VERGLEICHENDE SOZIALFORSCHUNG MIT MEHREBENENMODELLEN IN R

Forschungspraktikum I und II
Dr. Christian Czymara
Comparative social research

AGENDA

- Why are country comparisons interesting?
- How to compare countries
- How to combine different data sets
- Tutorial: Relationship between economic conditions and life satisfaction

COMPARATIVE RESEARCH

COMPARATIVE RESEARCH

- In a way, all empirical research is comparative
- For example, comparing groups, households, schools, corporations, products, …
- Here:
  - comparative ≙ multi-national
  - contexts ≙ countries

→*comparing countries*

RESEARCH QUESTIONS

- Many research questions deal with differences of individual outcome due to some other individual characteristic(s)
- But are such relationships universal (i.e., do they hold in different contexts)?
- If not: Why?
- Moreover: Other research is interested in differences in individual outcomes due to *contexts*

EXAMPLE

- Individual level relationship (micro-level only)
  - For example: Job loss affects well-being
- Country characteristic and individual outcome (macro-micro link)
  - For example: Economic conditions shape well-being
- Moreover, the former may depend on the latter
  - For example: Job loss has an stronger effect on well-being in economically poor countries

LIFE SATISFACTION ACROSS COUNTRIES

- Substantive differences of average life satisfaction across countries

- Two possible explanations:
  - Explanation 1: *Individuals* living in country A differ in a relevant way from those in country B (*composition effect*)
  - Explanation 2: *Country A itself* differs from country B in a way that affects individuals (*context effect*)

- Explanation 1: Individual-level relationships (difference in y across countries result of a different distribution of *individual-level* x)

- Explanation 2: Influences beyond the individual-level

LIFE SATISFACTION AND INCOME DISSATISFACTION

- Hypothesis: Less income satisfaction leads to less life satisfaction
- Supported by the data: Income explains almost 50 percent of the differences between countries (more next session)
- Which kind of explanation is this, composition or country effect?
- What would be an example of the other effect?

SUMMING UP

- Individual life satisfaction (y) varies across European countries
- Possible explanations:
  - Income (x) important for life satisfaction and individual income levels differ between countries (composition effect)
  - National unemployment rate, social system etc. (z) influence life satisfaction (country effect)

TYPES OF COMPARATIVE RESEARCH

COMPARATIVE RESEARCH

- Countries as object of research
- Countries as units of analysis
- Countries as contexts

COUNTRIES AS OBJECT OF RESEARCH

- Interest in certain countries
- Choice of countries crucial and often based on theoretical considerations
- In depth case studies of few countries
- Use of quantitative and / or qualitative methods possible
- For example: Comparison of social systems in Sweden, Germany and the US
- Potential problems: Generalizability? Causality?

COUNTRIES AS UNITS OF ANALYSIS

- Relate macro-level characteristics to another
- ”Translate countries into variables“
- Larger samples of countries necessary
- Quantitative methods for statistical inference
- For example: Correlation between national unemployment rate and vote share of extreme right-wing parties
- Problem: *ecological fallacy*
- Hence, we will also not (only) look at macro-level relationships

ECOLOGICAL FALLACY

- Inference on the individual level based on macro-level relationships
- Example: Unemployment and far-right voting
- Finding: Far-right parties more successful where unemployment is high
- Can you conclude that the unemployed are more likely to vote far-right?

FICTIONAL EXAMPLE

	Country A	Country B
Unemployment rate (%)	20	40
Far-right vote share	2	4

UNDERLYING DISTRIBUTION

Country A	Unemployed	Employed	Total
Far-right vote share	0	2	2
Other parties vote share	20	78	98
Total	20	80	100

Country B	Unemployed	Employed	Total
Far-right vote share	0	4	4
Other parties vote share	40	56	96
Total	40	60	100

UNDERLYING DISTRIBUTION

Country A	Unemployed	Employed	Total
Far-right vote share	**0**	2	2
Other parties vote share	**20**	78	98
Total	20	80	100

Country B	Unemployed	Employed	Total
Far-right vote share	**0**	4	4
Other parties vote share	**40**	56	96
Total	40	60	100

CONCLUSION

- No unemployed voted far-right in either country
- Hypothesis refuted on the individual-level
- More fine-grained data needed (here: voting within each labor market status groups)
- However, one might still argue that unemployment has a *context effect*
- E. g.: Living with many unemployed affects voting behavior of the employed

COUNTRIES AS CONTEXTS

- How do individuals in certain countries act?
- Do national contexts affect (or moderate) individual actions?
- Possible with qualitative (generally few countries) & quantitative methods (many countries)

HIERARCHICAL LINEAR MODELS

- Combines the *country as contexts* approach with the *countries as units of analysis* approach
- Countries as contexts
  - Modelling individual-level differences between different contexts
- Countries as units of analysis
  - Understanding countries as units that have variables
  - Modelling how country-level variables affect the individual-level outcome variables
  - Many countries needed

QUANTITATIVE COUNTRY COMPARISONS

- Using micro- (x) or macro-level (z) characteristics as independent variables to explain a micro-level outcome (y)
- Moreover, one can test whether certain individual level relationships (x → y) depend on country characteristics (z)
  - For example, is the effect of unemployment on life satisfaction especially strong in countries with low levels of social security?
  - So-called *cross-level interactions*

NEW INSIGHTS OF COUNTRY COMPARISONS

- Generalizability of individual-level findings
  - Does the effect of x on y hold across different political, cultural, economic, etc. contexts?
  - If not, how and why does this effect vary between countries?

  →Adapt theory

NEW INSIGHTS OF COUNTRY COMPARISONS

- Reasons for differences between countries
  - Are the *individuals different* in a specific way that relates to x and y?
  - Are there *aspects specific to the countries* which affect its residents?

(SOME) CHALLENGES OF QUANTITATIVE COMPARATIVE RESEARCH

CHALLENGES

- Are the variables we measure actually comparable across countries?
- Possible issues: Do the translations capture the same concept? Do the used concepts have the same meaning in different countries? …
- Technically, this is about *measurement equivalence*
- Measurement equivalence can be tested with structural equation modelling (not covered)

EXAMPLE: LEFT-RIGHT-ORIENTATION (THORISDOTTIR ET AL. 2007)

- “Left” and “right” considered core aspect of political identity
- However, left-right-scale differently understood in Western and Eastern (post-Sovjet) Europe
- *Resistance to change* correlates with right-wing conservatism in both regions
- *Acceptance of inequality* is associated with right-wing orientation in West Europe only
- *Openness to experience* related to left-wing orientation in Western Europe and right-wing orientation in Eastern Europe
- *Needs for security* associated with right-wing orientation in Western Europe and left-wing orientation in Eastern Europe

CHALLENGES

- *„There is a curious inconsistency in the way researchers interpret the results from […] replications […]. Failure to reproduce a finding in the same culture […] leads the investigator to question the reliability, validity and comparability of the research procedures […]. But failure to corroborate the same finding in a different culture often leads to claim of having discovered "cultural" differences."* (Finifter in Kohn 1987: p. 720)

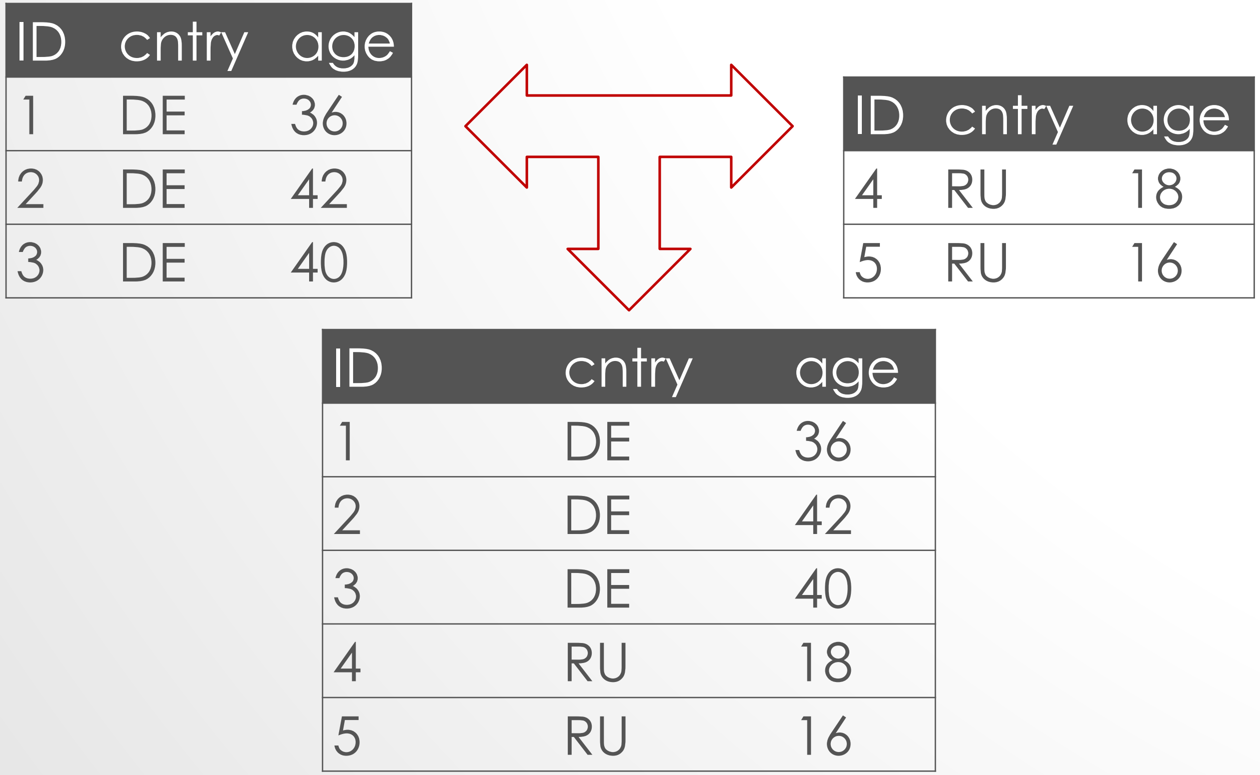
COMBINING DATA IN R

IMPORTING DIFFERENT FILE TYPES

- There are numerous ways to store data, each needs a different import function in R
- Stata's dta files: `read_dta()` (haven package)
- Excel xlsx files: `read_excel()` (readxl package)
- CSV files: `read.csv()` (base R)
- (Rdate files: `load()` (base R)
- And a lot more…

BINDING DATA

- Binding means combining rows (`rbind()`) or columns (`cbind()`) of two tables
- For example: Two countries



ID	cntry	age
1	DE	36
2	DE	42
3	DE	40

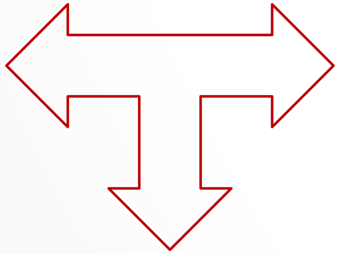
ID	cntry	age
4	RU	18
5	RU	16

ID	cntry	age
1	DE	36
2	DE	42
3	DE	40
4	RU	18
5	RU	16

BINDING ROWS

- Binding countries means adding *rows* → `rbind()`

ID	cntry	age
1	DE	36
2	DE	42
3	DE	40

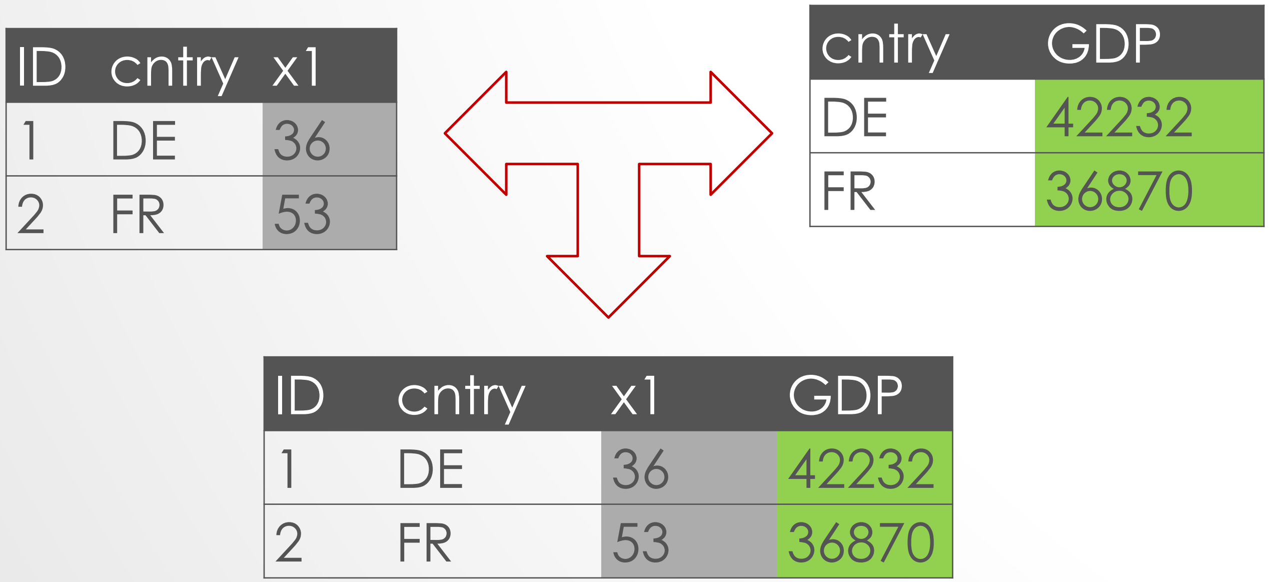


ID	cntry	age
4	RU	18
5	RU	16

ID	cntry	age
1	DE	36
2	DE	42
3	DE	40
4	RU	18
5	RU	16

BINDING COLUMNS

- Binding variables means adding *columns* → `cbind()`



ID	cntry	x1
1	DE	36
2	FR	53

cntry	GDP
DE	42232
FR	36870

ID	cntry	x1	GDP
1	DE	36	42232
2	FR	53	36870

BINDING DATA

- A drawback of `rbind()` is that it will only work when both tables have the same number of columns
- … and `cbind()` only when both data sets have the same number of rows
- Hence, `rbind()` will only work when both data sets have the exact same variables (as in the example)
- … and `cbind()` is useful when you have the exact same respondents in two datasets (hardly the case)

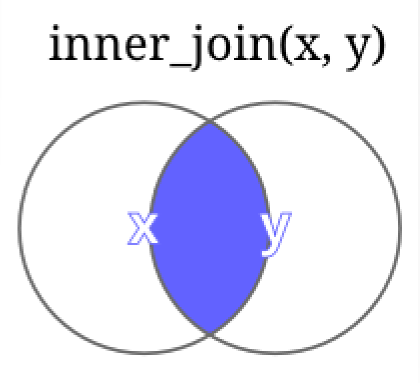
JOIN()

- The functions of the join() family of the dplyr package combine two (or more) tables / data sets
- Let us call table 1 master data. It is the one to which we add other data (e. g.: the ESS)
- Table 2 should be added to data set 1, let us call it using data (e. g.: additional country-level data)
- Finally, we need to know based on which column(s) we want to merge both data sets, let us call this the key variable
- The general syntax is: `join_type(masterData, usingData, by = keyVariable)`
- For example: `innerJoinDf <- inner_join(ESS, countrydata, by = “ID")`

`DPLYER`'S JOIN TYPES

- Inner Join (`inner_join()`): Combines observations of data 1 and 2 that are available in *both* data sets
- Left Join (`left_join()`): Adds data 2 to data 1
- Right Join (`right_join()`): Adds data 1 to data 2
- Full Join (`full_join()`): Combines observations of data 1 and 2 that are available in *either* data set
- Semi Join (`semi_join()`): Similar to `inner_join()`
- Anti Join (`anti_join()`): Only keeps observations of data 1 that are *not* available in data 2

INNER_JOIN()



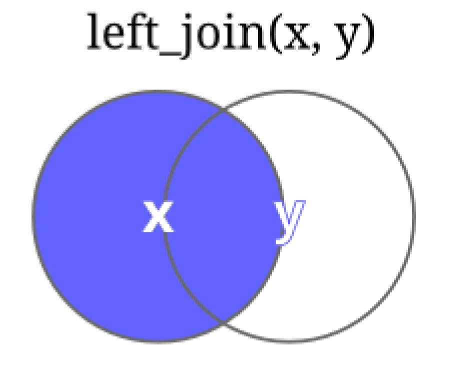
- Adds master data to using data based on key variable
- Only includes observations that exist in *both data*
- E. g.: `inner_join(master, using, by = “cntry”)`

cntry	x1
AT	1
BE	2
DE	3

cntry	x2
AT	A
BE	B
ES	C

cntry	x1	x2
AT	1	A
BE	2	B

LEFT_JOIN()



left_join(x, y)

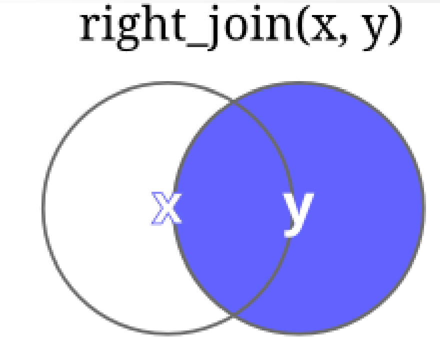
- Adds using data to master data based on key variable
- Only includes observations that are included in the *master data*
- Generates NA if observation missing in using data
- E. g.: `left_join(master, using, by = "cntry")`

cntry	x1
AT	1
BE	2
DE	3

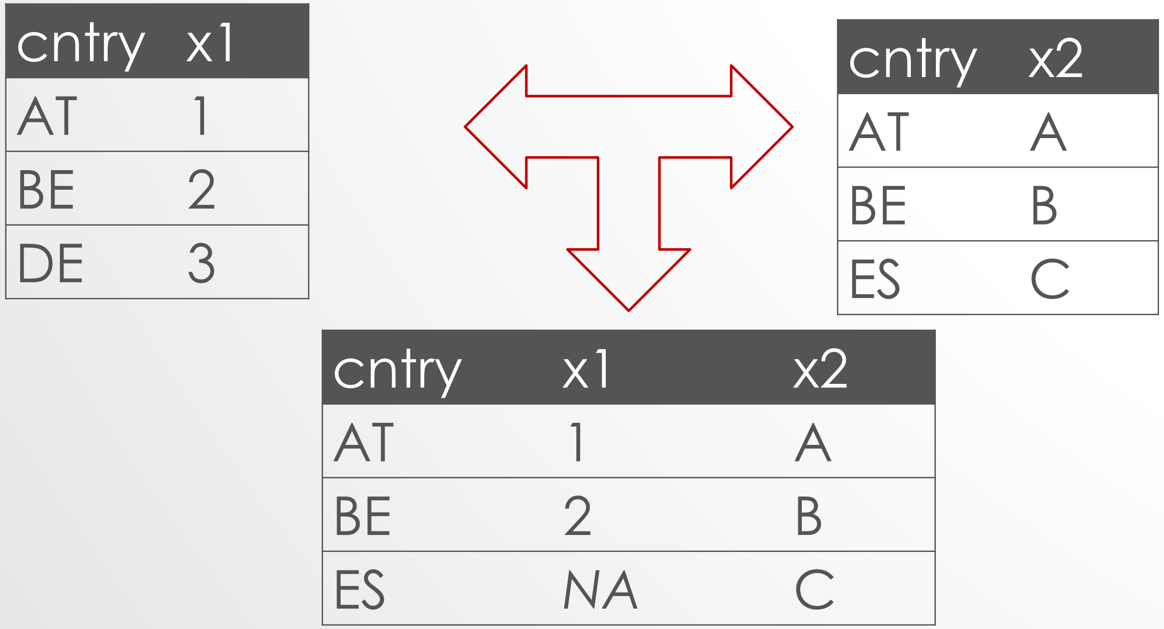
cntry	x2
AT	A
BE	B
ES	C

cntry	x1	x2
AT	1	A
BE	2	B
DE	3	*NA*

RIGHT_JOIN()



- Adds master data to using data based on key variable
- Only includes observations that are included in the *using data*
- Generates NA if observation missing in master data
- E. g.: `right_join(master, using, by = "cntry")`

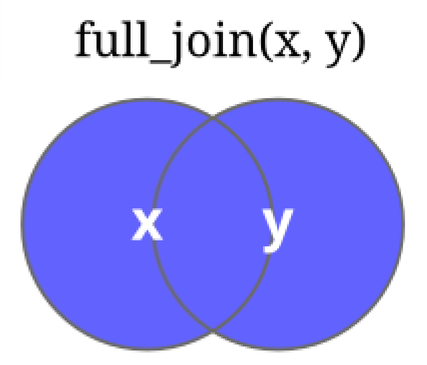


cntry	x1
AT	1
BE	2
DE	3

cntry	x2
AT	A
BE	B
ES	C

cntry	x1	x2
AT	1	A
BE	2	B
ES	*NA*	C

FULL_JOIN()



- Adds master data to using data based on key variable
- Includes all observations that exist in *either data*
- E. g.: `full_join(master, using, by = “cntry”)`

cntry	x1
AT	1
BE	2
DE	3

cntry	x2
AT	A
BE	B
ES	C

cntry	x1	x2
AT	1	A
BE	2	B
DE	3	*NA*
ES	*NA*	C

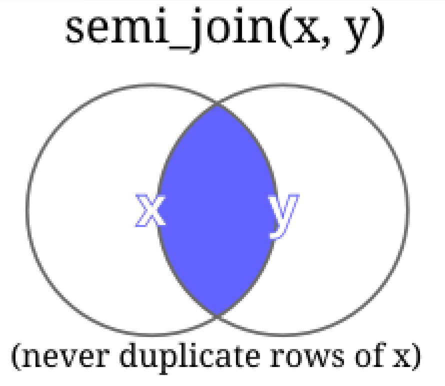
SEMI_JOIN()

- Adds master data to using data based on key variable
- Only includes observations that exist in *both data*
- … but only keeps variables that exist in the master data
- E. g.: `semi_join(master, using, by = “cntry”)`

cntry	x1
AT	1
BE	2
DE	3

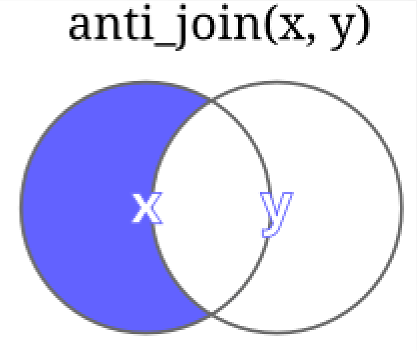
cntry	x2
AT	A
BE	B
ES	C

cntry	x1
AT	1
BE	2



ANTI_JOIN()

- Keeps observations of the master data that do not match the using data
- Generates NA if missing in master data
- E. g.: `anti_join(master, using, by = “cntry”)`



cntry	x1
AT	1
BE	2
DE	3

cntry	x2
AT	A
BE	B
ES	C

cntry	x1
DE	3

JOIN WITH MULTI-NATIONAL DATA

- The logic of each join function similarly applies for cross-national data, where we have several observations per key variable value (i. e.: multiple respondents per country)
- In this case, each respondent of the country in data 1 will get the country's value in data 2
- For example: `inner_join()`:

Respondent_ID	cntry	age
1	AT	34
2	AT	57
3	BE	35
4	BE	64
5	DE	24
6	DE	36

cntry	x2
AT	A
BE	B
ES	C

Respondent_ID	cntry	age	x2
1	AT	34	A
2	AT	57	A
3	BE	35	B
4	BE	64	B

…

MULTIPLE KEY VARIABLES

- Sometimes you may want to combine data sets based on multiple key variables (e. g. countries and years):

```
→ left_join(data1, data2, by=c(“cntry", "year"), 
match="all")
```

- Or you may want to combine more than two data sets:

```
→left_join(data1, data2, by = “cntry”) %>%
     left_join(., data3, by = "cntry")) %>%
     left_join(., data4, by = "cntry"))
```

- Of couse, don’t forget to assign these operations to an object

LITERATURE

- Kohn (1987): Cross-National Research as an Analytic Strategy, American Sociological Review, Vol. 52 (6), 713-731
- More on joining: http://rstudio-pubs-static.s3.amazonaws.com/227171_618ebdce0b9d44f3af65700e833593db.html