VHLSS Cleaning Procedure

Tuong-Vy Phan, Nguyen Thi Hong Tram & Huy Le Vu

2025-12-23

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, because the survey has evolved significantly since its inception in 1993—with changes in sampling design, questionnaire structure, and variable naming—longitudinal analysis requires rigorous data cleaning and harmonization.

The Purpose of this Book

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 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

Overview

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

Data Structure

2.1 Outline of the survey

- Section 1. Basic demographic characteristics related to living standards
- Section 2. Education
- Section 3. Labour - 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

- Sample size: 46,995 households
- Number of survey areas: 3,133 areas (selected from the master sample of the 2019 Population and Housing Census, updated as needed)

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 $t-1$
- 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 Basic Statistics

Library

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

Import data

```
vhlss <- read_dta("clean/vhlss_14_18.dta")

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

Summary statistics

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

vhlss <- vhlss %>%
  head(9399) %>%
  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 3

Cleaning Procedure

Below is the code used to clean VHLSS 2014 to 2018 at household level

3.1 Clear all settings

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

3.2 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 "$MyProject/data"
    gl clean "$MyProject/clean"
    gl temp  "$MyProject/temp"
}
```

3.3 Clean data for each year

3.3.1 Make sure each file is uniquely defined

```
forval i = 14 (2) 18 {
    foreach m in Ho1 Ho2 Ho3 Ho4 {
        use "$data/VHLSS_20`i'/hhold/`m'.dta", clear
        duplicates drop tinh huyen xa diaban hoso, force
        tempfile uniq_`m'_20`i'
        save `uniq_`m'_20`i'', replace
    }
}
```

3.3.2 Merge household file

```
forval i = 14 (2) 18 {
    use `uniq_Ho1_20`i'', clear
    merge 1:1 tinh huyen xa diaban hoso using `uniq_Ho2_20`i''
    drop _merge

    merge 1:1 tinh huyen xa diaban hoso using `uniq_Ho3_20`i''
    drop _merge

    merge 1:1 tinh huyen xa diaban hoso using `uniq_Ho4_20`i''
    drop _merge

    save "$temp/ho_20`i'.dta", replace
}
```

3.3.3 Generate variables

```
forval i = 14 (2) 18 {
    use "$data/VHLSS_20`i'/hhold/Muc1A.dta", clear
    keep if m1ac3 == 1 // keep if that individual is household head
    keep tinh huyen xa diaban hoso m1ac2 m1ac5
    duplicates drop tinh huyen xa diaban hoso, force
        // Gender of household head
    merge 1:1 tinh huyen xa diaban hoso using "$temp/ho_20`i'.dta"
    keep if _m == 3
    drop _m
```

```
        // Total expenditure
    egen tongchi_01 = rowtotal(chisxkd_1 chisxkd_2 chisxkd_3 chisxkd_4 chisxkd_5 chisxkd_6 chisxk
    save "$temp/ho_20`i'.dta", replace
}
```

3.3.4 Take variables from individual file

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

```
use "$data/VHLSS_2014/hhold/Muc3C.dta", clear
  collapse (sum) m3c11 m3c13 m3c14 m3c15, by(tinh huyen xa diaban hoso)
  merge 1:1 tinh huyen xa diaban hoso using "$temp/ho_2014.dta"
  drop _m
save "$temp/ho_2014.dta", replace

use "$data/VHLSS_2014/hhold/Muc3B.dta", clear
    collapse (sum) m3c5b m3c6b, by(tinh huyen xa diaban hoso)
    merge 1:1 tinh huyen xa diaban hoso using "$temp/ho_2014.dta"
    drop _m
save "$temp/ho_2014.dta", replace

use "$data/VHLSS_2014/hhold/Muc7.dta", clear
    keep tinh huyen xa diaban hoso m7c1 m7c2 m7c8 m7c17
    merge 1:1 tinh huyen xa diaban hoso using "$temp/ho_2014.dta"
    drop _m
save "$temp/ho_2014.dta", replace
```

3.4 Create panal data

3.4.1 Import the meta data

Since same variable can have different name in each year, we need to change them into 1 name to append together. This metadata will be used to rename, keep and label all wanted variables.

```
import delimited "$data/VHLSS_codebook_9k - ho.csv", varnames(1) encoding(UTF-8) clear
```

3.4.2 Rename and label code

```
local N = _N

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

    local renamed_vars ""
    forval iii = 1/`N' {
        local vvvcode = code[`iii']
        local vvv_i = code_`i'[`iii']  // Chuyển từ vvv`i' thành vvv_i để tránh lỗi cú
        local v_desc = description[`iii']

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

                  // 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

tempfile combined_data
save `combined_data', emptyok

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

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

```
      gen year = `i'

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

    save `combined_data', replace
}

save "$clean/vhlss_14_18.dta", replace
```

Chapter 4

Important note during cleaning

4.1 Linking VHLSSs

VHLSS is a rotating panel dataset. It is possible to construct a panel across three survey rounds (e.g., 2014-2016-2018) if the data is designed based on the same Population and Housing Census. For instance, the household system from 2010 to 2018 was designed from the 2009–2010 Census. However, because VHLSS 2020 was designed based on the 2019 Census, it is not possible to create a panel linking VHLSS 2016–2018 with 2020.

4.2 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.

4.3 Inconsistent industry codes

VSIC1993 is the basis of industry codes used in the 2002 through 2006 VHLSSs, the 2000 through 2007 enterprise data, and the 1999 population census. VSIC2007 is used in the 2008 through 2018 VHLSSs, the 2008 through 2017 enterprise data, and the 2009 population census. VSIC2018 is used in the 2019 population census.

4.4 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 $= \frac{\text{Weight household}}{\text{Houshold size}}$