101030141902 101010301419 101010301419
71010103014190410101030141903101010301419101010301419
8 10101230181301 10101230181301 101012301813101012301813
9 10101230181302 10101230181302 101012301813101012301813
101010123018130310101230181303101012301813101012301813
```


## Example:

Pane I data ID06=" 101010301419" and ID04=" 101010301419" Household members of IDO4=="101010301419" in 2004 and 2006

## MEM04 :

```
> d<-MEM04[MEM04$ ID04=="101010301419", c (51, 10:15)]
> colnames (d)<-c ("PID04", "sex", "relation", "month", "year", "age", "marital")
>d
    PIDO4 sex relation month year age marital
\begin{tabular}{llllllll}
10 & 10101030141901 & 2 & 1 & 1 & 1952 & 52 & 3 \\
11 & 10101030141902 & 1 & 3 & 1 & 1980 & 24 & 1 \\
1210101030141903 & 1 & 3 & 2 & 1987 & 17 & 1
\end{tabular}
```


## MEM06:

```
>d1<-MEM06[MEM06$ ID06=="101010301419", c (23, 7:12)]
> colnames (d1)<-c ("PID06", "sex", "relation", "month", "year", "age", "marital")
> d1
    PID06 sex relation month year age marital
\begin{tabular}{rlllrrrr}
5 & 10101030141901 & 2 & 1 & 1 & 1952 & 54 & 3 \\
610101030141902 & 1 & 3 & 1 & 1980 & 26 & 2 \\
7 & 10101030141903 & 2 & 3 & 11 & 1980 & 25 & 2 \\
8 & 10101030141904 & 1 & 3 & 2 & 1987 & 19 & 1
\end{tabular}
```

$\checkmark \quad$ In the above case, the household consisted of a widowed mother and two sons of age 24 and 17 in 2004. The elder son got married during 2004 and 2006, and his wife was surveyed in 2006.

```
\ Topic:Household size in 2004
```

Compar ing LIN06 with HLD04, the household sizes in 2004 of LIN06 are consistent in 4, 066 households.

```
> t<-tapply (LIN06$PID04, LIN06$ID04, length)
```

> ip. hhsz<-data. frame (ID04=names ( t ), ip. hhsz=as. vector ( t ), row. names=NULL)
> ip. hhsz2<-merge(ip. hhsz, HLD04[c ("ID04", "tsnguoi")], by="ID04", all. x=T)
$>$ head (ip. hhsz2)

|  | ID04 ip. hhsz tsnguoi |  |  |
| :--- | :--- | ---: | ---: |
| 1 | 101010301415 | 5 | 5 |
| 2 | 101010301419 | 3 | 3 |
| 3 | 101012301813 | 4 | 4 |
| 4 | 101051100213 | 6 | 6 |
| 5 | 101051100215 | 6 | 6 |
| 6 | 101051100219 | 4 | 4 |

> table(ip. hhsz2\$ip. hhsz==ip. hhsz2\$tsnguoi, useNA="ifany")
FALSE TRUE 〈NA〉
14006679
> Topic 2: Difficulties of matching individual data
Case of ID04=="101010301415"

## MEM04:

```
>d<-MEM04[MEM04$ ID04=="101010301415", c (51, 10:15)]
> colnames (d)<-c ("PID04", "sex", "relation", "month", "year", "age", "marital")
>d
    PIDO4 sex relation month year age marital
\begin{tabular}{rlllrlll}
5 & 10101030141501 & 1 & 1 & 5 & 1954 & 50 & 2 \\
610101030141502 & 2 & 2 & 7 & 1957 & 46 & 2 \\
710101030141503 & 1 & 3 & 111977 & 26 & 1 \\
810101030141504 & 1 & 3 & 31980 & 24 & 1 \\
910101030141505 & 1 & 3 & 31986 & 18 & 1
\end{tabular}
```

LIN06:

```
> d1<-LIN06[LIN06$ID04=="101010301415", c ("PID04", "m1bc4", "m1bc5", "m1bc6")]
> colnames (d1)<-c ("PID04", "sex", "age", "member in 2006")
> d1<-d1[order (d1$PID04),]
>d1
\begin{tabular}{llll}
\multicolumn{4}{c}{ PID04 sex age member in 2006} \\
5 & 10101030141501 & 1 & 57 \\
1 & 10101030141502 & 2 & 46 \\
3 & 10101030141503 & 1 & 26
\end{tabular}
```

$\checkmark$ In the above case, the age 57 of PIDO4=" 10101030141501 " in LIN06 should be read as 50 . It may be possible to revise errors manually in such way. However, how to make it programmable?

## Example of usage at household level:

Changes of consumption expenditure per capita between 2004 and 2006
\# Household-level panel
> hp5<-ip5[!duplicated(ip5\$ID06),c("ID06","ID04")]
> HLD06["pcexp06"]<-HLD06\$chidsbq
> HLD04["pcexp04"]<-HLD04\$chidsbq
\# Linked the pcexp data of HLD06 and HLD04
> pcexp.panel<-merge(hp5,HLD06[c("ID06","pcexp06")])
> pcexp.panel<-merge(pcexp.panel,HLD04[c("ID04","pcexp04")])
$>\operatorname{dim}($ pcexp.panel)
[1] 28214
\# Generated the variable of increasing rate of pcexp between 2004 and 2006.
> pcexp.panel["rate"]<-(pcexp.panel\$pcexp06/pcexp.panel\$pcexp04-1)*100
$>$ head (pcexp. panel)
ID04 ID06 pcexp06 pcexp04 rate

| 1 | 101010301419 | 101010301419 | 1226 |
| ---: | :--- | ---: | ---: |
| 2 | 101012301813 | 101012301813 | 1111 |
| 3 | 101051100215 | 101051100215 | 430 |
| 4 | 101051100219 | 101051100219 | 946.48649 |
| 4 | 973 | 15.28150 |  |


| 5 | 101053301214 | 101053301214 | 778 |
| :--- | :--- | :--- | :--- |
| 6 | 101053301215 | 101053301215 | 968 |

## Summary:

$\checkmark$ Regarding VHLSS 2004 and 2006, household-level panel consisted of 4,193 household identifiers was generated.
$\checkmark$ Also, individual-level panel consisted of $\mathbf{1 1 , 4 1 1}$ person identifier was generate.
$\checkmark \quad$ To make panel more complete, data entry and data check of identification items should be carried out more carefully.

## 6. Population and Household

### 6.1 Ethnicity

$\checkmark$ Kinh people are the majority ethnic group of Vietnam.
Regarding household head's ethnicity, 88\% of households are Kinh. The share of Kinh households is high especially in Red River Delta (99\%), North Central (91\%), South Central Coast (95\%), South East (92\%) and Mekong River Delta (93\%).
On the other, the share of Kinh is only $11 \%$ in rural of North West.

```
> d<-ttchung [c ("ID", "wt", "tinh", "ttnt", "tsnguoi", "dantoc", "phdich")]
# Generated the variable kinh: 1=Kinh, 2=Non-kinh
>d["kinh"]<-ifelse(d$dantoc==1, 1, 2)
> table(d$kinh)
    1 2
77451444
# Weighted number of households by region and urban/rural
> denominator<-addmargins(tapply (d$wt, list(substr (d$tinh, 1, 1),d$ttnt), sum, na. rm=T))
# Weighted number of Kinh households by region and urban/rural
>d1<-subset (d, kinh==1)
> numerator<-addmargins(tapply (d1$wt, list(substr(d1$tinh, 1, 1), d1$ttnt), sum, na. rm=T))
```

\# Share of Kinh households by region and urban/rural
> share<-round (numerator/denominator*100, 1) [c (9, 1:8) , c (3, 1, 2) ]
> colnames (share) <-c ("Total", "Urban", "Rural")
$>$ rownames (share) <-c ("Vietnam", region. name)
$>$ share

Total Urban Rural
$\begin{array}{llll}\text { Vietnam } & 87.8 & 94.3 & 85.3\end{array}$
Red River Delta $\quad 99.4100 .0 \quad 99.2$
$\begin{array}{llll}\text { North East } & 62.4 & 85.7 & 55.8\end{array}$
$\begin{array}{llll}\text { North West } & 24.2 \quad 84.6 \quad 11.1\end{array}$

| North Central | 91.2 | 98.2 | 90.0 |
| :--- | :--- | :--- | :--- |
| South Central Coast | 94.8 | 98.7 | 93.1 |
| Central Highlands | 71.1 | 86.2 | 64.3 |
| South East | 92.1 | 92.4 | 91.6 a |
| Mekong River Delta | 93.2 | 95.0 | 92.7 |

```
\checkmark ~ O u t ~ o f ~ 1 , 4 4 4 ~ n o n - k i n h ~ s a m p l e ~ h o u s e h o l d s , ~ 3 8 1 ~ h o u s e h o l d s ~ n e e d e d ~ t h e ~ i n t e r p r e t a t i o n ~ s e r v i c e
    when conducting the survey.
    people.
> m<-tapply(d$wt, list (d$kinh, d$phdich), length)
>m[is.na(m)]<-0
>m<- addmargins(m)[c(3,1,2), c(3, 1, 2)]
> rownames (m)<-c ("Total", "Kinh", "Non-kinh")
colnames (m)<-c ("Total", "Interpretation required", "No")
>m
            Total Interpretation required No
\begin{tabular}{llr} 
Total & 9189 & 3818808 \\
Kinh & 7745 & 07745
\end{tabular}
```

    It suggests that about one fourth non-Kinh households have communication problem with Kinh
    
### 6.2 Religion

Ethnicity and religion are the very important factors when analyzing social and cultural structure of the country. In VHLSS, question on religion is not found in the questionnaire.

According to 2009 Population Census, only a small fraction of the Vietnamese adheres to institutional religions;

| Non religious, | $80.8 \%$ |
| :--- | :--- |
| Buddhism, | $9.3 \%$ |
| Christianity, | $7.2 \%$ |
| $\quad$ Roman Catholic, | $6.7 \%$ |
| $\quad$ Protestant, | $0.5 \%$ |
| Hòa Hảo, | $1.5 \%$ |
| Cao Đài, | $1.1 \%$ |

"Although according to a 1999 census most Vietnamese list themselves as having no religious affiliation, religion, as defined by shared beliefs and practices, remains an integral part of Vietnamese life, dictating the social behaviours and spiritual practices of Vietnamese individuals in Vietnam and abroad. The triple religion (Vietnamese: tam giáo), referring to the syncretic combination of Mahayana Buddhism, Confucianism, and Taoism remains a strong influence on the beliefs and practices of the Vietnamese, even if the levels of formal membership in these religious communities may not reflect that influence. One of the most notable and universal spiritual practices common to Vietnamese is ancestor veneration, a practice shared with Chinese and most other Asian cultures. Practically all Vietnamese, regardless of formal religious affiliation, have an altar in their home or business where prayers are offered to their ancestors. These offerings and practices are done frequently during important traditional or religious celebrations (e.g., death anniversaries), the starting of a new business, or even when a family member needs guidance or counsel. Belief in ghosts and spirits is common; it is commonly believed that failing to perform the proper rituals for one's ancestors will cause them to become hungry ghosts (Vietnamese: ma đói)." (Wikipedia)

### 6.3 Age heaping

Age heaping is popular in developing countries’ censuses and surveys.
The below chart shows the number of sample household members of VHLSS 2006 by sex and single year of age.
It is understood that age heaping is not significant in Vietnam. The reason might be that they use twelve signs of Chinese Zodiac in daily life, which is listed in the questionnaire.

Chart Number of sample household members by sex and single year of age (2006)


```
> t<-tapply(muc1a$PID,list(muc1a$m1ac5,muc1a$m1ac2),length)
> dim(t)
[1] 104 2
> colnames(t)<-c("Male","Female")
> m<-data.frame(age=rownames(t),t)
> head(m)
    age Male Female
0}00224 22
1 1 265 224
2 2 254 220
3 3 239 267
4 4 238 210
5 5 273 253
> write.csv(m,"pop_by_age_sex.csv",row.names=F)
```

TABLE OF THE LUNAR-SOLAR CALENDAR YEAR EQUIVALENCE

| Mouse | 1900 | 1912 | 1924 | 1936 | 1948 | 1960 | 1972 | 1984 | 1996 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Buffaio | 1901 | 1913 | 1925 | 1937 | 1949 | 1961 | 1973 | 1985 | 1997 |
| Tiger | 1902 | 1914 | 1926 | 1938 | 1950 | 1962 | 1974 | 1986 | 1998 |
| Cat | 1903 | 1915 | 1927 | 1939 | 1951 | 1963 | 1975 | 1987 | 1999 |
| Dragon | 1904 | 1916 | 1928 | 1940 | 1952 | 1964 | 1976 | 1988 | 2000 |
| Snake | 1905 | 1917 | 1929 | 1941 | 1953 | 1965 | 1977 | 1989 | 2001 |
| Horse | 1906 | 1918 | 1930 | 1942 | 1954 | 1966 | 1978 | 1990 | 2002 |
| Goat | 1907 | 1919 | 1931 | 1943 | 1955 | 1967 | 1979 | 1991 | 2003 |
| Monkey | 1908 | 1920 | 1932 | 1944 | 1956 | 1968 | 1990 | 1992 | 2004 |
| Cock | 1909 | 1921 | 1933 | 1945 | 1957 | 1969 | 1981 | 1993 | 2005 |
| Dog | 1910 | 1922 | 1934 | 1946 | 1958 | 1970 | 1982 | 1994 | 2006 |
| Pig | 1911 | 1923 | 1935 | 1947 | 1959 | 1971 | 1983 | 1995 |  |

Note: Buffalo is the second symbol, taking place of Ox in the Chinese zodiac.
Also the fourth symbol of Rabbit is replaced with Cat.

### 6.4 Literacy

$\checkmark \quad$ In section 2A (Education) of the questionnaire, literacy was asked for those who finished 4-th grade or below, and never go to school. Those who finished 5-th grade (primary school) or upward are regarded as literacy.
$\checkmark \quad$ Generated the variable literate for all household members
1: literate
2: Illiterate
$>\mathrm{d}<-$ muc2a
> literate<-ifelse (d\$m2ac1>=5|d\$m2ac2==1, 1, 2)
\# Number of records by literacy
$>$ table(literate)
literate
12
332115860
$\checkmark \quad$ Literacy rate of Vietnam is $86 \%$.
$>$ round $($ prop. table $($ tapply $(d \$ w t$, literate, sum, na. $\mathrm{rm}=\mathrm{T})) * 100,1)$
12
86.113 .9
$\checkmark \quad$ Merged literate with MUC1A, and appended data of urban/rural and ethnicity in TTCHUNG.

```
>df<-merge(cbind(muc1a, literate),ttchung[, c("ID", "ttnt", "dantoc")], key="ID", all. x=T)
>df["kinh"]<-ifelse(df$dantoc==1, 1, 2)
```

$\checkmark \quad$ Literacy rate by region and urban/rural
Literacy rate is low in rural of North West (67\%) and rural of Central Highlands (77\%).

```
> denominator<-addmargins(tapply (df$wt, list (substr (df$tinh, 1, 1), df$ttnt), sum, na. rm=T))
>d2<-subset (df, literate==1)
> numerator<-addmargins(tapply (d2$wt, list (substr (d2$tinh, 1, 1), d2$ttnt), sum, na. rm=T))
> rate<-round (numerator/denominator*100, 1)[c (9, 1:8), c (3, 1, 2)]
> colnames(rate)<-c ("Total", "Urban", "Rural")
> rownames(rate)<-c ("Vietnam", region. name)
> rate
    Total Urban Rural
Vietnam 
Red River Delta }\quad90.0\quad91.9 89.4 
North East }\quad85.8 92.3 84.2 
North West 70.5 90.7 67.2
North Central 87.2 92.6 86.3
South Central Coast 87.4 89.5 86.5
Central Highlands }\quad79.7 87.3 76.5 
South East 
Mekong River Delta 83.6 85.8 83.1
```

$\checkmark \quad$ Literacy rate by ethnicity
Literacy rate of non-Kinh people is $15 \%$ points lower than Kinh people.
Range of Kinh people's literacy rates by region is narrower compared to non-Kinh people.

```
> denominator<-addmargins(tapply (df$wt, l ist(substr (df$t inh, 1, 1), df$k inh), sum, na. rm=T))
d2<-subset (df, literate==1)
> numerator<-addmargins(tapply (d2$wt, l i st (substr (d2$t inh, 1, 1), d2$k inh), sum, na. rm=T))
> rate<-round (numerator/denominator*100, 1)[c(9, 1:8), c (3,1,2)]
> colnames(rate)<-c ("Total", "Kinh", "Non-kinh")
> rownames(rate)<-c ("Vietnam", region. name)
> rate
                                    Total Kinh Non-kinh
Vietnam 86.1 88.5 72.3
Red River Delta 90.0 90.0 88.9
```

| North East | 85.890 .3 | 79.8 |
| :--- | :--- | :--- |
| North West | 70.593 .3 | 65.3 |
| North Central | 87.288 .4 | 76.7 |
| South Central Coast | 87.488 .9 | 64.5 |
| Central Highlands | 79.790 .4 | 59.2 |
| South East | 88.289 .2 | 77.9 |
| Mekong River Delta | 83.684 .8 | 69.2 |

### 6.5 Polygamy

## Summary of Marriage System

VHLSS allows the cases of multiple spouses. Examples of micro data of VHLSS 2006 are shown later.

- The enumerator’s manual "Operational Handbook of VHLSS 2004" describes household members as follows;
"If the household owner has many wives, then enumerators write the names of the first wife and her children, then the second wife and her children and go on in this way." (at page 21)
- In Vietnam, men belonging to the old generation (born around 1930s or earlier) usually had more than one wife. That was the customs of those days.
- According to the Law of Marriage and Family, which has begun since 1959 in North Vietnam at the time and was applied for South Vietnam after the end of War in 1975, one husband has one wife only. Until now, this provision is still in effect. Therefore, people got married before the implementation of the Law could have more spouses if their marriage was carried out before that time.
- Vietnam has the Law of Family Register System. Accordingly, the second spouse cannot be registered as an official spouse at the Local Authority. However, VHLSS is flexible and still accepts the second spouse if the family answer that they are spouses.


## VHLSS 2006 - Example of polygamy

$\checkmark$ Data file MUC1A: List of household members
$>$ head (muc1a)
tinh huyen xa diaban hoso matv m1ac2 m1ac3 m1ac4a m1ac4b m1ac5 m1ac6 m1ac7

| 1 | 101 | 01 | 03 | 014 | 15 | 1 | 2 | 1 | 7 | 1957 | 48 | 3 | 12 |
| ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | ---: | :--- | :--- | :--- | :--- |
| 2 | 101 | 01 | 03 | 014 | 15 | 2 | 1 | 3 | 11 | 1977 | 28 | 1 | 12 |
| 3 | 101 | 01 | 03 | 014 | 15 | 3 | 1 | 3 | 3 | 1980 | 26 | 1 | 12 |
| 4 | 101 | 01 | 03 | 014 | 15 | 4 | 1 | 3 | 3 | 1986 | 20 | 1 | 12 |
| 5 | 101 | 01 | 03 | 014 | 19 | 1 | 2 | 1 | 1 | 1952 | 54 | 3 | 12 |


| 6 | 101 | 0103 | 014 | 19 | 2 |  | 3 | 1 | 1980 | 26 | 2 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | mlac8 m1ac9 mlac10a m1ac10b ID wt xaid PID |  |  |  |  |  |  |  |  |  |  |  |


| 1 | 1 | $N A$ | $N A$ | $N A$ | 101010301415 | 3107.318 | 1010103 | 10101030141501 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 2 | 1 | $N A$ | $N A$ | $N A$ | 101010301415 | 3107.318 | 1010103 | 10101030141502 |
| 3 | 1 | $N A$ | $N A$ | $N A$ | 101010301415 | 3107.318 | 1010103 | 10101030141503 |
| 4 | 1 | $N A$ | $N A$ | $N A$ | 101010301415 | 3107.318 | 1010103 | 10101030141504 |
| 5 | 1 | $N A$ | $N A$ | NA 101010301419 | 3107.318 | 1010103 | 10101030141901 |  |
| 6 | 1 | NA | NA | NA | 101010301419 | 3107.318 | 1010103 | 10101030141902 |

$\checkmark \quad$ Generated subset of spouse.
$>d f<-m u c 1 a[, c(18,6: 8,11,12)]$
> colnames (df) [3:6]<-c ("sex", "relation", "age", "marital")
$>$ head (df)

| ID matv sex relation age marital |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | 101010301415 | 1 | 2 | 1 | 48 | 3 |
| 2 | 101010301415 | 2 | 1 | 3 | 28 | 1 |
| 3 | 101010301415 | 3 | 1 | 3 | 26 | 1 |
| 4 | 101010301415 | 4 | 1 | 3 | 20 | 1 |
| 5 | 101010301419 | 1 | 2 | 1 | 54 | 3 |
| 6 | 101010301419 | 2 | 1 | 3 | 26 | 2 |
| $>$ | spouse<-subset (df, relation==2) |  |  |  |  |  |

$\checkmark \quad$ Made frequency table of the number of spouse within the household.
The results show that there are 4 sample households with two spouses.
>ns<-(tapply (spouse\$ID, spouse\$ID, length))
$>$ table(ns)
ns
12
$7340 \quad 4$

```
\checkmark ~ E x a m p l e ~ o f ~ h o u s e h o l d ~ w i t h ~ m u l t i p l e ~ s p o u s e s ~ w i t h i n ~ t h e ~ h o u s e h o l d ~
# IDs of such households
> names(ns)[ns==2]
[1] "103117100913" "105253900215" "109060100814" "201150700215"
>df[df$ID=="103117100913",]
```

ID matv sex relation age marital

| 1333 | 103117100913 | 1 | 1 | 1 | 80 |
| ---: | :--- | :--- | :--- | :--- | :--- |
| 1334 | 103117100913 | 2 | 2 | 2 | 78 |
| 1335 | 103117100913 | 3 | 2 | 2 | 63 |
| 1336 | 103117100913 | 4 | 1 | 3 | 27 |
| 1337 | 103117100913 | 5 | 2 | 3 | 25 |
| 1338 | 103117100913 | 6 | 1 | 6 | 6 |
| 1339 | 103117100913 | 7 | 1 | 6 | 3 | > df[df\$ID=="105253900215", ]

ID matv sex relation age marital

| 3137 | 105253900215 | 1 | 1 | 1 | 42 | 2 |
| :--- | :--- | :--- | :--- | :--- | :--- | ---: |
| 3138 | 105253900215 | 2 | 2 | 2 | 34 | 2 |
| 3139 | 105253900215 | 3 | 2 | 3 | 11 | $N A$ |
| 3140 | 105253900215 | 4 | 2 | 2 | 30 | 2 |
| 3141 | 105253900215 | 5 | 1 | 3 | 5 | $N A$ |
| 3142 | 105253900215 | 6 | 2 | 3 | 2 | $N A$ |

> df[df\$ID=="109060100814", ]
ID matv sex relation age marital

| 4692 | 109060100814 | 1 | 1 | 1 | 47 | 2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 4693 | 109060100814 | 2 | 2 | 2 | 49 | 2 |
| 4694 | 109060100814 | 3 | 2 | 2 | 46 | 2 |
| 4695 | 109060100814 | 4 | 1 | 3 | 14 | 1 |
| 4696 | 109060100814 | 5 | 2 | 3 | 12 | $N A$ |

> df[df\$ID=="201150700215", ]
ID matv sex relation age marital

| 7705 | 201150700215 | 1 | 1 | 1 | 33 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 7706 | 201150700215 | 2 | 2 | 2 | 42 |
| 7707201150700215 | 3 | 2 | 2 | 43 | 2 |
| 7708 | 201150700215 | 4 | 1 | 3 | 18 |
| 7709201150700215 | 5 | 2 | 3 | 11 | 2 |
| 7710201150700215 | 6 | 2 | 3 | 7 | $N A$ |
| 7711201150700215 | 7 | 2 | 3 | 15 | 2 |
| 7712201150700215 | 8 | 1 | 7 | 13 | 1 |
| 7713201150700215 | 9 | 2 | 7 | 8 | $N A$ |
| 7714201150700215 | 10 | 1 | 7 | 9 | $N A$ |
| 7715201150700215 | 11 | 2 | 3 | 4 | $N A$ |

7716201150700215 12 $12 \quad 3 \quad 2 \quad$ NA

Among 4 cases, the first case with spouses' ages 63 years old and 78 years. Remaining 3 cases, location of which are Red River Delta and North East regions, are spouses with ages around 40-50 years old.

According to Ms. Van,
"Therefore, the first case may belong to the custom of old generation. The other 3 cases may be they violate the Law, but with some special reasons (for example, to have children if the first wife could not produce kids, ...). Because of the preference of son, if the wife has the daughters only, the man could consider to have the second wife, though it is not permitted by the Law."

### 6.6 Relationship to the household head

## Summary

- According to "Operational Handbook of VHLSS 2004", the variable of person number MATV follows the next rule;
- The first person to be written down in the questionnaire is the household owner although this person may not be the respondent or even when this person is not present at home. Household owner is always coded number 1.
- Then followed by household owner's spouse and children who are not yet married from the oldest one to the youngest. If the household owner has many wives, then enumerators write the names of the first wife and her children, then the second wife and her children and go on in this way.
- Then followed by household owner's children who are married, their spouse and children (if any).
- Then come father, mother, adopted brothers and sisters, maternal and paternal grandparents, maternal and paternal grandchildren (whose parents are both not living in the surveyed household) and other relatives of the household owner and his/her spouse.
- Finally, non-relatives.
- The code of relationship differs between VHLSS 2002 and 2004/2006.

For VHLSS 2002

| Code | Relationship to head |
| :--- | :--- |
| 1 | Head |
| 2 | Wife/husband |
| 3 | Child |
| 4 | Child in law |
| 5 | Parents |
| 6 | Sister/brother |
| 7 | Grandfather/grandmother |
| 8 | Grandchild |
| 9 | Other relationship |

For VHLSS 2004 and 2006

| Code | Relationship to head |
| :--- | :--- |
| 1 | Head |
| 2 | Spouse |
| 3 | Children |


| 4 | Parents |
| :--- | :--- |
| 5 | Grandfather/Grandmother |
| 6 | Grandchild |
| 7 | Other relation |
| 9 | Missing |

- The code for "Child in law", that is, son's wife or daughter's husband, seems not to be clearly described in the manual of VHLSS 2004.

For this, Ms. Van describes, "Explaining for enumerator is that code 3 is the children of the head. In other words, code 3 is for the one who is blood/adopted children of the head. Code 7 is for other relationship."

However, micro data of VHLSS 2006 suggests that there are two cases of code=3 (child) and code $=7$ (other relation). Next are examples.

| > df[df\$ID=="715032100614", ] |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ID matv sex relation age marital |  |  |  |  |  |  |
| 30336 | 715032100614 | 1 | 1 | 1 | 50 | 2 |
| 30337 | 715032100614 | 2 | 2 | 2 | 53 | 2 |
| 30338 | 715032100614 | 3 | 1 | 3 | 18 | 1 |
| 30339 | 715032100614 | 4 | 1 | 3 | 15 | 1 |
| 30340 | 715032100614 | 5 | 2 | 3 | 14 | 1 |
| 30341 | 715032100614 | 6 | 1 | 3 | 29 | 2 |
| 30342 | 715032100614 | 7 | 2 | 3 | 26 | 2 |
| 30343 | 715032100614 | 8 | 1 | 6 | 7 | NA |
| 30344 | 715032100614 | 9 | 1 | 6 | 5 | NA |
| 30345 | 715032100614 | 10 | 1 | 6 | 2 | NA |
| 30346 | 715032100614 | 11 | 1 | 6 | 1 | NA |
| 30347 | 715032100614 | 12 | 2 | 3 | 26 | 2 |
| 30348 | 715032100614 | 13 | 1 | 3 | 28 | 2 |
| 30349 | 715032100614 | 14 | 1 | 6 | 3 | NA |
| 30350 | 715032100614 | 15 | , | 3 | 21 | 2 |
| 30351 | 715032100614 | 16 | 2 | 3 | 21 | 2 |
| 30352 | 715032100614 | 17 | 2 | 4 | 76 | 3 |

In the above case of relationship code=3, MATV=7 and 16 are regarded as sons' wives, and MATV=13 is a daughter's husband.

```
>df[df$ID=="303090700214",]
    ID matv sex relation age marital
14437 303090700214 1 1 1 56 2
```

| 14438303090700214 | 2 | 2 | 2 | 53 | 2 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 14439303090700214 | 3 | 1 | 3 | 16 | 1 |
| 14440303090700214 | 4 | 2 | 3 | 14 | 1 |
| 14441303090700214 | 5 | 1 | 3 | 22 | 2 |
| 14442303090700214 | 6 | 2 | 7 | 23 | 2 |
| 14443303090700214 | 7 | 2 | 6 | 6 | NA |
| 14444303090700214 | 8 | 2 | 6 | 4 | NA |
| 14445303090700214 | 9 | 2 | 6 | 1 | NA |
| 14446303090700214 | 10 | 1 | 3 | 20 | 2 |
| 14447303090700214 | 11 | 2 | 7 | 22 | 2 |
| 14448303090700214 | 12 | 1 | 6 | 4 | NA |
| 14449303090700214 | 13 | 1 | 6 | 2 | NA |
| 14450303090700214 | 14 | 1 | 6 | 0 | NA |
| 14451303090700214 | 15 | 2 | 7 | 22 | 2 |
| 14452303090700214 | 16 | 1 | 3 | 18 | 2 |
| 14453303090700214 | 17 | 2 | 7 | 20 | 2 |

In the above case of relationship code=7, MATV=6, 11 and 17 are regarded as sons' wives. MATV=15 is supposed to be a son's wife while the son (her husband) is away from home.

In summary,

| Types of code for spouse of son/daughter | Number of sample households |
| :--- | :---: |
| Case 1: code 3 was applied for daughter's husband. | 115 |
| Case 1: code 3 was applied for son's wife. | 668 |
| Case 2: code 7 was applied for daughter's husband. | 36 |
| Case 2: code 7 was applied for son's wife. | 289 |

```
# Made use of data frame ft, which was created in 6.7.1.
> child. couple<-function(df, k, j) {
# return 2 if the code of son/daughter' s spouse is 3.
# return 4 if the code of son/daughter' s spouse is 4.
+ flag<-0
+ if(j>=2&df[k, j+27]==2&df[k,j+28]==2&df[k,j+44]*df[k,j+45]==2&
+ (df[k, j+10]==3&df[k, j+11]==3)) {
+ flag<-2
+ }else if(j>=2&df[k, j+27]==2&df[k,j+28]==2&df[k,j+44]*df[k, j+45]==2&
+ (df[k, j+10]==3&df[k, j+11]==7)){
+ flag<-4
+ }
+ return(flag)
```

```
+ }
>childinlaw<-data.frame(ID=ft$ID, matrix (0, nrow=nrow(ft),ncol=5))
> colnames (childinlaw)<-c ("ID", "m3", "f3", "m7", "f7", "other")
>for(k in 1:nrow(ft)){
+ for(j in 1:ft[k, 2]) {
# Add 1 to childinlaw() if the spouse if female
+ x<-child. couple(ft,k, j)+(ft[k, j+45]==2)
+ if(x>0) childinlaw[k, x]<-childinlaw[k, x]+1
+ }
+}
sum(childinlaw$m3)
[1] }11
>sum(childinlaw$f3)
[1] 668
> sum(childinlaw$m7)
[1] 36
>sum(childinlaw$f7)
[1] 289
>sum(childinlaw$other)
```

[1] 0

- When discussing the family type, the sequence of person number MATV is indispensable for determining a couple.
- The two household members within the household who satisfy the next conditions are presumed as a couple in VHLSS 2006.

1) Their person numbers (MATV) are consecutive.
2) Their marital statuses are married.
3) Their sexes are opposite.
4) Their relationship codes are the same, or the first person's code is 3 (Child) and the second person's code is 7 (Other relation).

- "Brother/sister", "other relatives" and "non-relatives" are not distinguished clearly in the questionnaire.


## Background of the relationship codes

It is said that the traditional Vietnamese family is patriarchal, especially among Kinh people. Traditional values of Vietnamese lifestyle were deeply affected by Confucian ethics. During thousand years the Chinese invaded and maintained control Vietnam, Vietnamese culture was permeated by their Confucian philosophical beliefs.

Regarding the reason why the response categories of relationship was simplified in VHLSS 2004/2006, Ms. Van replied, "We distinguish the direct relationship within family. The direct relationship with the head is important to be distinguished clearly."

## For reference:

## Relationship to the household head in Population Census

- In Vietnam 1999 Population Census, the response categories of the question on the relationship to the household head are as follows. Code 3 "Son/daughter" is limited only for biological child.

1 Household head
2 Spouse
3 Son/Daughter
4 Parents
5 Other

- The definition of each response category is described in the Enumerators' manual;

Question 2: What is the relationship of (name) to the household head?
If respondent is the head of household, interviewers fill in code "1". For the rest of household members, interviewers fill in corresponding codes 2 to 5 depending on their relationship with the head of household. Definition of each relationship to head of household is, as follow:

Head of household: is a household representative who is recognized by all members of household.
In households where there are only children because their father and mother work in ministry of defense and/or ministry of public security (enumerating in special plan), head of household is the oldest child.

For students of professional schools or college, students of stay-in general schools, interviewers register them by their room/apartment as one household. Head of household is a person who is recognized by all household's members. Other household's members would have "other" relationship to the head. Spouse: if a person has two spouses or more (wives or husbands) and lives together with them in one household, these persons are all considered as spouses (wives or husbands) of the head of household (see question 15). Interviewers select code " 2 " if a person is a spouse of the head of household.

Biological child: is child by birth of the head of household. Interviewers select code " 3 " if a person is defined as biological child.

Parent: Father (mother) of the head of household includes biological father (mother), adopted father (mother), father-in-law and mother-in-law of the head of household.

Other relationship: Interviewers should report specific relationship with the head of household, such as stepchild of the head of household, daughter-in-law/son-in-law, adopted child, other brother, older sister, younger brother/sister, aunt, uncle, grandparent, grandchild, domestic employees, maid, friend, etc. A person is defined as "other relationship" with the head of household has code " 5 ".

### 6.7 Family type

It is said that the Vietnamese household traditionally followed the extended multi-generational pattern. The parents, their sons and their wives, their children, and unmarried siblings usually constituted a Vietnamese household.

## Objectives:

1) To determine the family type of each household.
2) To analyze living standards by family type.

### 6.7.1 Family type based on the number of persons by generation

## Data file for determining the family type (1)

$\checkmark \quad$ Designed household-level data file consist of number of persons by relationship as well as data of relationship, marital status, sex and age of each person.

Layout of record

| ID | Number of persons by relationship |  |  |  |  |  |  |  |  |  | Relationship of each person |  |  |  |  | Marital status of each person |  |  |  |  | Sex of each person |  |  |  |  | Age of each person |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | T | 1 | 2 | 3 | 4 | 5 | 6 | 7 | x | 9 | 1 | 2 | 3 | $\ldots$ | 17 |  |  |  |  |  | 1 | 2 | 3 | $\ldots$ | 17 | 1 | 2 | 3 | $\ldots$ | 17 | 1 | 2 | 3 | $\ldots$ | 17 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

```
\checkmark Verified MUC1A
>df<-muc1a[, c(18,6:8,11,12)]
> colnames (df) [3:6]<-c ("sex", "relation", "age", "marital")
> table(df$sex, useNA="ifany")
    2
19157 19914
> table(df$relation, useNA="'ifany")
\begin{tabular}{rrrrrrrr}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 9 \\
9189 & 7348 & 18275 & 952 & 45 & 2203 & 1057 & 2
\end{tabular}
> table(subset (df, age>=13) $mar ital, useNA=" i fany")
```

```
    1 2 3 4 5
10880 18029 2028 265 135
>table(is.na(df$age))
FALSE
39071
```

$\checkmark \quad$ Created data frame ft: family type

```
# household size
> hhsize<-tapply (df$ID, df$ID, length)
> ft<-data. frame(ID=names (hhsize), hhsz=hhsize, row. names=NULL)
# Number of persons by relationship
>for(j in c(1:7,9)){
+ d<-subset(df, relation==j)
+ nj<-tapply (d$ID, d$ID, length)
+ dj<-data. frame(ID=names(nj),nj)
+ ft<-merge(ft, dj, by. x="ID", by. y="ID", all. x=T)
+ colnames (ft)[ncol(ft)]<-paste ("n", j, sep="")
+ }
dim(ft)
[1] 9189 10
 head(ft)
    ID hhsz n1 n2 n3 n4 n5 n6 n7 n9
101010301415 4 1 NA 3 NA NA NA NA NA
2101010301419 4 1 NA 3 NA NA NA NA NA
3101010301424 4 1 NA NA NA NA NA 3 NA
4101010901913 2 1 1 NA NA NA NA NA NA
5101010901915 3 1 1 1 1 NA NA NA NA NA
6101010901919 3 1 1 1 1 NA NA NA NA NA
>ft.old<-ft
# Relationship, marital status, sex and age by person number
> for(j in 1:17){
+ d<-subset(df,matv==j)[,c(1,4,6,3,5)]
+ colnames (d) <-c ("ID", paste (c ("r", "m", "s", "a"), j, sep=""))
```

```
+ ft<-merge(ft, d, by. x="ID", by. y="ID", a|l.x=T)
+ }
>dim(ft)
[1] 9189 78
ft.old2<-ft
colnames (ft)
```

| [1] "ID" | "hhsz" | "n1" | "n2" | "n3" | "n4" | "n5" | "n6" | "n7" | "n9" |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| [11] "r1" | " m 1 " | "s1" | "a1" | "r2" | "m2" | "s2" | "a2" | "r3" | "m3" |
| [21] "s3" | "a3" | "r4" | "m4" | "s | "a4" | "r5" | "m5" | "s5" | "a5" |
| [31] "r6" | "m6" | "s6" | "a6" | "r7" | "m7" | "s7" | "a7" | "r8" | "m8" |
| [41] "s8" | "a8" | "r9" | "m9" | "s9" | "a9" | "r10" | "m10" | "s10" | "a10" |
| [51] "r11" | "m11" | "s11" | "a11" | "r12" | "m12" | "s12" | "a12" | "r13" | "m13" |
| [61] "s13" | "a13" | "r14" | "m14" | "s14" | "a14" | "r15" | "m15" | "s15" | "a15" |
| 71] "r16" | "m16" | "s16" | "a16" | "r17" | "m17" | "s17 | a 1 |  |  |

\# Ordered the variables as designed
$>f t<-f t[, c(1: 10, \operatorname{seq}(11,75, b y=4), \operatorname{seq}(12,76, \operatorname{by}=4), \operatorname{seq}(13,77, \operatorname{by}=4), \operatorname{seq}(14,78, b y=4))]$
$>$ colnames (ft)
[1] "ID" "hhsz" "n1" "n2" "n3" "n4" "n5" "n6" "n7" "n9"
[11] "r1" "r2" "r3" "r4" "r5" "r6" "r7" "r8" "r9" "r10"
[21] "r11" "r 12" "r13" "r14" "r15" "r16" "r 17" "m1" "m2" "m3"

[41] "m14" "m15" "m16" "m17" "s1" "s2" "s3" "s4" "s5" "s6"
[51] "s7" "s8" "s9" "s10" "s11" "s12" "s13" "s $14 "$ "s $15 "$ "s16"
[61] "s17" "a1" "a2" "a3" "a4" "a5" "a6" "a7" "a8" "a9"
[71] "a10" "a11" "a12" "a13" "a14" "a15" "a16" "a17"
$>\operatorname{head}(f t[, c(1: 16,28: 33,45: 50,62: 67)])$
ID hhsz n1 n2 n3 n4 n5 n6 n7 n9 r1 r2 r3 r4 r5 r6 m1 m2 m3 m4 m m6 s1 s2 s3

$\begin{array}{llllllllllllllllll}2 & 101010301419 & 4 & 1\end{array}$




s4 s5 s6 a1 a2 a3 a4 a5 a6
11 NA NA 48282620 NA NA

21 NA NA 54262519 NA NA
$3 \quad 1$ NA NA 7428241 NA NA
4 NA NA NA 7774 NA NA NA NA
5 NA NA NA 483920 NA NA NA
6 NA NA NA 625921 NA NA NA

```
>ft.old3<-ft
>ft[is.na(ft)]<-0
>head (ft[, c(1:16, 28:33,45:50,62:67)])
```

ID hhsz n1 n2 n3 n4 n5 n6 n7 n9 r1 r2 r3 r4 r5 r6 m1 m2 m3 m4 m m6 s1 s2 s3

```
1101010301415 4 4 1 0 3 3 0 0 0 0 0 0 0 1 1 3 3 3 3 0 0 0 0
```



```
3 101010301424 4 1 0 0 0 0 0 0 0
4101010901913 2 1 1 1 0 0 0 0 0 0 0 0 1 2 0 0 0
5 101010901915 
6 101010901919 3 1 1 1 1 0 0 0 0 0 0 1 2 2 3 0
    s4 s5 s6 a1 a2 a3 a4 a5 a6
1 1 0 0 48 28 26 20 0 0
2 1 0 0 54 26 25 19 0 0
3 11 0 0 74 28 24 1 0 0
4 0
5}0
6 0 0 0 62 59 21 0 0 0
```


## Data file for determining the family type (2)

$\checkmark$ Designed household-level data file consist of number of persons by generation as well as number of couples by generation.

For each household;

| Generation based on the household head |  | Number of persons | Number of couples |
| :---: | :--- | :--- | :--- |
| 1 | Grandparents | p1 | c1 |
| 2 | Parents | p2 | c2 |
| 3 | Head/spouse | p3 | c3 |


| 4 | Children | p4 | c4 |
| :--- | :--- | :--- | :--- |
| 5 | Grandchildren | p5 | c5 |
| 6 | Brother/sister | p6 | c6 |
| 7 | Other | p7 | c7 |

## $\checkmark \quad$ Defined function couple( $\mathbf{d f}, \mathbf{k}, \mathbf{j}$ )

df: data frame of family type, defined in the above
k : record number of df
j : person number of df
couple=1 if $j$-th and ( $j+1$ )-th persons are a couple
couple=0 else

- The two household members within the household who satisfy the next conditions are presumed as a couple in VHLSS 2006.

1) Their person numbers (MATV) are consecutive.
2) Their marital statuses are married.
3) Their sexes are opposite.
4) Their relationship codes are the same, or the first person's code is 3 (Child) and the second person's code is 7 (Other relation).
```
> couple<-function(df, k, j) {
+ flag<-0
+ if(j==1&df[k, j+11]==2) {
+ flag<-1
+ return(flag)
+ } else if(j>=2&df[k, j+27]==2&df[k,j+28]==2&df[k,j+44]*df[k,j+45]==2&
+ (df[k, j+10]==df[k, j+11]|(df[k, j+10]==3&df[k, j+11]==7))){
+ flag<-1
+ }
+ return(flag)
+ }
```

Examples;

```
>couple(ft, 2, 2)
```

[1] 1

```
>couple(ft, 5, 1)
```

[1] 1
$>\operatorname{couple}(f t, 1,1)$
[1] 0
$>\operatorname{couple}(f t, 3,1)$
[1] 0
$>$ couple (ft, 3, 2)
[1] 1

## $\checkmark$ Defined function generation(df, $\mathbf{k}, \mathbf{j}$ )

df: data frame of family type, defined in the above
k : record number of df
j : person number of df
generation: j-th person’s generation code defined in the above table
> generation<-function(df, $\mathrm{k}, \mathrm{j}$ ) \{
$+\mathrm{g}<-7$
$+\operatorname{if}(\mathrm{df}[\mathrm{k}, \mathrm{j}+10]==1 \mid \mathrm{df}[\mathrm{k}, \mathrm{j}+10]==2) \mathrm{g}<-3$

+ else if(df[k, j+10]==3) g<-4
+ else if(df[k, j+10]==4) g<-2
+ else if (df $[k, j+10]==5) \mathrm{g}<-1$
+ else if (df[k, j+10]==6) g<-5
+ \# Child in law
+ else if (df $[k, j+10]==7 \& d f[k, j+9]==3 \& c o u p l e(d f, k, j-1)==1) g<-4$
+ \# Brother/sister
$+\mathrm{flag}<-0$
+ for ( n in ( $\mathrm{j}+1$ ): $\mathrm{df}[\mathrm{k}, 2]$ ) \{
$+\operatorname{if}(\mathrm{df}[\mathrm{k}, \mathrm{n}+10]==5) \mathrm{flag}\langle-\mathrm{flag}+1$
+ \} \# end of for
+ if (g==7\&flag>0) g<-6
+ return (g)
+ \} \# end of function


## Remarks:

Code=6 (Brother/sister) is identified only if Grandparents exist in the household.

A part of brothers/sisters might be grouped as code=7.
$\checkmark$ Created household-level data file ft2 consist of number of persons by generation as well as number of couples by generation.

```
>ft2<-data. frame(ID=ft$ID, matr ix (0, nrow=nrow (ft), ncol=14))
>colnames (ft2)<-c ("ID", paste ("p", 1:7, sep=""), paste("c", 1:7, sep=""))
>for(k in 1:nrow(ft)){
+ for(j in 1:ft[k, 2]){
+ g<-generation(ft, k, j)
+ ft2[k,g+1]<-ft2[k, g+1]+1
+ ft2[k,g+8]<-ft2[k,g+8]+couple(ft,k, j)
+ }
+ }
>head(ft2)
    ID p1 p2 p3 p4 p5 p6 p7 c1 c2 c3 c4 c5 c6 c7
```



```
2 101010301419 0
3101010301424 0}0
4101010901913 0}0
5101010901915}0000021100 0 0 0 0 0 1 1 0 0 0 0 0 
6101010901919 0
>ft2.old<-ft2
# Generated the variable of hhsz: household size
> ft2["hhsz"]<-rowSums (ft2[, 2:8])
```

\# Number of persons by generation
> sapply (ft2[, 2:8], sum)

| p1 | p2 | p3 | p4 | p5 | p6 | p7 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| 45 | 952 | 16537 | 18600 | 2203 | 4 | 730 |

```
# Generated the variable of ct: total number of married couples with household
> ft2["ct"]<-rowSums(ft2[, 9:15])
```

\# Number of married couples by generation

```
sapply (ft2[, 9:15], sum)
```

| $c 1$ | $c 2$ | $c 3$ | $c 4$ | $c 5$ | $c 6$ | $c 7$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2 | 115 | 7331 | 1109 | 3 | 0 | 40 |

\# Frequency of sample households by total number of married couples within household
> addmargins (table (ft2\$ct))

| 0 | 1 | 2 | 3 | 4 | 5 | Sum |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1384 | 7068 | 687 | 43 | 6 | 1 | 9189 |

## Grouping sample household based on the number of persons by generation

$\checkmark$ Excluded 482 households with $\mathrm{p} 7>0$, and grouped 8,707 rest of households based on the number of household members by generation.

Two-generation household is the majority, which occupies $65 \%$, followed by three-generation household (19\%) and one-generation household (10\%).

Table Type of households by number of household members' generations

| Number of generation | Household consist of the below generations | Code | Number of households | Percentage (\%) |
| :---: | :---: | :---: | :---: | :---: |
|  | All households |  | 9,189 | 100.0 |
| 1 | Total | h10 | 914 | 9.9 |
|  | Single | h11 | 344 | 3.7 |
|  | Plural members (head, spouse only) | h12 | 570 | 6.2 |
|  | Plural members including siblings | h13 | 0 | - |
| 2 | Total | h20 | 5,927 | 64.5 |
|  | Head - Children | h21 | 5,855 | 63.7 |
|  | Parents - Head | h22 | 72 | 0.8 |
| 3 | Total | h30 | 1,779 | 19.4 |
|  | Head - Children - Grandchildren | h31 | 1,074 | 11.7 |
|  | Head - Grandchildren | h32 | 96 | 1.0 |
|  | Parents - Head - Children | h33 | 606 | 6.6 |
|  | Grandparents - Parents - Head | h34 | 0 | - |
|  | Grandparents - Head | h35 | 3 | 0.0 |
| 4 | Total | h40 | 84 | 0.9 |


|  | Parents - Head - Children - Grandchildren | h41 | 46 | 0.5 |
| :---: | :--- | :--- | ---: | ---: |
|  | Parents - Head -Grandchildren | h42 | 4 | 0.0 |
|  | Grandparents - Parents - Head - Children | h43 | 3 | 0.0 |
|  | Grandparents -Head - Children | h44 | 31 | 0.3 |
|  | Grandparents <br>  <br>  <br> Grandchildren (Parents) - Head - (Children) - | h50 | 3 | 0.0 |
|  | Households with other relatives and/or non-relatives | h60 | 482 | 5.2 |

\# Generated the variable code of household type in data frame ft2.

```
> d<-subset(ft2, p7==0)
> d["code"]<-0
>d["code"]<-ifelse(d$p3==1& d$hhsz==1, 11, d$code)
>d["code"]<-ifelse(d$p1==0& d$p2==0& d$p3>=2& d$p4==0& d$p5==0& d$p6==0,12, d$code)
>d["code"]<-ifelse(d$p1==0& d$p2==0& d$p3>0& d$p4==0& d$p5==0&d$p6>0,13,d$code)
>d["code"]<-ifelse(d$p1==0& d$p2==0& d$p4>0& d$p5==0, 21,d$code)
>d["code"]<-ifelse(d$p1==0& d$p2>0& d$p4==0& d$p5==0, 22,d$code)
>d["code"]<-ifelse(d$p1==0& d$p2==0& d$p4>0& d$p5>0, 31, d$code)
>d["code"]<-ifelse(d$p1==0& d$p2==0& d$p4==0& d$p5>0,32,d$code)
> d["code"]<-ifelse(d$p1==0& d$p2>0& d$p4>0& d$p5==0,33, d$code)
>d["code"]<-ifelse(d$p1>0& d$p2>0& d$p4==0& d$p5==0, 34, d$code)
>d["code"]<-i felse(d$p1>0& d$p2==0& d$p4==0& d$p5==0,35,d$code)
>d["code"]<-ifelse(d$p1==0& d$p2>0& d$p4>0& d$p5>0, 41,d$code)
>d["code"]<-ifelse(d$p1==0& d$p2>0& d$p4==0& d$p5>0, 42, d$code)
>d["code"]<-ifelse(d$p1>0& d$p2>0& d$p4>0& d$p5==0, 43, d$code)
>d["code"]<-ifelse(d$p1>0& d$p2==0& d$p4>0& d$p5==0, 44, d$code)
>d["code"]<-ifelse(d$p1>0& d$p5>0, 50, d$code)
>(t<-table(d$code))
```

| 11 | 12 | 21 | 22 | 31 | 32 | 33 | 35 | 41 | 42 | 43 | 44 | 50 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 344 | 570 | 5855 | 72 | 1074 | 96 | 606 | 3 | 46 | 4 | 3 | 31 | 3 |

$>\operatorname{sum}(\mathrm{t})$
[1] 8707
$>$ round $(\mathrm{t} / \mathrm{nrow}(\mathrm{ft} 2) * 100,1)$
$\begin{array}{lllllllllllll}11 & 12 & 21 & 22 & 31 & 32 & 33 & 35 & 41 & 42 & 43 & 44 & 50\end{array}$
$\begin{array}{lllllllllll}3.7 & 6.2 & 63.7 & 0.8 & 11.7 & 1.0 & 6.6 & 0.0 & 0.5 & 0.0 & 0.0\end{array} 0.3 \quad 0.0$

```
# Added code to ft2
>ft2<-merge(ft2, d[c ("ID", "code")], by="ID", alI=T)
> ft2[is.na(ft2$code), "code"]<-60
> table(ft2$code, useNA="ifany")
\begin{tabular}{rrrrrrrrrrrrrr}
11 & 12 & 21 & 22 & 31 & 32 & 33 & 35 & 41 & 42 & 43 & 44 & 50 & 60 \\
344 & 570 & 5855 & 72 & 1074 & 96 & 606 & 3 & 46 & 4 & 3 & 31 & 3 & 482
\end{tabular}
> ft2.old2<-ft2
```


### 6.7.2 Family type based on the generation and number of married couples

## Grouping sample household based on the generation and number of married couples

$\checkmark$ Excluded 482 sample households with other relation and 1,262 households without married couple, then grouped 7,445 rest of households by the number of married couples' generation and number of couples.

Nuclear family household is the major type, which occupies 61\%, followed by the below types;
$>$ Household consist of head's couple and children's couple (6.5\%)
$>$ Household consist of head's couple, their parent(s) and their child(ren) (5.3\%)
$>$ Household without head's couple but with children's couple(s) (4.4\%)
$>$ Household consist of head's couple, their child(ren) and their grandchild(ren) (1.9\%)

As for number of married couples' generation, households including one-generation couple(s) is $74 \%$, and household including two-generation couples is $7.2 \%$. On the other, household without married couple is $14 \%$.

Table Type of household by number of married couples' generation and number of married couples

| Number of couples' generation | Number <br> of couples | Household consist of the below couples | Code2 | Number of households | Percentage <br> (\%) |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | All households |  | 9,189 | 100.0 |
|  |  | All households with marriage couples and without other relation |  | 7,445 | 81.0 |
| 1 generation |  | Total |  | 6,774 | 73.7 |
|  | 1 | Total |  | 6,769 | 73.7 |
|  |  | $\checkmark$ Subtotal of a head's couple | 10 | 6,350 | 69.1 |
|  |  | $\checkmark \quad$ Nuclear family (Sum of 11 and 12) |  | 5,555 | 60.5 |
|  |  | Head's couple only | 11 | 570 | 6.2 |
|  |  | Head's couple and their children | 12 | 4,985 | 54.2 |
|  |  | Head's couple, their children and their grandchildren | 13 | 177 | 1.9 |
|  |  | Head's couple and their parents | 14 | 26 | 0.3 |
|  |  | Head's couple, their parents and their children | 15 | 488 | 5.3 |
|  |  | Head's couple, their parents, their children and their grandchildren | 16 | 15 | 0.2 |


|  |  | head's couple, their grandparents and their parents | 17 | 0 | - |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | head's couple, their grandparents, their parents and their children | 18 | 1 | 0.0 |
|  |  | Head's couple and their grandchildren | 19 | 57 | 0.6 |
|  |  | Head's couple, their grandparents and their children | 20 | 24 | 0.3 |
|  |  | Head's couple, their parents and their grandchildren | 21 | 3 | 0.0 |
|  |  | Other | 29 | 4 | 0.0 |
|  |  | $\diamond$ Subtotal of no head's couple | 30 | 419 | 4.6 |
|  |  | No head's couple but children's couple | 31 | 400 | 4.4 |
|  |  | No head's couple but parents' couple | 32 | 19 | 0.2 |
|  |  | No head's couple but siblings' couple | 33 | 0 | - |
|  |  | No head's couple but grandchildren's gouple | 34 | 0 | - |
|  | 2 and more | Total | 40 | 5 | 0.1 |
|  |  | No head's couple but children's couples | 41 | 5 | 0.1 |
|  |  | Head's couple and siblings' couples | 42 | 0 | - |
|  |  | No head's couple but siblings' couple | 43 | 0 | - |
| 2 generation |  | Total | 50 | 664 | 7.2 |
|  |  | Head's couple and children's couple | 51 | 598 | 6.5 |
|  |  | Head's couple and parents' couple | 52 | 64 | 0.7 |
|  |  | Head's couple and grandchildren's couple | 53 | 0 | - |
| ' |  | Head's couple and grandparents' couple | 54 | 2 | 0.0 |
| 3 generation |  | Head's couple, parents' couple and children's couple | 60 | 7 | 0.1 |
| 0 generation |  | No married couple without other relation | 70 | 1,262 | 13.7 |
| x |  | Households with other relations | 80 | 482 | 5.2 |

$>$ table (ft2\$p7>0)
FALSE TRUE
8707482
$>$ table $(\mathrm{ft} 2 \$ \mathrm{p} 7==0 \& \mathrm{ft} 2 \$ \mathrm{ct}==0)$
FALSE TRUE
79271262
$>\mathrm{d}<-$ subset (ft2, $\mathrm{p} 7==0 \& \mathrm{ct}>0$ )
$>\operatorname{dim}(\mathrm{d})$
[1] $7445 \quad 18$

```
>d["code2"]<-0
>d["code2"]<-ifelse(d$c1==0&d$c2>0&d$c4>0&d$c5==0,60,d$code2)
>d["code2"]<-ifelse (d$c1==0&d$c2==0&d$c3>0&d$c4>0&d$c5==0, 51, d$code2)
>d["code2"]<-ifelse (d$c1==0&d$c2>0&d$c3>0&d$c4==0&d$c5==0,52, d$code2)
>d["code2"]<-ifelse (d$c1==0&d$c2==0&d$c3>0&d$c4==0&d$c5>0, 53, d$code2)
>d["code2"]<-ifelse(d$c1>0&d$c2==0&d$c3>0&d$c4==0&d$c5==0, 54, d$code2)
> d["code2"]<-ifelse(d$ct==1, 1, d$code2)
> d["code2"]<-ifelse (d$ct==1&d$c3==1, 29, d$code2)
> d["code2"]<-ifelse (d$ct==1&d$c3==0, 30, d$code2)
>d["code2"]<-ifelse (d$c1==0&d$c2==0&d$c3==0&d$c4==1&d$c5==0&d$c6==0,31, d$code2)
>d["code2"]<-ifelse(d$c1==0&d$c2==1&d$c3==0&d$c4==0&d$c5==0&d$c6==0,32, d$code2)
>d["code2"]<-ifelse(d$c1==0&d$c2==0&d$c3==0&d$c4==0&d$c5==0&d$c6==1, 33, d$code2)
>d["code2"]<-ifelse (d$c1==0&d$c2==0&d$c3==0&d$c4==0&d$c5==1&d$c6==0,34, d$code2)
> d["code2"]<-ifelse (d$ct>=2&d$c1==0&d$c2==0&d$c3==0&d$c4>=2&d$c5==0&d$c6==0, 41, d$code2)
> d["code2"]<-ifelse (d$ct>=2&d$c1==0&d$c2==0&d$c3==1&d$c4>=2&d$c5==0&d$c6>=1, 42, d$code2)
> d["code2"]<-ifelse (d$ct>=2&d$c1==0&d$c2==0&d$c3==0&d$c4>=2&d$c5==0&d$c6>=2, 43, d$code2)
>d["code2"]<-ifelse(d$ct==1&d$c3==1&d$hhsz==2, 11, d$code2)
>d["code2"]<-ifelse (d$ct==1&d$c3==1&d$p1==0&d$p2==0&d$p4>0&d$p5==0&d$p6==0, 12, d$code2)
> d["code2"]<-ifelse(d$ct==1&d$c3==1&d$p1==0&d$p2==0&d$p4>0&d$p5>0&d$p6==0, 13, d$code2)
>d["code2"]<-ifelse(d$ct==1&d$c3==1&d$p1==0&d$p2>0&d$p4==0&d$p5==0&d$p6==0,14,d$code2)
>d["code2"]<-ifelse (d$ct==1&d$c3==1&d$p1==0&d$p2>0&d$p4>0&d$p5==0&d$p6==0, 15, d$code2)
>d["code2"]<-ifelse(d$ct==1&d$c3==1&d$p1==0&d$p2>0&d$p4>0&d$p5>0&d$p6==0, 16, d$code2)
> d["code2"]<-ifelse (d$ct==1&d$c3==1&d$p1>0&d$p2>0&d$p4==0&d$p5==0&d$p6==0, 17, d$code2)
>d["code2"]<-ife|se (d$ct==1&d$c3==1&d$p1>0&d$p2>0&d$p4>0&d$p5==0&d$p6==0, 18, d$code2)
>d["code2"]<-ifelse(d$ct==1&d$c3==1&d$p1==0&d$p2==0&d$p4==0&d$p5>0&d$p6==0, 19,d$code2)
> d["code2"]<-ifelse(d$ct==1&d$c3==1&d$p1>0&d$p2==0&d$p4>0&d$p5==0&d$p6==0, 20, d$code2)
> d["code2"]<-ifelse (d$ct==1&d$c3==1&d$p1==0&d$p2>0&d$p4==0&d$p5>0&d$p6==0, 21, d$code2)
table(d$code2)
\begin{tabular}{rrrrrrrrrrrrrrr}
11 & 12 & 13 & 14 & 15 & 16 & 18 & 19 & 20 & 21 & 29 & 31 & 32 & 41 & 51 \\
570 & 4985 & 177 & 26 & 488 & 15 & 1 & 57 & 24 & 3 & 4 & 400 & 19 & 5 & 598
\end{tabular}
    52 54 60
    64 2 7
round(table(d$code2)/nrow (ft2)*100, 1)
\begin{tabular}{lllllllllllllll}
11 & 12 & 13 & 14 & 15 & 16 & 18 & 19 & 20 & 21 & 29 & 31 & 32 & 41 & 51
\end{tabular}
6.2 54.2 1.9 0.3 5.3 0.2 0.0 0.6 0.3 0.0
52 54 60
```

```
0.7 0.0 0.1
# Added code2 to ft2
> ft2<-merge(ft2, d[c ("ID", "code2")],by="ID", all=T)
> ft2["code2"]<-ifelse(ft2$p7>0, 80, ft2$code2)
>ft2["code2"]<-ifelse(ft2$p7==0&ft2$ct==0,70, ft2$code2)
> table(ft2$code2, useNA="'ifany")
\begin{tabular}{rrrrrrrrrrrrrrr}
11 & 12 & 13 & 14 & 15 & 16 & 18 & 19 & 20 & 21 & 29 & 31 & 32 & 41 & 51 \\
570 & 4985 & 177 & 26 & 488 & 15 & 1 & 57 & 24 & 3 & 4 & 400 & 19 & 5 & 598
\end{tabular}
    52 54 60 70 80
    64 2 7 1262 482
ft2. old3<-ft2
```


## Family type by region

Comparing family type between Red River Delta and Mekong River Delta, the share of nuclear family is higher in Red River Delta. Consequently, the average household size in Red River Delta is smaller than Mekong River Delta.

```
\checkmark ~ C o m p o s i t i o n ~ o f ~ t y p e ~ o f ~ h o u s e h o l d s ~ b y ~ n u m b e r ~ o f ~ h o u s e h o l d ~ m e m b e r s ' ~ g e n e r a t i o n ~ a n d ~ b y ~ r e g i o n
> family<-cbind(ft2, ttchung[, c("tinh", "ttnt", "dantoc", "tsnguoi", "thunhap", "chids", "wt")])
> t<-tapply(family$wt, list(substr (family$tinh, 1, 1), family$code), sum)
t[is.na(t)]<-0
> region. name<-c ("Red River Delta", "North East", "North West", "North Central",
+ "South Central Coast","Central Highlands", "South East","Mekong River Delta")
> rownames (t)<-region. name
round (prop. table (t, 1)*100, 1)
    11
Red River Delta 4.6 10.3 62.0 0.7 10.6 1.3 6.9 0.0 0.5 0.1 0.0 0.1 0.0 2.9
North East }\quad1.9 5.1 64.8 0.9 10.0 0.6 9.3 0.0 0.5 0.0 0.1 0.6 0.1 6.1
North West 1.6 3.6 66.6 0.6 10.8 0.4 5.5 0.0 0.5 0.0 0.0 0.0 0.0 10.4
North Central 4.1 7.3 65.9 0.5 10.4 0.4 7.6 0.1 0.7 0.0 0.0 0.4 0.0 2.6
South Central Coast 3.5 6.4 65.2 0.6 10.5 0.7 8.4 0.1 0.4 0.0 0.0 0.3 0.0 3.9
Central Highlands 1.9 4.1 72.0 0.2 7.8 0.7 5.9 0.0 0.4 0.0 0.0 1.0 0.0 5.8
South East }\quad5.5 3.5 61.8 0.7 13.6 1.0 5.0 0.0 0.6 0.0 0.0 0.3 0.0 8.0
Mekong River Delta 4.0 5.2 59.7 1.3 16.5 1.8 4.4 0.0 0.3 0.1 0.0 0.1 0.0 6.4
```

$\checkmark$ Composition of type of households by number of married couples' generation and number of married couples, and by region

```
> t<-tapply (family$wt, list(substr (family$tinh, 1, 1), family$code2), sum)
l[is.na(t)]<-0
round(prop. table (t, 1)*100, 1)
    11
110.3 52.9 1.8 0.4 5.6 0.2 0 0.9 0.1 0.1 0.0 4.1 0.0 0.0 5.4 0.7 0.0 0.1 14.6 2.9
2 5.1 55.5 1.5 0.7 7.5 0.1 0 0.5 0.4 0.0 0.1 3.3 0.1 0.0 7.7 1.1 0.0 0.1 10.2 6.1
3 3.6 57.7 1. 8 0.6 3.9 0.0 0 0.4 0.0 0.0 0.0 1.8 0.0 0.0 10.4 1.0 0.0 0.2 8.2 10.4
4 7.3 56.7 2.2 0.3 6.7 0.3 0 0.2 0.4 0.0 0.1 3.7 0.2 0.1 5.1 0.4 0.0 0.0 13.9 2.6
5 6.4 56.4 1.6 0.0 6.9 0.4 0 0.2 0.2 0.0 0.1 4.1 0.5 0.1 5.4 0.6 0.0 0.0 13.3 3.9
```

$64.162 .81 .90 .25 .10 .2 \quad 0 \quad 0.50 .70 .00 .12 .10 .00 .0$


$\checkmark \quad$ Average household size by region
>hhsz<-tapply (family\$tsnguoi*family\$wt, substr (family\$tinh, 1, 1), sum)/

+ tapply (family\$wt, substr (family\$tinh, 1, 1), sum)
> names (hhsz) <-region. name
$>$ round (hhsz, 2)

| Red River Delta | North East | North West | North Central |
| ---: | ---: | ---: | ---: |
| 3.83 | 4.27 | 4.81 | 4.27 |
| Senth Central Coast | Central Highlands | South East | Mekong River Delta |
| 4.27 | 4.88 | 4.31 | 4.19 |

### 6.8 Education

- Compared school attendance rate (SRA) between male and female.
- Data frames and variables used;
> muc1a 39,071 records

```
    \diamond MIAC2 sex
```

    \(\diamond\) M1AC5 age
    \(\diamond \mathrm{wt}\)
    \(\rangle\) PID
    > muc2a 39,071 records
    M M2AC5 Currently attending school?
            1 Yes
                2 Vacation
            3 No
    « wt
    \(\star\) PID
    >d<-merge (muc1a[, c ("PID", "m1ac2", "m1ac5")],muc2a[, c ("PID", "m2ac5", "wt")])
$>\mathrm{d}<$-subset (d, mlac5>=5\&m1ac5<=24)
$>\operatorname{dim}(\mathrm{d})$
[1] 158355
$>\mathrm{m}<-$ matrix (as. vector (by (d\$wt, list (d\$m1ac5, d\$m2ac5, d\$m1ac2), sum) ), nrow=20, ncol=6)
$>$ rownames $(m)<-5: 24$
> colnames (m) <-outer (c ("Yes", "Vacation", "No") , c ("M", "F"), paste, sep="-")
$>$ round (m)
Yes-M Vacation-M No-M Yes-F Vacation-F No-F
$51256176 \quad 126693196230227789 \quad 98104183635$
$6 \quad 280540 \quad 206955102211298711 \quad 17222560424$
$\begin{array}{llllllll}7 & 306948 & 279764 & 17811 & 307781 & 180568 & 8203\end{array}$
$8 \quad 300101 \quad 283586 \quad 14080306068 \quad 262342 \quad 21313$
$\begin{array}{lllllllll}9 & 357408 & 332475 & 18410 & 358804 & 304737 & 17060\end{array}$
$10452439394041 \quad 20013400681 \quad 365906 \quad 9140$
$11382831 \quad 354612 \quad 37193475993 \quad 386252 \quad 28489$
$12420381 \quad 36745240795483817 \quad 40604252599$
$13446986 \quad 46090394640498525 \quad 46253686768$
$14457009 \quad 387374159974438114 \quad 418141141029$
$15455399 \quad 413641258209365051 \quad 384540202645$
$16445887 \quad 335297376000412619 \quad 345752329362$
$17335312 \quad 269601483310348163 \quad 272138360918$
$18271186 \quad 155011582531207864 \quad 139435492442$
$19270177 \quad 46581645679277946 \quad 40373504163$
$2022657939338576048215340 \quad 33407535029$

|  | 185678 | 22555641390189922 | 22000575593 |
| :---: | :---: | :---: | :---: |
|  | 154610 | 11020638537132590 | 9215572484 |
| 23 | 142607 | 1778261748655919 | 4953679878 |
| 24 | 55446 | 160464797148171 | 10649603295 |
| ```# School attendance rate by sex and single year age >df<-as.data. frame (m)``` |  |  |  |
| $>\mathrm{df}[$ "male. SAR"]<-(df[, 1]+df[, 2])/rowSums (df[, 1:3])*100 |  |  |  |
| $>d f[" f e m a l e . S A R "]<-(d f[, 4]+d f[, 5]) / r o w S u m s ~(d f[, 4: 6]) * 100$ |  |  |  |
| $>(\mathrm{sar}<-\mathrm{round}(\mathrm{df}[, 7: 8], 1))$ |  |  |  |
| male. SAR female. SAR |  |  |  |
| 5 | 66.1 | 64.0 |  |
| 6 | 82.7 | 88.6 |  |
| 7 | 97.1 | 98.3 |  |
| 8 | 97.6 | 96.4 |  |
| 9 | 97.4 | 97.5 |  |
| 10 | 97.7 | 98.8 |  |
| 11 | 95.2 | 96.8 |  |
| 12 | 95.1 | 94.4 |  |
| 13 | 90.6 | 91.7 |  |
| 14 | 84.1 | 85.9 |  |
| 15 | 77.1 | 78.7 |  |
| 16 | 67.5 | 69.7 |  |
| 17 | 55.6 | 63.2 |  |
| 18 | 42.3 | 41.4 |  |
| 19 | 32.9 | 38.7 |  |
| 20 | 31.6 | 31.7 |  |
| 21 | 24.5 | 26.9 |  |
| 22 | 20.6 | 19.9 |  |
| 23 | 20.6 | 8. 2 |  |
| 24 | 8.1 | 8. 9 |  |

## $\checkmark$ Summary:

Female's school attendance rate (SAR) is slightly higher than male's SAR.


## 7. Agricultural land

### 7.1 History

- In Vietnam, land belongs to the entire people, which was approved by the National Assembly in 1980. The State has the powers and responsibilities for uniform administration of land as representative.
- In 1986, Doi Moi (Renovation) policy was launched.
- 1993 Land Law allowed ownership of a right to use land. This right is called the Land Use Right (LUR).
- In 1993, agricultural land was allocated free of charge for farmers for 20 years.
- 2003 Amended Land Law introduced market mechanism for LUR. Market price was applied for the price of LUR.


### 7.2 Rice cultivated area by region

Comparison of the rice cultivated area by region shows that mean rice cultivated area per sample household conditional on non-zero rice production is smaller in Red River Delta as compared to Mekong River Delta. The coefficient of variation is also smaller in Red River Delta.

It is said that in North Vietnam agricultural land was equally distributed among farmers after the World War II, which resulted in the small mean area and the coefficient of variation. On the other, in South Vietnam landlord system had continued until the unity in 1976.

Table Share of sample households cultivated rice, mean and coefficient of variation of rice cultivated area per sample household (VHLSS 2006)

| Region | Share of households <br> cultivated rice (\%) | Mean rice cultivated area <br> for the past 12 months $\left(\mathrm{m}^{2}\right)$ | Coefficient of <br> variation |
| :--- | ---: | ---: | ---: |
| Red River Delta | 67.6 | 4,019 | 0.63 |
| Mekong River Delta | 37.2 | 23,141 | 1.12 |

Note: For example, in a period of one year, on a land lot of $360 \mathrm{~m}^{2}$ people grow two rice crops, total rice cultivated area is twice the arable area, that is $720 \mathrm{~m}^{2}$.

```
# Used data frame MUC4B11
>d<-subset (muc4b11, m4b11ma>=5)
> dim(d)
[1] 6477 22
# Aggregated the rice cultivated area for the past 12 months by household
> t<-tapply (d$m4b11c3, d$ID, sum)
# Created data frame land consist of variables of ID, area, wt for all households.
> land<-data.frame(ID=names (t), area=t, row. names=NULL)
> land<-merge(land, ttchung[c ("ID", "wt")], alI=T)
> land[is. na(land)]<-0
# Mean rice cultivated area by region
tapply(land$area, substr (land$ID, 1, 1), function(x) mean(x))
\begin{tabular}{cccccccr}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
2718. 764 & 2871.756 & 5305.886 & 3175.130 & 2344.995 & 2342.837 & 1607.811 & 8620.547
\end{tabular}
# Coefficient of variation
>tapply(land$area, substr(land$ID, 1, 1), function(x) sqrt(var (x))/mean(x))
    1
1.029775 1.047827 1. 154190 1. 316736 1.316790 2. }056574 3.938203 2.248628
# Generated data frame land2, a subset of land, where household cultivated rice.
> land2<-subset(land, area>0)
> dim(land2)
[1] 4824 3
# Share of sample households cultivated rice
> tapply(land2$wt, substr (land2$ID, 1, 1), length)/tapply (land$wt, substr (land$ID, 1, 1), length)
\begin{tabular}{llllllll}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8
\end{tabular}
0.6764403 0. 7031131 0.7715618 0.6607495 0.5516432 0.4106529 0.1506734 0.3725174
# Mean rice cultivated area per sample household
> tapply(land2$area, substr (land2$ID, 1, 1), function(x) mean(x))
\begin{tabular}{rrrrrrrr}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
4019.223 & 4084.343 & 6876.813 & 4805.346 & 4250.928 & 5705.151 & 10670.838 & 23141.324
\end{tabular}
# Coefficient of variation
> tapply(land2$area, substr(land2$ID, 1, 1), function(x) sqrt(var(x))/mean(x))
    1 
0.6274495 0.6892151 0.8941031 0.8980131 0.7128223 1.0720696 1.2222574 1.1211101
```


[^0]:    muc3h.dta : 921431
    muc3i.dta : 2162216
    muc4a.dta: 3907152
    muc4b0.dta : 2408016
    muc4b11.dta : 1496119
    muc4b12.dta : 1549912
    muc4b13.dta : 393213
    muc4b14.dta: 979113
    muc4b15.dta: 837910
    muc4b16.dta : 5251311
    muc4b161.dta : 1659211
    muc4b21.dta : 1657714
    muc4b22.dta: 1022218
    muc4b31.dta: 2799
    muc4b32.dta : 27717
    muc4b41.dta: 393213
    muc4b42.dta : 228320
    muc4b51.dta : 365314
    muc4b52.dta : 270619
    muc4c.dta: 437933
    muc4c2.dta: 293339
    muc4d.dta: 918922
    muc5a1.dta : 13172610
    muc5a2.dta: 28856415
    muc5b1.dta : 11411410
    muc5b2.dta: 1149528
    muc5b3_4.dta : 918923
    muc6.dta : 918967
    muc6a.dta : 1589712
    muc6b.dta : 8589011
    muc7.dta : 918951
    muc8.dta : 918937
    muc8_vayvon.dta : 498720
    ttchung.dta : 9189141

