

Getting to know the SUSADICA dataset in Stata

A practical orientation to the structure and use of the dataset

Alisher Kosimov

**Data Reuse on Sustainable Agricultural Development in Central Asia
(SUSADICA – Data Reuse)**

Purpose of the module

This module provides a **practical introduction** to the **SUSADICA dataset** in **Stata**. It helps users:

- Understand the **dataset's structure** and **organization**
- Navigate variables using **names, labels**, the **questionnaire**, and the **codebook**
- Explore **key identifiers** and **panel features**
- Recognize and interpret different **variable types**
- Apply **good practices** for **reproducible data work**

Outcome: You will be prepared to confidently explore, understand, and use the SUSADICA dataset for further analysis.

What you need for this module

- **Stata** installed on your computer
- **SUSADICA dataset** (.dta file)
- **Questionnaire**, codebook or variable list
- (data package is available in Harvard Dataverse repository at <https://doi.org/10.7910/DVN/DJWILN>)
- **Project folder** for do-files, logs, outputs, and clean data

Stata, R or Python: Why we start with Stata?

Feature	 STATA	 R	 Python
Learning curve	✓ Easier to learn; simple, consistent syntax with graphical user interface and command line.	⚠ Steeper learning curve; less consistent as a full programming language	⚠ Steeper learning curve; highly versatile and used for general-purpose programming.
Core functionality	✓ A comprehensive statistical toolbox is included in the base package, with extensive built-in functionality.	Requires installing and managing multiple community-contributed packages for specific analyses, which can sometimes conflict.	Primarily relies on external libraries like pandas and statsmodels for data analysis.
Reproducibility	★ Strong focus on backwards compatibility and version control, which is essential for academic research.	⚠ Can be less reliable for reproducing older analyses due to frequent package updates.	⚠ Requires more effort and specific tooling to ensure reproducibility across different computing environments.
Community	Niche but highly engaged user community, particularly in the social sciences; excellent paid and community support.	Large and active open-source community, though support is not centralized.	Massive community spanning all areas of programming, not limited to statistics.
Visualization	Produces publication-quality graphics; less flexible and customizable than R/Python.	★ Highly advanced and customizable graphics via packages like ggplot2 .	★ Strong visualization capabilities with libraries like Matplotlib and Seaborn .
Cost	⚠ Paid license.	✓ Free & open-source.	✓ Free & open-source.



Keeps the
command
history

Lists the variables
after we open the
dataset

Results window
shows the output after we run commands

Shows information
about the selected
variable

Commands can be typed directly

	b52_total	a3	c1	bbl_2022	country	year								
1347	23	Secondary general	Willing	.	uzb	2022								
1348	19	Higher	Neutral	.	uzb	2022								
1349	14.4	Secondary general	Neutral	.	uzb	2022								
1350	28	Secondary general	Very willing	No	uzb	2022								
1351	19	Secondary general	Willing	.	uzb	2022								
1352	27.2	Higher	Very willing	No	uzb	2022								
1353	30.5	Secondary general	Willing	.	uzb	2022								
1354	25.1	Secondary special (college)	Very willing	No	uzb	2022								
1355	23	Higher	Unwilling	.	uzb	2022								
1356	20.3	Secondary general	Willing	No	uzb	2022								
1357	22.6	Secondary general	Willing	.	uzb	2022								
1358	33	Secondary general	Unwilling	Yes	uzb	2022								
1359	57.5	Higher	Unwilling	Yes	uzb	2022								
1360	67.3	Secondary general	Unwilling	Yes	uzb	2022								
1361	73	Secondary general	Neutral	Yes	uzb	2022								
1362	42.8	Secondary general	Unwilling	Yes	uzb	2022								
1363	48	Secondary professional (vocational school)	Willing	Yes	uzb	2022								
1364	92	Secondary professional (vocational school)	Neutral	Yes	uzb	2022								
1365	34	Higher	Unwilling	Yes	uzb	2022								
1366	50	Secondary professional (vocational school)	Neutral	Yes	uzb	2022								
1367	47	Secondary special (college)	Unwilling	Yes	uzb	2022								
1368	48	Higher	Neutral	Yes	uzb	2022								
1369	32	Secondary professional (vocational school)	Unwilling	Yes	uzb	2022								
1370	64	Secondary general	Unwilling	Yes	uzb	2022								
1371	86	Higher	Willing	Yes	uzb	2022								
1372	71.5	Higher	Unwilling	Yes	uzb	2022								
1373	77	Higher	Willing	Yes	uzb	2022								
1374	74.5	Secondary professional (vocational school)	Unwilling	Yes	uzb	2022								
1375	100	Secondary general	Unwilling	Yes	uzb	2022								
1376	59	Secondary professional (vocational school)	Willing	Yes	uzb	2022								
1377	44	Secondary general	Unwilling	Yes	uzb	2022								
1378	36	Secondary professional (vocational school)	Willing	Yes	uzb	2022								
1379	63	Secondary professional (vocational school)	Willing	Yes	uzb	2022								
1380	20	Higher	Unwilling	.	uzb	2022								
1381	91	Higher	Neutral	Yes	uzb	2022								
1382	47	Secondary general	Unwilling	Yes	uzb	2022								
1383	49.5	Higher	Unwilling	Yes	uzb	2022								
1384	43.4	Secondary general	Unwilling	Yes	uzb	2022								

 Filter variables here

☒ Name	Label
☒ b52_total	Total land current (ha)
☒ a3	Farm manager education level
☒ c1	Risk willingness (1-5)
☐ farmid	Farm ID number
☒ bb1_2022	Farm in cluster (1/0)
☒ country	2a Country
☒ year	Survey year, 2019/2022
☐ district	Districts ordered
☐ isnearfield	Interview near field (1/0)
☐ mahalla_id	mahalla_id
☐ icode	Interviewer Code
☐ farm_speciali...	Farm specialization (Cat)
☐ irrigated_02	Farm > 80% irrigated (1/0)
☐ a1_2019	In this farm you are..(Cat)
☐ a1_2022	Farm manager ownership role
☐ a2	Year of birth
☐ a4	Farm manager agri education
☐ a5	Agri experience (years)
☐ a6	Main job before farm manage
☐ a6_other	Please, specify other
☐ a7_2022	Household members (numbe

Variables Snapshots

Properties

Variables

Name	bb1_2022
Label	Farm in cluster (1/0)
Type	byte
Format	%10.0g
Value label	lbbb1_2022
Notes	

◀ Data

Filename	3_SUSADICA_clean.dta
Label	
Notes	
Variables	2,475
Observations	1,864
Size	13.95M
Memory	64M
Sorted by	year

Missing values: Why do data go missing?

Possible reasons for missing values:

- **Non-response behavior**
(e.g., sensitive questions left unanswered)
- **Data collection issues**
(survey errors, equipment failure)
- **Measurement limitations**
(variables difficult to observe or record)
- **Not applicable / skip logic –**
(question does not apply to that farm e.g., *“If not a cluster member → skip the cluster section”*)

Important: In skip-logic cases, missing values are **expected** and should not be treated as errors.

Understanding *why* data are missing is important for
correct interpretation and analysis.

Key takeaways

- ✓ The SUSADICA dataset is organized around **farm-year observations** and includes both **pooled cross-sectional** and **panel components**.
- ✓ Survey *year*, *country*, and *farm ID* are **key identifiers** for understanding and working with the data.
- ✓ **Variable names** follow the questionnaire structure, with **wave-specific suffixes** (e.g., `_2019`, `_2022`) and **harmonized variables** without suffixes.
- ✓ **Questionnaire** and **codebook** are essential tools for interpreting variables.
- ✓ Commands such as **describe**, **tab**, **codebook**, **summarize**, and **lookfor** help navigate and understand variables efficiently.
- ✓ Correct analysis begins with understanding **variable types**, **labels**, **codes**, and **missing values**.
- ✓ Missing values often arise from **skip patterns** and are expected rather than errors.
- ✓ Good research practice requires **organized, reproducible, and non-destructive data work**.

Thank you!