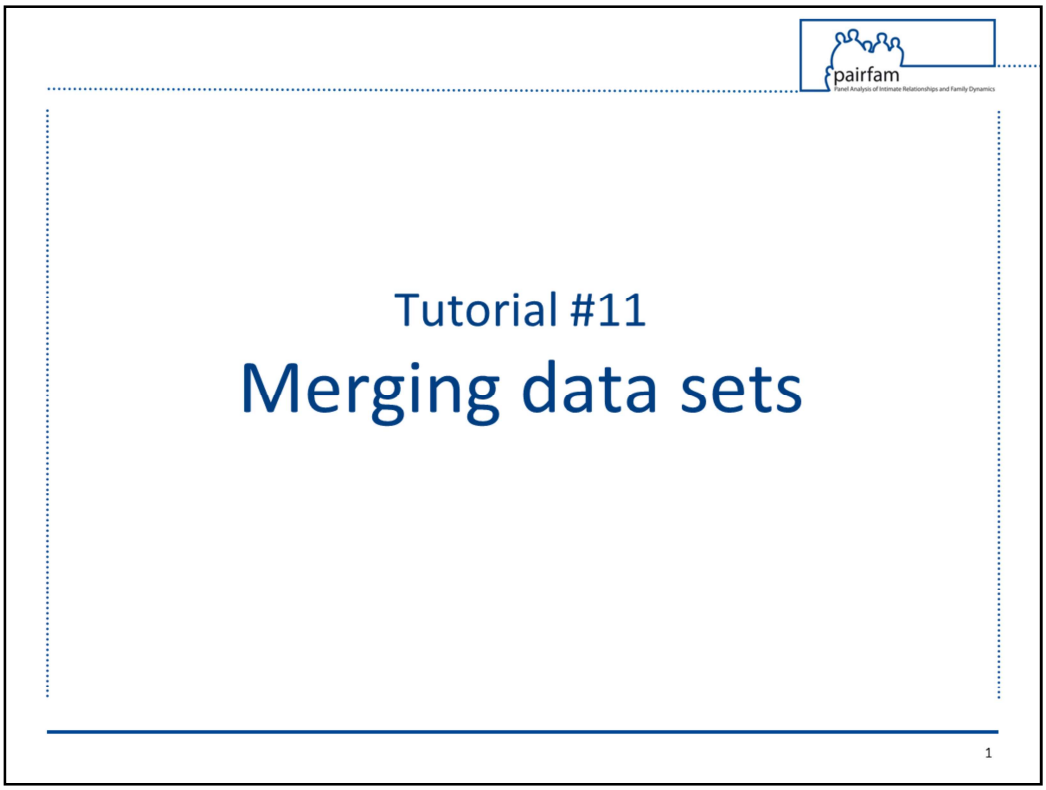
The corresponding video tutorials are available online:
https://www.youtube.com/playlist?list=PL7BcpOtSe5u_zQctYXz4ee79Zc9r4mfnr

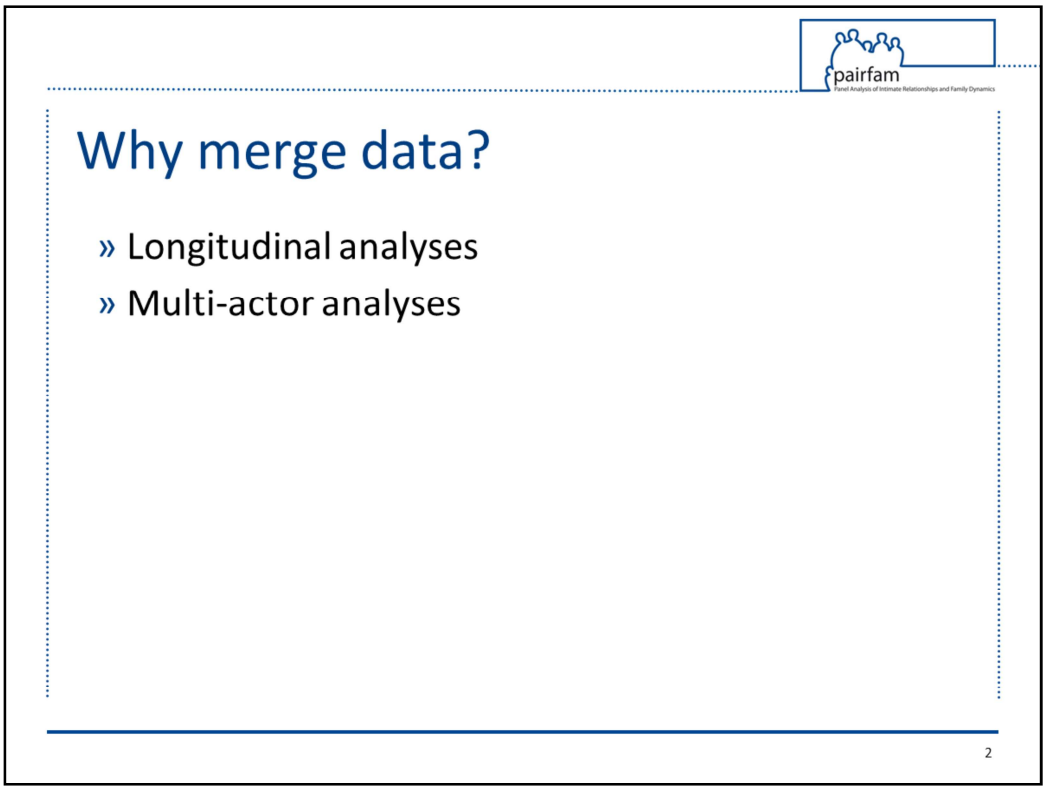


pairfam tutorial

11. Merging data sets

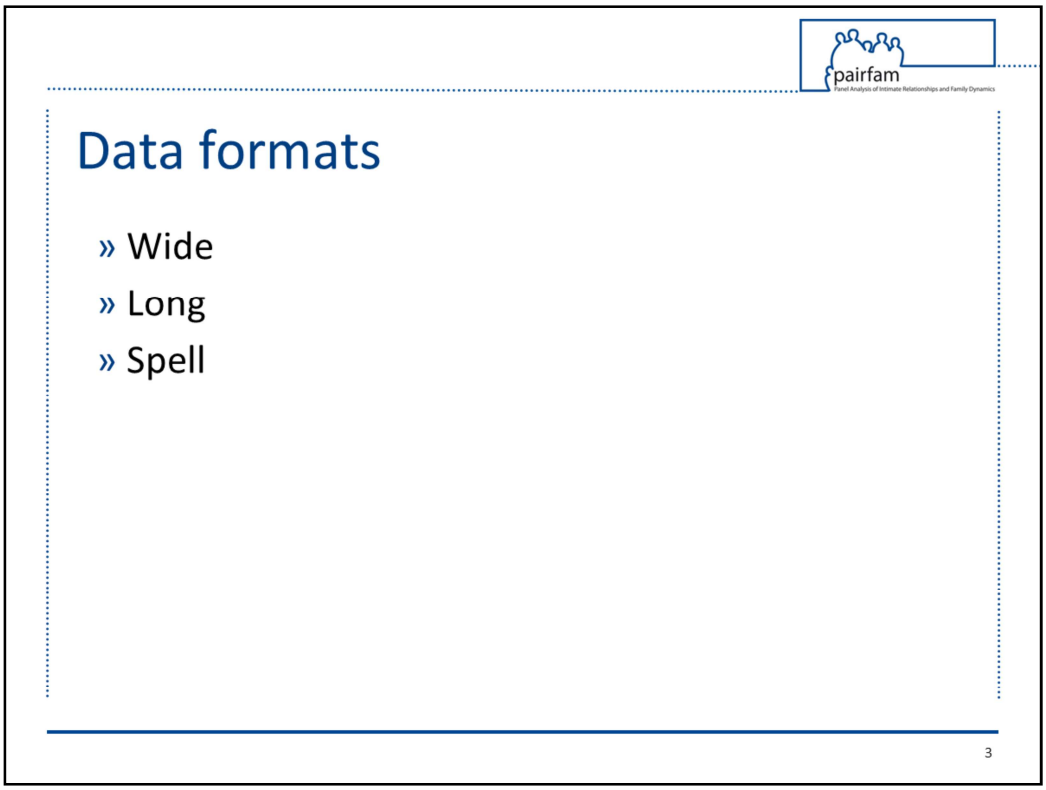
Kristin Hajek & Madison Garrett, May 2022

The eleventh tutorial covers different matching procedures to help you merge the *pairfam* data sets.



Why should we merge the *pairfam* data?

As the survey data is stored as separate cross-sectional data sets and organized by different actors, you will likely need to merge different actor data sets and/or several waves to answer your specific research question. To analyze change over time, you will need to merge the different waves, for example waves 1 to 13 of the anchor data. For multi-actor analyses, you will need to merge the data sets of different actors, for example the anchor with the partner data.



Which data formats exist?

There are various data formats: wide, long, and spell.

In the wide format, each row represents one case with a unique identifier, and time is structured in columns. This means that a separate variable contains information for each time point, for example the variables “satisfaction with life in wave 1“ and “satisfaction with life in wave 2“, and so on.

In the long format, each case covers several rows with a unique combination of identifier and time. Time is structured in rows, meaning that one variable contains information for multiple time points. In this format, one variable “satisfaction with life“ contains the response to life satisfaction in wave 1 in the first row, life satisfaction in wave 2 in the second row, and so on.

Spell format also contains several rows per case. However, the rows don‘t represent different waves, but different statuses of an individual. The biodata sets are stored in spell format. The biopart data set, for example, contains one row per anchor relationship. In the following, the focus will be on merging the data in wide and long format.

Common types of matching in Stata

- » append: Appends new observations over time or of different samples (e.g., DemoDiff)
- » 1:1 merge: Matches two uniquely related actors (e.g., anchor and current partner)
- » 1:m merge: Matches one reference actor with multiple corresponding actors (e.g., anchor and parents)
- » reshape: Change data into wide/long format

4

What are my options to match different data sets in Stata?

The append command appends new observations over time (e.g., waves 1 to 13) or of different samples (e.g., anchor base and *DemoDiff* data), creating a new data set in long format. With the command 1:1 merge, two uniquely related actors can be merged into wide format (e.g., anchor with current partner). 1:m stands for 1:many, and matches one reference actor with multiple corresponding actors, for example one anchor with several parents. By using the reshape command, data sets can be reshaped from wide into long format, or vice verca.

append: anchor1 + anchor1_DD

anchor1

	id	wave	var1	var2	var3
1	100000	1	1	2	3
2	101000	1	5	4	3
3	102000	1	-3	3	-3

anchor1_DD

	id	wave	var1	var2	var3
1	800100000	1	2	3	-10
2	800101000	1	4	3	-10
3	800102000	1	2	-3	-10

anchor1 + anchor1_DD

	id	wave	var1	var2	var3
1	100000	1	1	2	3
2	101000	1	5	4	3
3	102000	1	-3	3	-3
4	800100000	1	2	3	-10
5	800101000	1	4	3	-10
6	800102000	1	2	-3	-10

5

How do I append DemoDiff cases?

On the left you can see three hypthetical rows of the *anchor1* and *anchor1_DD* data sets, which each represent 3 respondents with a unique *id* in wave 1. They have been asked the same questions (columns *var1*, *var2*, and *var3*). By using the append command, we can merge these two data sets into long format - the result is shown on the right. This format contains the same amount of variables and a combined number of rows from both data sets: The wave 1 *DemoDiff* cases have been added to the pairfam base sample from wave 1. Please note that for the append command to run correctly, respondent *id*'s must be unique in both data sets and variable names must be identical.

1:1 merge: anchor1 + partner1

anchor1

	id	wave	var1	var2	var3
1	100000	1	1	2	3
2	101000	1	5	4	3
3	102000	1	-3	3	-3

partner1

	id	pid	wave	pvar1	pvar2	pvar3
1	100000	100101	1	2	3	4
2	101000	101101	1	4	3	2
3	102000	102101	1	2	-3	2

anchor1 + partner1

	id	wave	var1	var2	var3	pid	pvar1	pvar2	pvar3
1	100000	1	1	2	3	100101	2	3	4
2	101000	1	5	4	3	101101	4	3	2
3	102000	1	-3	3	-3	102101	2	-3	2

6

What if I want to use the 1:1 merge command?

For the merge command to run correctly, repondent *id*'s must be identical in both data sets, whereas variable names must differ. When merging anchor data with partner data from wave 1, the number of cases (rows) does not change: The combined data set results in the same three rows, or cases. However, the number of variables has increased. In addition to the variables *id*, *wave*, *var1*, *var2*, and *var3*, the merged data set also includes *pid*, *pvar1*, *pvar2*, and *pvar3*.

1:m merge: anchor2 + parent2

anchor2

	id	wave	var1	var2
1	100000	2	1	2
2	101000	2	5	4
3	102000	2	-3	3
4	103000	2	5	5

parent2

	id	parid	wave	parvar3
1	100000	100301	2	1
2	100000	100302	2	2
3	100000	100304	2	3
4	101000	101301	2	4
5	101000	101302	2	5
6	102000	102303	2	-3

anchor2 + parent2

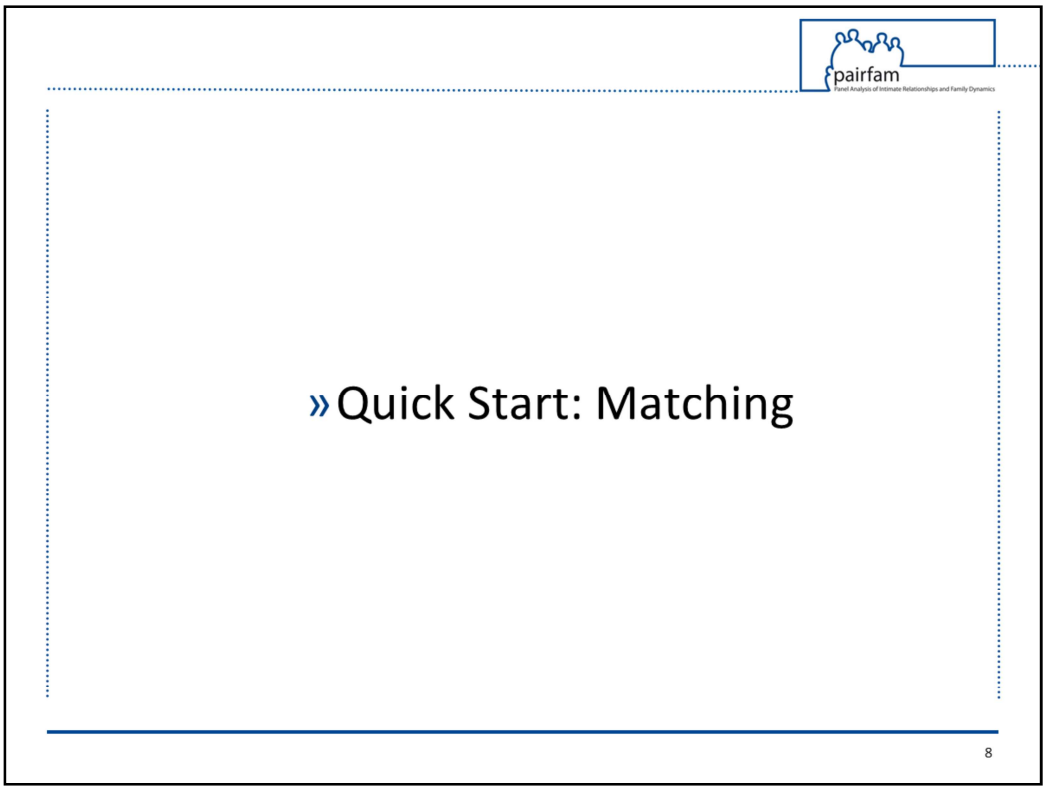
	id	wave	var1	var2	parid	parvar3
1	100000	2	1	2	100301	1
2	100000	2	1	2	100302	2
3	100000	2	1	2	100304	3
4	101000	2	5	4	101301	4
5	101000	2	5	4	101302	5
6	102000	2	-3	3	102303	-3
7	103000	2	5	5	.	.

7

When do I use the 1:m merge command?

As several parents and children of one anchor person may have also been interviewed, it is possible that several parents and children of one anchor respondent are available in the data. Looking at the anchor and parent data from wave 2 on the left side, we see that the anchor with the *id* 100000 is represented with three rows in the parent data, meaning three parents of this anchor have been interviewed. The variable *parid* gives the identification number of the parent. In order to merge the data from all three parents to the anchor data, we need to run the 1:many merge: merging one anchor to “many“ parents. This process enlarges both the number of cases as well as the number of variables.

Always check the resulting data carefully after matching different data sets, for example with the browse command, in order to make sure that the merging process was successful.



Further and more detailed examples on how to match the pairfam data sets can be found in the Quick Start “Matching“.

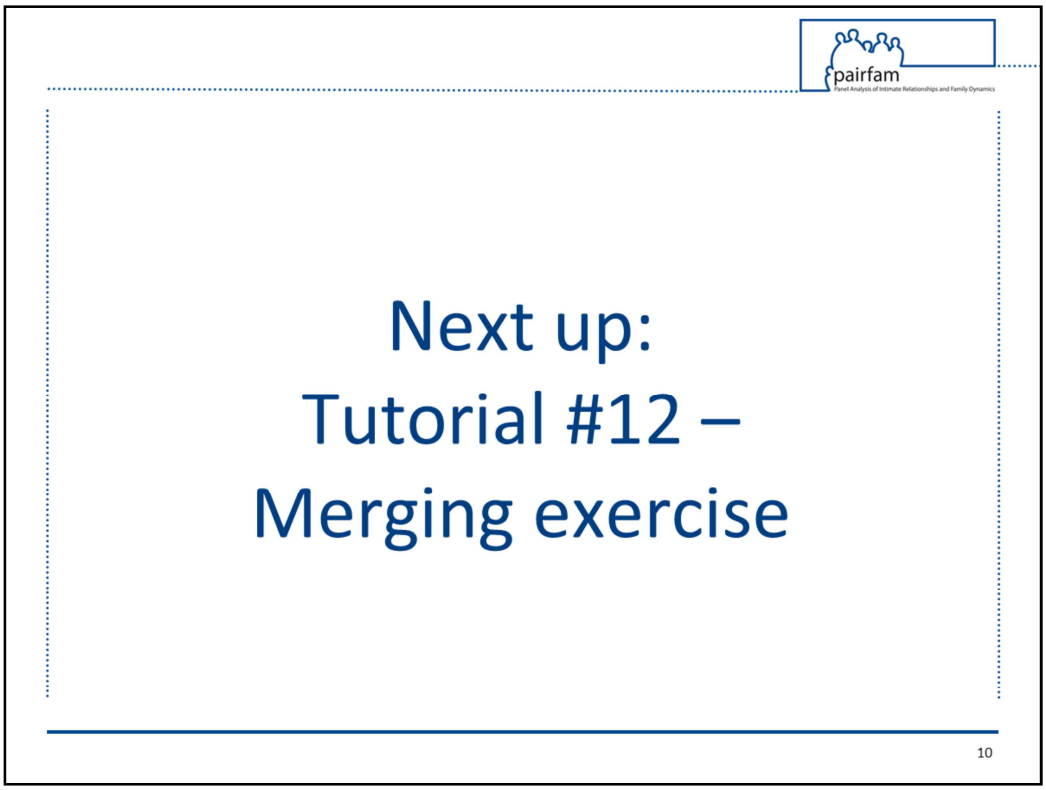
Gender-specific analyses

- Dyads of anchor and partner in wide format
- Problem: gender comparison
- Solution: gender-specific coding of variables

9

One last note on gender-specific analyses: If you have merged anchor and partner data into wide format, the data set will contain one row per anchor-partner dyad and several variables from both the anchor and partner respondents (for example, both the anchor‘s and the partner‘s satisfaction with the relationship). When conducting gender comparisons, please be aware that the anchor (and partner) can be male or female.

One solution to this problem is gender-specific variable coding. For example, you could create a new variable representing women‘s satisfaction with the relationship by using the anchor‘s information if the anchor is female, and the partner‘s information if the partner is female. And then create a new variable representing men‘s satisfaction with the relationship with the same logic. Please keep in mind that there are also homosexual relationships represented in the *pairfam* data.



This marks the end of the eleventh tutorial.

The next tutorial covers a hands-on exercise on how to merge different *pairfam* data sets.