

Arnaud Natal^o

RUME-NEEMISIS survey: How to Handle Panel Data?

Abstract: This short note aims to provide the key to handling the RUME (2010), NEEMISIS-1 (2016-17), and NEEMISIS-2 (2020-21) panel data sets, a first-hand survey panel collected within the Observatory of Rural Dynamics and Inequalities. After briefly reviewing the data collected in ten villages in rural South India and cross-sectional identifiers, this note provides the details necessary to assemble these surveys longitudinally. An example from the *Stata* software and construction keys for future waves are provided.

Keywords: India, panel data, methodology.

^oarnaud.natal@u-bordeaux.fr. ORCID: 0000-0003-1301-2281. Univ. Bordeaux, and IFP.

1 Introduction

This short note aims to provide the key to handling the RUME (2010), NEEMSI-1 (2016-17), and NEEMSI-2 (2020-21) panel data sets, a first-hand survey panel collected within the Observatory of Rural Dynamics and Inequalities. After briefly reviewing the data collected in ten villages in rural South India and cross-sectional identifiers, this note provides the details necessary to assemble these surveys longitudinally. An example from the *Stata* software and construction keys for future waves are provided.

2 Data

The RUME survey was conducted in 2010 in ten villages in coastal/central Tamil Nadu in the Cuddalore district and Kallakurichi district (ex-Viluppuram district), a mostly agricultural area. The villages benefit from the proximity of two large industrial towns (Neyveli and Cuddalore) and a regional business centre (Panruti). The RUME survey randomly selected 405 households using a stratified sample framework based on three dimensions: proximity to small towns (Panruti, Villupuram, and Cuddalore), an agroecological criterion, and caste affiliation. Thus, half of the villages have irrigated land (the other half is dry), and within villages, half of the sample was selected from the most upper and middle caste part of the village, “Ur”. In contrast, the other half comes from the “Colony” part, where Dalits mainly live. The random route sample method was operated to choose households: enumerators, by a team of two, interviewed a household every five houses. For more details regarding the RUME wave, see Gu erin et al. (2023).

NEEMSI has been built on the RUME survey and has two waves carried out respectively in 2016-17 (NEEMSI-1) and 2020-21 (NEEMSI-2). NEEMSI added a new survey unit compared to RUME: the individual or “ego” level. In 2016-17, two household members were directly addressed individual questionnaires: the respondent of the household questionnaire (called “ego 1”) and one younger household member (“ego 2”). In 2020-21, an additional ego was added, bringing the number of individuals responding to the individual questionnaire to three. Of the 405 households surveyed in 2010, 388 were re-surveyed in 2016-17 with NEEMSI-1, and 104 new households were randomly added –using the same method as for RUME– to refresh the sample. The final sample in 2016-17 consists of 492 households. The NEEMSI-2 (2020-21) wave is the third wave of data collection. The survey recovered 485 households from the NEEMSI-1 survey and randomly selected 147 new households –using

the same method as for RUME– to increase the sampled population. The final sample in 2020-21 consists of 632 households. For more details regarding the NEEMSI-1 wave, see Nordman et al. (2017). For the NEEMSI-2 wave, see Nordman et al. (2021).

These surveys (RUME, NEEMSI-1, and NEEMSI-2) constituted an original longitudinal data collection tool on more than 600 households from 10 villages in rural Tamil Nadu. For more information regarding the RUME-NEEMSI panel data, see Di Santolo et al. (2023).

Table 1: Population of the RUME (2010), NEEMSI-1 (2016-17), and NEEMSI-2 (2020-21) waves. – Authors’ calculations.

	RUME (2010)	NEEMSI-1 (2016-17)	NEEMSI-2 (2020-21)
Number of households			
<i>Cross-sectional</i>	n=405	n=492	n=632
<i>Panel 2010 / 2016-17</i>	n=388		
<i>Panel 2016-17 / 2020-21</i>			n=485
<i>Panel 2010 / 2016-17 / 2020-21</i>	n=382		
Number of individuals			
<i>Cross-sectional</i>	n=1928	n=2696	n=3647
<i>Panel 2010 / 2016-17</i>	n=1826		
<i>Panel 2016-17 / 2020-21</i>			n=2628
<i>Panel 2010 / 2016-17 / 2020-21</i>	n=1783		

3 A reminder on the cross-sectional identifier

3.1 Household identifier

In RUME (2010), NEEMSI-1 (2016-17) and NEEMSI-2 (2021-21), the household is the main unit of analysis, but some questions are asked at the level of individuals. In all the datasets, HHID- is a unique household identifier.

Wave	Household unique identifier
RUME (2010)	HHID2010
NEEMSI-1 (2016-17)	HHID2016
NEEMSI-2 (2020-21)	HHID2020

A unique household identifier means that each household has a different identifier. This identifier makes it possible to link the household between the different databases of the same wave.

3.2 Individual identifier

For all the waves inside each household, INDID- is a unique individual identifier.

Wave	Individual relative unique identifier
RUME (2010)	INDID2010
NEEMSI-1 (2016-17)	INDID2016
NEEMSI-2 (2020-21)	INDID2020

However, this variable is not unique in absolute terms. It is unique only within a household. For instance, in the 2010 wave, several individuals are coded “F1” in the data, but only one per household.

Thus, to obtain a unique and absolute individual identifier, i.e., each individual has a different identifier, HHID- and INDID- should be combined.

Wave	Individual absolute unique identifier
RUME (2010)	INDID2010 combined with HHID2010
NEEMSI-1 (2016-17)	INDID2016 combined with HHID2016
NEEMSI-2 (2020-21)	INDID2020 combined with HHID2020

As previously stated, the two NEEMSI waves add the individual level through the individual “ego” questionnaire. The variable `ego id` identified the number of the “ego”. In 2016-17, they are two, three in 2020-21.¹ This unique “ego” identifier within each household takes the value “1”, “2”, or “3” depending whether the individual is “ego 1”, “ego 2”, or “ego 3”. Thus, the individual data can be merged with the household data either by the unique combination HHID- and INDID- or by the unique combination HHID- and `ego id`.

4 The longitudinal identifier

4.1 Presentation

As previously stated:

- of the 405 households interviewed in 2010, 388 were re-interviewed in 2016-17 as part of the NEEMSI-1 collection,
- of the 492 households interviewed in 2016-17, 485 were re-interviewed in 2020-21 as part of the NEEMSI-2 collection,

¹Two household members were directly addressed individual questionnaires: the respondent of the household questionnaire (called “ego 1”) and one younger household member (“ego 2”). In 2020-21, an additional ego was added, bringing the number of individuals responding to the individual questionnaire to three.

- and 382 households were surveyed in the three waves (2010, 2016-17, and 2020-21).

In order to identify households and individuals, two key variables were created: HHID_panel and INDID_panel. These two variables allow households and individuals to be merged within the panel.

At the household level, HHID_panel is a unique household identifier across the waves, i.e., each household surveyed with RUME, NEEMSIS-1, and/or NEEMSIS-2 has a different identifier.

At the individual level, inside each household, INDID_panel is a unique longitudinal individual identifier. However, as for the cross-sectional identifiers, this variable is not unique in absolute terms. It is unique only within a household. Thus, INDID_panel, combined with HHID_panel, is a unique individual identifier across the waves.

4.2 Construction

HHID_panel is constructed by merging the first letter of the village of origin (see Table 2) with a unique code. The unique code was given by starting at the value “1” and adding 1 for the next household. For instance, in Elamthampattu, HHID_panel goes from “ELA1” to “ELA65”. That is, since 2010, 65 unique households have been surveyed in this village under either RUME, NEEMSIS-1 and/or NEEMSIS-2.

INDID_panel is constructed by merging “Ind_” with a unique code. The unique code was given by starting at the value “1” and adding 1 for the next household member. For instance, for the household “ELA12”, the first unique longitudinal individual identifier is “Ind_1” and the last is “Ind_6”. In other words, since 2010, in the household “ELA12”, six individuals were surveyed either RUME, NEEMSIS-1 and/or NEEMSIS-2.

4.3 Technical handling

These variables are available in four datasets, available upon request:²

- *RUME_NEEMSIS_panel-HH_wide.dta*
- *RUME_NEEMSIS_panel-HH_long.dta*
- *RUME_NEEMSIS_panel-indiv_wide.dta*
- *RUME_NEEMSIS_panel-indiv_long.dta*

²<https://neemsis.hypotheses.org/contact>

Table 2: Range of HHID_panel, NEEMSI-2 (2020-21) updates – Authors’ calculations.

Village	Code	1 to ... (n)
Elamthampattu	ELA	65
Govulapuram	GOV	67
Karumbur	KAR	65
Korattore	KOR	57
Kuvagam	KUV	67
Natham	NAT	63
Oraiur	ORA	69
Semakottai	SEM	65
Manamthavizhthaputhur	MANAM	65
Manappakam	MAN	69
Total		652

4.3.1 *-HH_wide.dta*

The first dataset *-HH_wide.dta* combined key household information (HHID, address, village, village area, and caste) in 2010, 2016-17 and 2020-21. The dataset is in “wide” format, i.e. the time dimension is in columns.

One observation is a unique household, with a unique HHID_panel. There are 652 observations, corresponding to the unique 652 households interviewed either in RUME (2010), NEEMSI-1 (2016-17) and/or NEEMSI-2 (2020-21). In other words, since the beginning of the data collection, i.e. 2010, 652 different households were interviewed. This includes both panel and non-panel households.³

For each row, HHID_panel is filled and linked with HHID2010, HHID2016, and/or HHID2020.

- If HHID2010 (HHID2016, or HHID2020) is empty, the household was not interviewed in 2010 (2016-17 or 2020-21).
- If several HHID- (-2010, -2016, and/or -2020) are filled, this means that the household is in a panel setting, with the waves filled.

For instance, if only HHID2010 and HHID2016 are filled, the household was interviewed in 2010 and 2016-17, but not in 2020-21.

Thus, the dataset *-HH_wide.dta* can be merged with the RUME, NEEMSI-1, or NEEMSI-2 datasets through the household cross-sectional identifier, to provide the unique household panel identifier in the RUME, NEEMSI-1 and/or NEEMSI-2 datasets. The merging can be

³Of these 652 households (see Table 1): 405 were interviewed in 2010, 492 were interviewed in 2016-17, 632 were interviewed in 2020-21, 388 were interviewed in 2010 and 2016-17, 485 were interviewed in 2016-17 and 2020-21, 382 were interviewed in 2010, 2016-17, and 2020-21.

achieved with *Stata* through the merge command (Acock, 2023).⁴

4.3.2 *-HH_long.dta*

The second dataset, *-HH_long.dta*, is the same as the first one (*-HH_wide.dta*), but in “long” format. While we had one observation per household with the time dimension in columns (“wide” format), here we have one observation per household per wave, i.e. the time dimension is in rows.

One row contains information for one household for a given wave. The total number of observation is 1529, corresponding to the sum of the 405 households interviewed in 2010, the 492 of 2016-17 and the 632 of 2020-21 (see Table 1).

HHID_panel is not unique per row. Indeed, a household can have up to three rows if it was surveyed in 2010, 2016-17 and 2020-21.

Switching from “wide” to “long” format is easily done with standard statistical software. For instance, with *Stata*, the reshape command allows this (Acock, 2023).

4.3.3 *-indiv_wide.dta*

-indiv_wide.dta combined key individual information (name, age, sex, village) in 2010, 2016-17 and 2020-21. The dataset is in “wide” format, i.e. the time dimension is in columns.

One observation is a unique individual, with a unique INDID_panel within a given household. There are 3805 observations, corresponding to the 3805 unique individuals interviewed either in RUME (2010), NEEMSI-1 (2016-17) and/or NEEMSI-2 (2020-21).

For each row, INDID_panel is filled and linked with INDID2010, INDID2016, and/or INDID2020.

- If INDID2010 (INDID2016, or INDID2020) is empty, the individual was not interview in 2010 (2016-17 or 2020-21).
- If several INDID- (-2010, -2016, and/or -2020) are filled, this means that the individual is in a panel setting.

For instance, if only INDID2010 and INDID2016 are filled, the individual was interviewed in 2010 and 2016-17, but not in 2020-21.

However, as already stated, INDID_panel is not itself a unique identifier, this variable must be combined with HHID_panel.

⁴See the section 5 for an example.

Thus, with the support of `HHID_panel`, the dataset `-indiv_wide.dta` can be merged with the RUME, NEEMSI-1, or NEEMSI-2 datasets through the individual cross-sectional identifier to provides the unique individual panel identifier in the RUME, NEEMSI-1 and/or NEEMSI-2 datasets. The merging can be achieved with *Stata* through the `merge` command (Acock, 2023).⁵

4.3.4 `-indiv_long.dta`

Last, `-indiv_long.dta`, is the same as `-indiv_wide.dta`, but in “long” format.

One row contains information for one household member for a given wave. The total number of observation is 8271, corresponding to the sum of the 1928 individual interviewed in 2010, the 2696 of 2016-17 and the 3647 of 2020-21 (see Table 1).

`INDID_panel` is not unique per row. Indeed, a household member can have up to three rows if it was surveyed in 2010, 2016-17 and 2020-21.

5 Example

As an illustration, we wish to bring together in a single database the participation in the labour market, measured with the variable `dummyworkedpastyear`. This variable is at the individual level.

Step 1: The first step consists of merging `-HH_wide.dta` with each dataset on the basis of the unique cross-sectional identifier to obtain the `HHID_panel`.

```
. ***** Step 1: Adding HHID_panel
. *** RUME (2010)
. use "RUME-HH", clear

. count
1,928

. merge m:m HHID2010 using "RUME_NEEMSI_panel-HH_wide.dta", keepusing(HHID_panel)
```

Result	# of obs.
not matched	247
from master	0 (_merge==1)
from using	247 (_merge==2)

⁵See the section 5 for an example.

```

    matched                                1,928  (_merge==3)
    -----

. keep if _merge==3
(247 observations deleted)

. count
1,928

. keep HHID2010 INDID2010 HHID_panel dummyworkedpastyear

. save "RUME-HH_v2", replace
file RUME-HH_v2.dta saved

.
. *** NEEMSI1-1 (2016-17)
. use "NEEMSI1-HH", clear

. count
2,696

. merge m:m HHID2016 using "RUME_NEEMSI1_panel-HH_wide.dta", keepusing(HHID_panel)

Result                                # of obs.
-----
not matched                            160
    from master                          0  (_merge==1)
    from using                            160 (_merge==2)

matched                                2,696  (_merge==3)
-----

. keep if _merge==3
(160 observations deleted)

. count
2,696

. keep HHID2016 INDID2016 HHID_panel dummyworkedpastyear

. save "NEEMSI1-HH_v2", replace
file NEEMSI1-HH_v2.dta saved

.
. *** NEEMSI2-2 (2020-21)
. use "NEEMSI2-HH", clear

. count
3,647

. merge m:m HHID2020 using "RUME_NEEMSI2_panel-HH_wide.dta", keepusing(HHID_panel)

```

```

Result                                # of obs.
-----
not matched                            20
  from master                          0  (_merge==1)
  from using                            20  (_merge==2)

matched                                3,647  (_merge==3)
-----

. keep if _merge==3
(20 observations deleted)

. count
3,647

. keep HHID2020 INDID2020 HHID_panel dummyworkedpastyear

. save "NEEMISIS2-HH_v2", replace
file NEEMISIS2-HH_v2.dta saved

. log close

```

Step 2: Merging INDID_panel. The second step consists of merging *-indiv_wide.dta* with each dataset on the basis of the unique cross-sectional identifier and HHID_panel to obtain the INDID_panel.

```

. ***** Step 2: Adding INDID_panel
. *** RUME (2010)
. use "RUME-HH_v2", clear

. count
1,928

. merge 1:m HHID_panel INDID2010 using "RUME_NEEMISIS_panel-indiv_wide.dta", keepusing(INDID_panel)

Result                                # of obs.
-----
not matched                            1,877
  from master                          0  (_merge==1)
  from using                            1,877  (_merge==2)

matched                                1,928  (_merge==3)
-----

. keep if _merge==3
(1,877 observations deleted)

. count

```

```

1,928

. keep HHID2010 INDID2010 HHID_panel INDID_panel dummyworkedpastyear

. rename dummyworkedpastyear dummyworkedpastyear2010

. save "RUME-HH_v3", replace
file RUME-HH_v3.dta saved

.
. *** NEEMSI1-1 (2016-17)
. use "NEEMSI1-HH_v2", clear

. count
2,696

. tostring INDID2016, replace
INDID2016 was byte now str2

. merge 1:m HHID_panel INDID2016 using "RUME_NEEMSI1-panel-indiv_wide.dta", keepusing(INDID_panel)

Result                                # of obs.
-----
not matched                            1,109
   from master                          0   (_merge==1)
   from using                            1,109 (_merge==2)

matched                                2,696 (_merge==3)
-----

. keep if _merge==3
(1,109 observations deleted)

. count
2,696

. keep HHID2016 INDID2016 HHID_panel INDID_panel dummyworkedpastyear

. rename dummyworkedpastyear dummyworkedpastyear2016

. save "NEEMSI1-HH_v3", replace
file NEEMSI1-HH_v3.dta saved

.
. *** NEEMSI2-2 (2020-21)
. use "NEEMSI2-HH_v2", clear

. count
3,647

. tostring INDID2020, replace

```

```

INDID2020 was byte now str2

. merge 1:m HHID_panel INDID2020 using "RUME_NEEMIS_panel-indiv_wide.dta", keepusing(INDID_panel)

Result                                     # of obs.
-----
not matched                                158
  from master                               0  (_merge==1)
  from using                                158 (_merge==2)

matched                                    3,647 (_merge==3)
-----

. keep if _merge==3
(158 observations deleted)

. count
3,647

. keep HHID2020 INDID2020 HHID_panel INDID_panel dummyworkedpastyear

. rename dummyworkedpastyear dummyworkedpastyear2020

. save "NEEMIS2-HH_v3", replace
file NEEMIS2-HH_v3.dta saved

. log close

```

Step 3: Merging new databases together. Last, the three new databases can be merged into a single one using HHID_panel and INDID_panel.

```

. ***** Step 3: Merging together
. *** Open data set
. use "RUME-HH_v3", clear

.
. *** Merge RUME and NEEMIS-1
. merge 1:1 HHID_panel INDID_panel using "NEEMIS1-HH_v3"

Result                                     # of obs.
-----
not matched                                972
  from master                               102 (_merge==1)
  from using                                870 (_merge==2)

matched                                    1,826 (_merge==3)
-----

. rename _merge merge_2016

```

```

.
. *** Add NEEMSIS-2
. merge 1:1 HHID_panel INDID_panel using "NEEMSIS2-HH_v3"
(label yesno already defined)

Result                                # of obs.
-----
not matched                            1,165
  from master                          158  (_merge==1)
  from using                            1,007 (_merge==2)

matched                                2,640 (_merge==3)
-----

. rename _merge merge_2020

.
. *** Statistics
. sort HHID_panel INDID_panel

. list HHID_panel INDID_panel dummyworkedpastyear2010 dummyworkedpastyear2016
      dummyworkedpastyear2020 in 1, noobs

+-----+
| HHID_p~1  INDID_~1  dum~2010  dum~2016  dum~2020 |
|-----|
|      ELA1      Ind_1      Yes      Yes      Yes |
+-----+

. log close

```

6 For the future

As previously stated, HHID_panel is constructed by merging the first letter of the village of origin (see Table 2) with a unique code. The unique code was given by starting at “1” and adding 1 for the next household. Thereby, for the next data collection, in addition to a cross-sectional identifier, it will be appropriate to:

- only linking the surveyed household with its longitudinal identifier HHID_panel if the household was already interviewed in RUME (2010), NEEMSIS-1 (2016-17), and/or NEEMSIS-2 (2020-21).
- assigning a new longitudinal identifier only for new households. Starting from the last identifier given in the surveyed household’s village of origin and adding 1. For

instance, if five new households are interviewed in the future NEEMSI-3 wave in Elamthampattu, they should be identified as “ELA66”, “ELA67”, “ELA68”, “ELA69”, and “ELA70” (see Table 2).

Regarding INDID_panel, as previously stated, it is constructed by merging “Ind_” with a unique code. The unique code was given by starting at “1” and adding 1 for the next household member. Thereby, for the next data collection, in addition to a cross-sectional identifier, it will be appropriate to:

- only linking the surveyed individual with its longitudinal identifier INDID_panel if the individual was already interviewed in RUME (2010), NEEMSI-1 (2016-17), and/or NEEMSI-2 (2020-21).
- assigning a new longitudinal identifier only for new individuals, starting from the last identifier given in the surveyed households and adding 1. For instance, if two new households members are interviewed in the future NEEMSI-3 wave in Elamthampattu, household “ELA1”, they should be identified as “Ind_6” and “Ind_7”, as the last household member is identified with “Ind_5”.

References

- Acock, A. C. (2023). *A Gentle Introduction to Stata* (6th ed.). Stata Press.
- Di Santolo, M., Guérin, I., Michiels, S., Mouchel, C., Natal, A., Nordman, C. J., & Venkatasubramanian, G. (2023). *Ten years of labour, migration and debt in India: Insights from original longitudinal household surveys in Tamil Nadu* (Working Paper). ODRIIS. Paris, France.
- Guérin, I., Roesch, M., Venkatasubramanian, G., Michiels, S., & Natal, A. (2023). *RUME 2010: Survey Manual* (tech. rep.). IRD-IFP. Paris, France.
- Nordman, C. J., Guérin, I., Venkatasubramanian, G., Michiels, S., Lanos, Y., Kumar, S., Raj, A., & Hilger, A. (2017). *NEEMSI-1: Survey Manual* (tech. rep.) (HAL Id: hal-03751302). IRD-IFP. Paris, France. <https://hal.science/hal-03751302>
- Nordman, C. J., Guérin, I., Venkatasubramanian, G., Michiels, S., Mouchel, C., Natal, A., & Di Santolo, M. (2021). *NEEMSI-2: Survey Manual* (tech. rep.). IRD-IFP. Paris, France.