Viet Nam Household Living Standards Survey
Codebook

Development and Policies Research Center (DEPOCEN)

2026-01-10

Contents

Preface

The Viet Nam Household Living Standards Survey (VHLSS) is one of the most comprehensive and widely used micro-datasets for socio-economic research in Viet Nam. However, as the survey has evolved significantly since its inception in 1993 through changes in sampling design, questionnaire structure, and variable naming, longitudinal analysis requires rigorous data cleaning and harmonization.

Through our work with this dataset, we noticed that while many researchers have cleaned and used the VHLSS, most work in isolation. This leads to inconsistent cleaning procedures and a narrow focus on specific modules. We encountered this same fragmentation even within our own team, and it is a challenge shared by the broader research community. Thus, this documentation was developed by the Development and Policies Research Center (DEPOCEN) to serve as a standardized guide for researchers and data analysts. Our goal is to ensure that the data cleaning process is:

- Reproducible: All steps are scripted to ensure consistency.
- Transparent: Every decision regarding outliers, missing values, and variable merging is documented.
- Accessible: By centralizing metadata and harmonization crosswalks, we aim to reduce the "entry barrier" for new researchers using VHLSS.

Who are we?

This codebook was developed by a research team at DEPOCEN, comprising **Tuong-Vy Phan, Huy Le Vu, and Nguyen Thi Hong Tram**, under the supervision of **Dr. Anh Ngoc Nguyen**. We would like to express their gratitude to **Dr. Doan Quang Hung** for providing the core variable list and to **Dr. Vu Hoang Linh** for sharing the data-cleaning scripts for the consumption module.

We are grateful for the comments and support of other DEPOCEN researchers, including Manh-Duc Doan, Quang-Thanh Tran, Nguyen Viet Lien, and Ha-My Bui.

Who is this for?

This guide is intended for economists, policy analysts, and students who are working with VHLSS microdata. We assume a basic understanding of statistical software (specifically Stata and R) and familiarity with household survey structures.

Structure of the Documentation

- Data Structure: An overview of the VHLSS history and sample design.
- Cleaning Procedure: A step-by-step workflow for moving from raw files to a master harmonized dataset, with provided Stata code.
- Notes & References: A collection of insights from the wider economic research community.

Acknowledgements

We would like to thank the General Statistics Office (GSO) for their work in conducting these surveys and the various scholars whose previous notes on VHLSS provided the foundation for this consolidated procedure.

Chapter 1

VHLSS Introduction

Purpose

To evaluate living standards for policy-making and socio-economic development planning, from 1993 to now the General Statistics Office (GSO) conducts the Viet Nam Household Living Standards Survey (VHLSS). The purpose of the VHLSS in order to systematically monitor and supervise the living standards of different population groups in Viet Nam; to monitor and evaluate the implementation of the Comprehensive Poverty Reduction and Growth Strategy; and to contribute to the evaluation of achievement of the Sustainable Development Goals (SDGs) and Vietnam's socio-economic development goals.

Availability

From 2002 to 2010, this survey has been conducted regularly by the GSO every two years. From 2011 to 2022, VHLSS are conducted annually. However, the odd-numbered year surveys only collect data on demographics, employment and income.

Survey period

- The survey was conducted in four periods in March, June, September and December. The period for collecting information in the locality is one month.
- The reference period of household income and expenditure was the last 12 months.

Coverage of the survey

Geographically, the survey covered the whole country. Scope of the survey included all selected enumeration areas and communes in 63 provinces and cities under central management.

Data collection method

Face-to-face interviews

Chapter 2

Dataset Basic Information

2.1 Outline of the survey

- Section 1. Basic demographic characteristics related to living standards
- Section 2. Education
- Section 3. Labor - Employment
- Section 4. Health and health care
- Section 5. Income
- Section 6. Consumption expenditure
- Section 7. Durable goods
- Section 8. Housing, electricity, water, sanitation facilities and use of Internet
- Section 9. Participation in poverty reduction programs
- Section 10. Business production activities
- Section 11. Commune general characteristics

2.2 Sample Size

Sample Overview

The sample for the the year t Residential Living Standards Survey (KSMS t) was designed in 2 steps as follows:

Step 1: Select Survey Areas

Select 3,133 areas from the master sample, structured as follows:

- 25% (783 areas): Re-selected from areas surveyed only in KSMS $t-2$
- 25% (783 areas): Re-selected from areas surveyed in both KSMS $t-2$ and KSMS $t-1$

Table 2.1: Sample Size by Survey Year (households)

Year	Total	Income	Expenditure
2008	45.945		9.189
2010	69.360		9.399
2012			9.399
2014	46.995	37.596	9.399
2016	46.995	37.596	9.399
2018	46.995	37.596	9.399
2020	46.980		

Notes: Total: Total surveyed households; Income: Households were asked about income and other issu

- 25% (783 areas): Re-selected from areas surveyed only in KSMS $t-1$
- 25% (783 areas): Newly selected from the master sample

Step 2: Select Households for Survey

For areas re-selected from KSMS $t-1/t-2$:

- Select all 15 households previously surveyed in $t-2$ and/or $t-1$
- If a household is no longer in the area, a replacement household will be chosen
- Additionally select 5 reserve households from the reserve lists of previous years (if insufficient, select adjacent households)

For newly selected areas:

- Update the list of all households in the area
- Select 20 households using the systematic random method from the updated list
- From these, select 15 main households and 5 reserve households

2.3 Rotating panel

The VHLSS uses a rotating panel design, allowing for the construction of a panel across three survey rounds (e.g., 2014-2016-2018), provided the data is based on the same Population and Housing Census sampling frame. For instance, the household systems from 2010 to 2018 were designed using the 2009 Census. However, since the 2020 VHLSS was based on the 2019 Census, it is not possible to create a panel linking VHLSS 2016-2018 with 2020.

2.4 Weight Data

At household level, each VHLSS will have a weihgt data for each year. Usually we would use $wt9$, which is the weight of 9,000 household. There are also $wt36$ and $wt45$ is for file with 36,000 or 45,000 household.

At individual level, we need to take the weight divided by the number of family member. Weight individual $= \frac{\text{Weight household}}{\text{Houshold size}}$

Chapter 3

Data Processing Procedure

Analyzing the VHLSS longitudinally presents challenges due to structural design changes, administrative adjustments, and varying module contents. This section outlines these inconsistencies and our strategies for data harmonization. For this version, we utilize the 9,000-household sample to construct a pooled individual-level dataset, focusing on a specific subset of variables.

The following sections provide a step-by-step guide to navigating this documentation.

3.1 Folder organization

The folder contains data, do file and other materials are organized as follows:

Sub-folder	Description	Action
`01_dofiles`	Code in Stata to clean the datasets	Run master.do only to modify cleaning
`02_data`	Raw dataset in Stata (.dta) format	Do not modify
`03_temp`	Temporary files	Do not modify
`04_clean`	Cleaned data in Stata (.dta) format	Download and use
`05_metadata`	Questionnaire and codebook	Access questionnaire and codebook here

After organizing folder, first thing we need to do is clear all settings and set working directory.

3.2 Clear all settings

```
cap log close
clear all
clear matrix
set more off
eststo clear
```

3.3 Set Working Directory

Replace username and path to your folder.

```
if "`c(username)'" == "XXX" {
gl MyProject "/Users/XXX/My Drive/DEPOCEN - VHLSS Data cleaning"
    gl data "$MyProject/02_data"
    gl temp  "$MyProject/03_temp"
    gl clean "$MyProject/04_clean"
    gl metadata "$MyProject/05_metadata"
}
```

3.4 Clean data for each year

Make sure each file is uniquely defined

As each survey year consists of multiple section-specific files, we must ensure all observations are uniquely identified prior to merging.

```
forval i = 14 (2) 18 {
    foreach m in Muc1A Muc2A Muc2X Muc3A Muc3C Muc4a {
        local filepath "$data/vhlss_20`i'/hh_9000/`m'.dta"
        capture confirm file "`filepath'"
        if _rc == 0 {
            use "`filepath'", clear
            if _N > 0 {
                duplicates drop tinh huyen xa diaban hoso matv, force
                tempfile uniq_`m'_20`i'
                save `uniq_`m'_20`i'', replace
            }
        }
    }
}
```

Merge individual file

```
forval i = 14 (2) 18 {
    use `uniq_Muc1A_20`i'', clear
        foreach m in Muc2A Muc2X Muc3A Muc3C Muc4a {
            capture merge 1:1 tinh huyen xa diaban hoso matv using `uniq_`m'_20`i''
            capture drop if _m == 2
            capture drop _m
        }
    save "$temp/individual_20`i'.dta", replace
}
```

Extract variables from household file

These variables do not appear in individual file, so we need to take them from household file.

```
forval i = 14 (2) 18 {
    use "$data/vhlss_20`i'/hh_9000/Ho1.dta", clear
    keep tinh huyen xa diaban hoso ttnt dantoc tsnguoi // take area, ethnic and household size va
    duplicates drop tinh huyen xa diaban hoso, force

    tempfile indiv1_`i'
    save `indiv1_`i''

    use "$data/vhlss_20`i'/hh_9000/Muc5a1.dta", clear
    keep tinh huyen xa diaban hoso m5a1ma m5a1c2a m5a1c2b m5a1c3a m5a1c3b // take food consumptio
    duplicates drop tinh huyen xa diaban hoso, force

    tempfile indiv2_`i'
    save `indiv2_`i''

    use "$data/vhlss_20`i'/hh_9000/Muc5a2.dta", clear
    keep tinh huyen xa diaban hoso m5a2ma m5a2c2a m5a2c2b // take food consumption variabls
    duplicates drop tinh huyen xa diaban hoso, force
    destring m5a2ma, replace

    tempfile indiv3_`i'
    save `indiv3_`i''

    use "$temp/individual_20`i'.dta", clear
        forval j = 1/3 {
            merge m:1 tinh huyen xa diaban hoso using `indiv`j'_`i''
```

```
            drop if _m == 2
            drop _m
        }
    save "$temp/individual_20`i'.dta", replace
}
```

Label define all variables

Due to the legacy GSO encoding used in the raw data, we converted the character sets into a readable format. Specifically, we extracted the correct value labels for all variables of interest into a separate do-file, which is then executed within the master script.

```
forval i = 14 (2) 18 {
    use "$temp/individual_20`i'.dta", clear

    // Run the fixed labels
    do "$temp/vhlss_labels_fixed.do"

    // Automatically attach labels to variables with the same name
    foreach v of varlist _all {
        capture label values `v' `v'
    }

    save "$temp/individual_20`i'.dta", replace
}
```

3.5 Append individual dataset

Import the meta data

Variable names and the ordering of questions frequently change between survey waves. For instance, the variable for `education` may be coded as *m2ac2a* in one year and *m2xc2a* in another. To facilitate analysis, we have compiled comprehensive metadata mapping variable names and labels for each year, which is available in the `05_metadata/VHLSS_codebook_9k_thanhvien - thanhvien.csv`. This dataset is restricted to variables of interest. To add further variables, download the metadata and ensure all new variables are harmonized to account for naming inconsistencies across years.

```
import delimited "$metadata/VHLSS_codebook_9k_thanhvien - thanhvien.csv", varnames(1) e
```

Rename and label code

```
local N = _N

forval i = 2014 (2) 2018 {
    // Open file
    file open myfile using `"$temp/label_in`i'.do"', write text replace

    local renamed_vars ""
    forval iii = 1/`N' {
        local vvvcode = code[`iii']
        local vvv_i = code_`i'[`iii']
        local v_desc = description[`iii']

        // Check if code_`i' is not empty
        if `"`vvv_i'"' != "" {
                // 1. Create the Rename command
            local result = `"ren `vvv_i' `vvvcode'"'
            file write myfile `"`result'"' _n

                // 2. Create the Label command
            local lab_cmd `"label variable `vvvcode' `"`v_desc'"'"'
            file write myfile `"`lab_cmd'"' _n

            local renamed_vars "`renamed_vars' `vvvcode'"
        }
    }
    file write myfile `"keep`renamed_vars'"'

    // close file
    file close myfile
}
```

```
clear all

tempfile combined_data
save `combined_data', emptyok

forval i = 2014 (2) 2018 {
    // Use file corresponding to each year
    use "$temp/individual_`i'.dta", clear

    // Run file .do corresponding to each year
    do `"$temp/label_in`i'.do"'
```

```
    gen year = `i'

    // Append data to initially created file
    append using `combined_data'

    save `combined_data', replace
}

save "$temp/vhlss_individual_14_18.dta", replace
```

3.6 Inconsistent province codes

As many users of Vietnamese data know, the number of provinces has changed significantly since the late 1980s. In most cases the changing of provincial boundaries was either a splitting or aggregating of existing provinces as opposed to districts being reallocated between provinces. The province codes also change within surveys and across data sources.

All province codes in this processed dataset are consistent across years. The implementation of this standardization can be found in Brian McCaig's website.

3.7 Generate new variables

Below are variables of interest that we create for our research purpose

Education

```
gen education = 0 if general_edu == 0 & vocational_edu == 0
replace education = 1 if general_edu == 1
replace education = 2 if general_edu == 2
replace education = 3 if general_edu == 3
replace education = 4 if vocational_edu == 4
replace education = 5 if vocational_edu == 5
replace education = 6 if vocational_edu == 6
replace education = 7 if vocational_edu == 7
replace education = 8 if general_edu == 8
replace education = 9 if general_edu == 9
replace education = 10 if general_edu == 10
replace education = 11 if general_edu == 11
replace education = 12 if general_edu == 12
```

```
lab var education "Bằng cấp cao nhất"
lab def education_lbl 0 "Không có bằng cấp" ///
                      1 "Tiểu học" ///
                      2 "THCS" ///
                      3 "THPT" ///
                      4 "Sơ cấp nghề" ///
                      5 "Trung cấp nghề" ///
                      6 "Trung học chuyên nghiệp" ///
                      7 "Cao đẳng nghề" ///
                      8 "Cao đẳng" ///
                      9 "Đại học" ///
                      10 "Thạc sĩ" ///
                      11 "Tiến sĩ" ///
                      12 "Khác"
lab val education education_lbl
```

Total income

```
foreach m in wage_1 holiday_bonus_1 bonus_1 wage_2 holiday_bonus_2 bonus_2 {
    replace `m' = 0 if missing(`m')
}

egen income = rowtotal(wage_1 holiday_bonus_1 bonus_1 wage_2 holiday_bonus_2 bonus_2)
replace income = . if income == 0

lab var income "Tổng thu nhập từ tất cả các nguồn (tiền lương, tiền thưởng, phụ cấp, etc.) trong
```

Food consumption

```
    // holiday consumption
egen quant_h = rowtotal(food_cons1_q food_cons2_q)
egen values_h = rowtotal(food_cons1_p food_cons2_p)

    // daily consumption
ren (food_cons3_q food_cons3_p) (quant_d values_d)

    // total consumption
foreach m in quant_d quant_h values_d values_h {
    replace `m' = 0 if `m' == .
}
```

```
gen food_quant = quant_d*350/30+quant_h
gen food_values = values_d*350/30+values_h

lab var food_quant "Số lượng thực phẩm tiêu thụ (kg)"
lab var food_values "Trị giá thực phẩm tiêu thụ (nghìn đồng)"

drop quant_* values_*

save "$clean/vhlss_individual_14_18.dta", replace
```

Chapter 4

Summary Statistics (of processed data)

Library

```
library(tidyverse)
library(gt)
library(gtExtras)
library(summarytools)
library(haven)
library(sjlabelled)
library(webshot2)
```

Import data

```
vhlss <- read_dta("clean/vhlss_individual_14_18.dta")

var_to_drop <- c("tinh", "huyen", "xa", "diaban", "hoso", "gioitinh", "dantoc")
```

Summary statistics

```
source("script/gt_summarytools.R")

vhlss <- vhlss %>%
  head(100) %>%
  select(where(is.numeric)) %>%
  select(-any_of(var_to_drop)) %>%
  mutate(across(everything(), ~ ifelse(.x < 0, NA, .x))) %>%
  copy_labels(vhlss)
```

```
gt_summarytools(data = vhlss, title = "VHLSS 2014-2018 Data Summary")
```

Chapter 5

Reference

VHLSS is a popular dataset and has been used by many scholars. Here are some of the notes from other economists that we should read:

- Notes on Vietnamese data by Brain McCaig
- Data and Rscript sharing by Trinh Thi Huong
- Notes on Vietnamese Microdata by Lan Anh Ngo
- Gaps in Household Income From VHLSS 2018 by Nguyen Chi Dung