Advanced Survey Design

Peter Lugtig & Bella Struminskaya

August 28 - September 1, 2023

Department of methods and statistics – Utrecht University

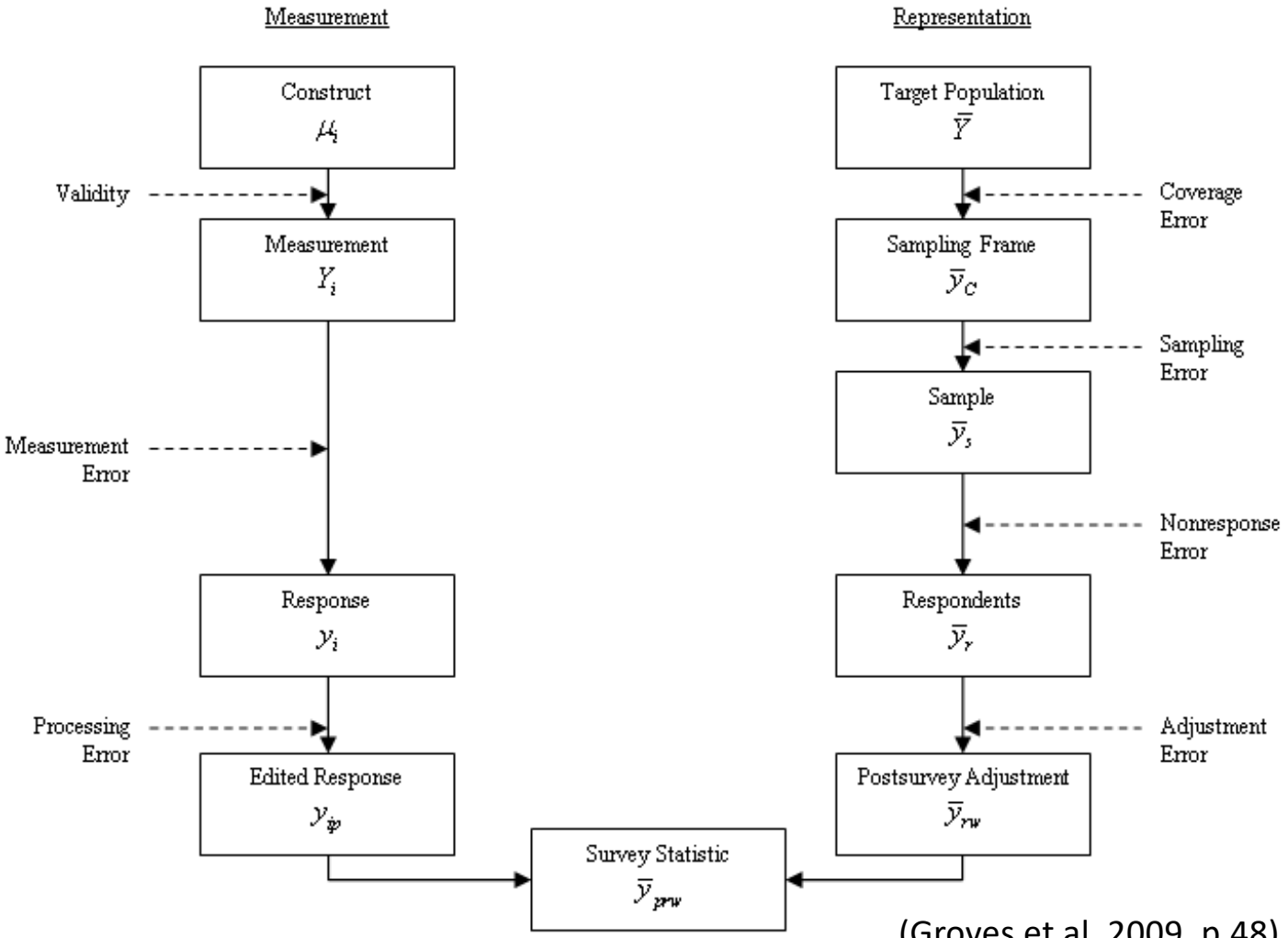
p.lugtig@uu.nl http://www.peterlugtig.com
b.struminskaya@uu.nl https://www.uu.nl/staff/BStruminskaya

Utrecht University

Data integration

Peter Lugtig – p.lugtig@uu.nl

TSE is focused on error (accuracy + precision)



(Groves et al. 2009, p.48)

Data quality framework of Biemer and Lyberg (2003)

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

Quality is "fitness for use"

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.

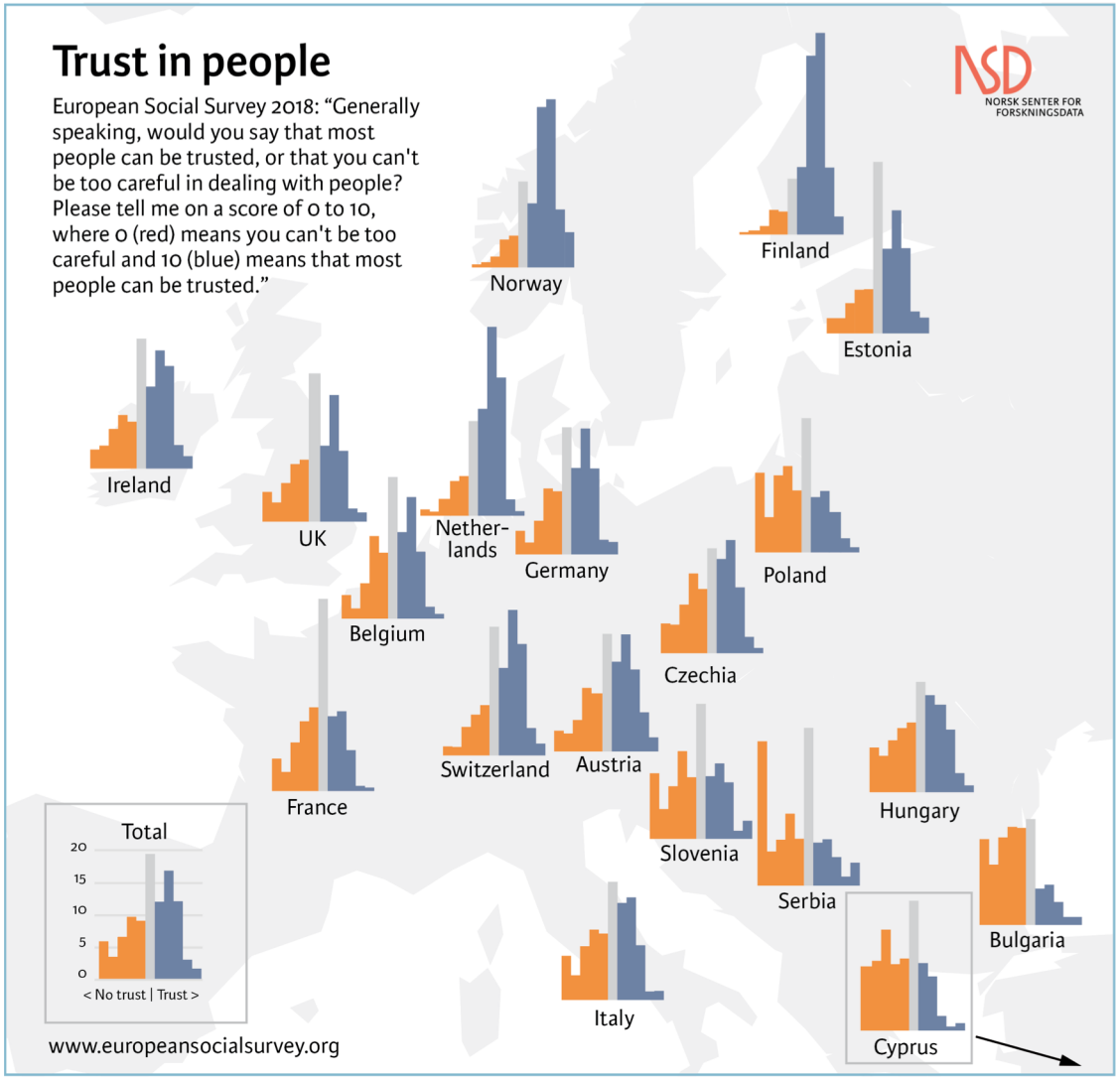
Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

The degree of confidence that users place in data products based on their image of the data provider.

Data quality framework

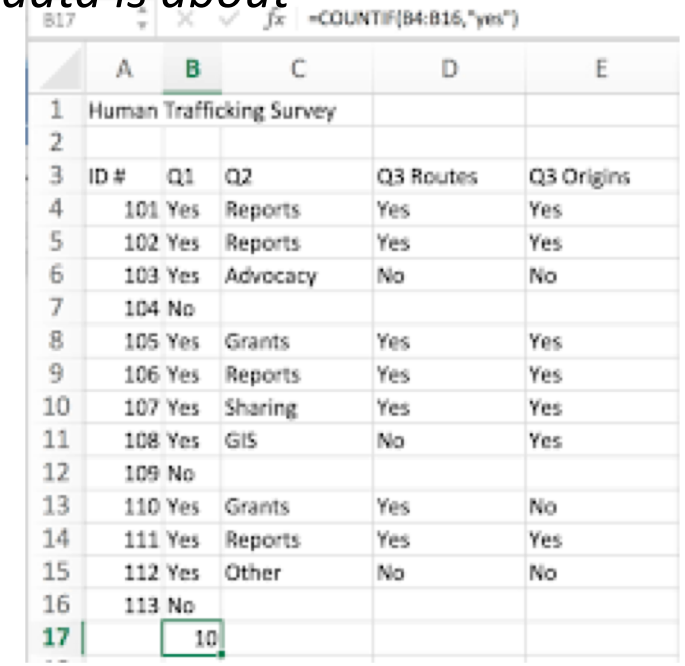
- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

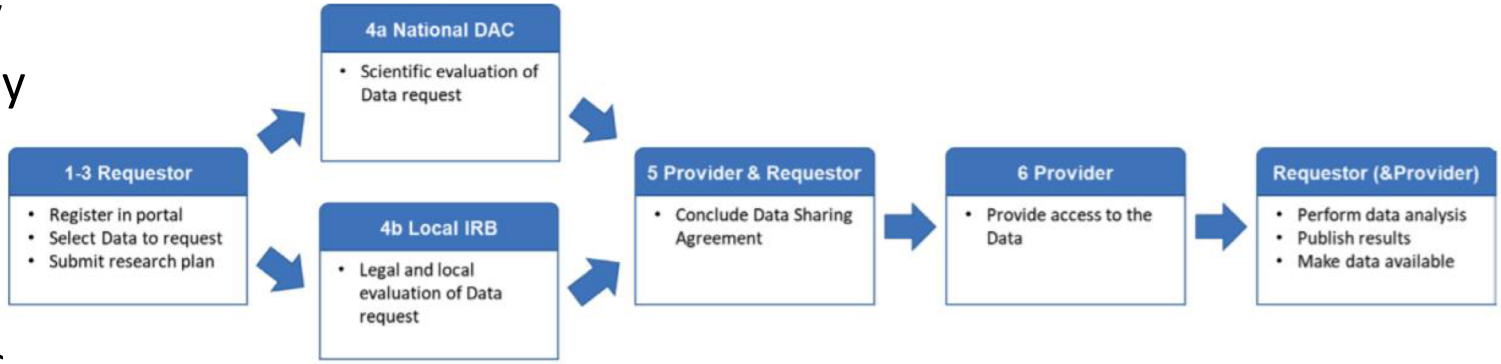
There is clear data documentation (metadata) so that we understand what data is about



	A	B	C	D	E
1	Human Trafficking Survey				
2					
3	ID #	Q1	Q2	Q3 Routes	Q3 Origins
4	101	Yes	Reports	Yes	Yes
5	102	Yes	Reports	Yes	Yes
6	103	Yes	Advocacy	No	No
7	104	No			
8	105	Yes	Grants	Yes	Yes
9	106	Yes	Reports	Yes	Yes
10	107	Yes	Sharing	Yes	Yes
11	108	Yes	GIS	No	Yes
12	109	No			
13	110	Yes	Grants	Yes	No
14	111	Yes	Reports	Yes	Yes
15	112	Yes	Other	No	No
16	113	No			
17		10			

Data quality framework

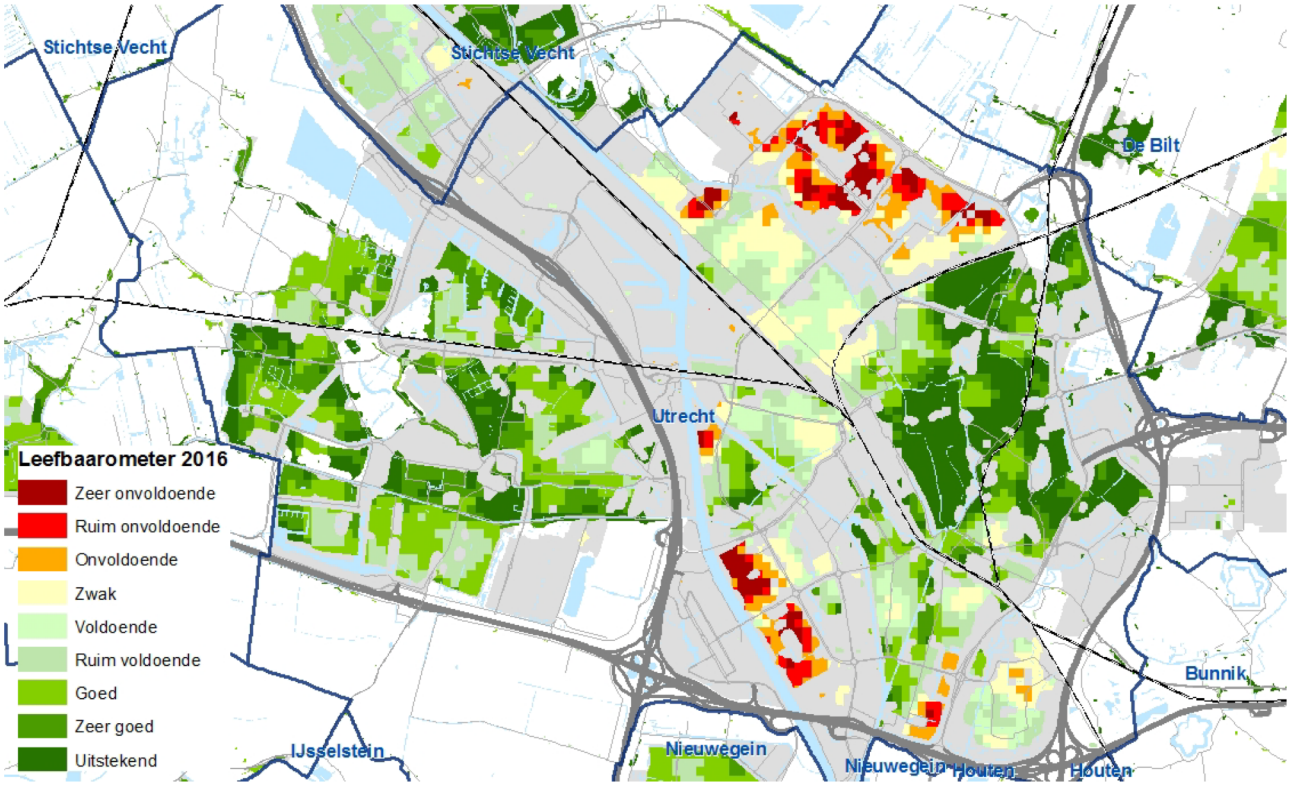
- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



The Netherlands HealthRI data access process

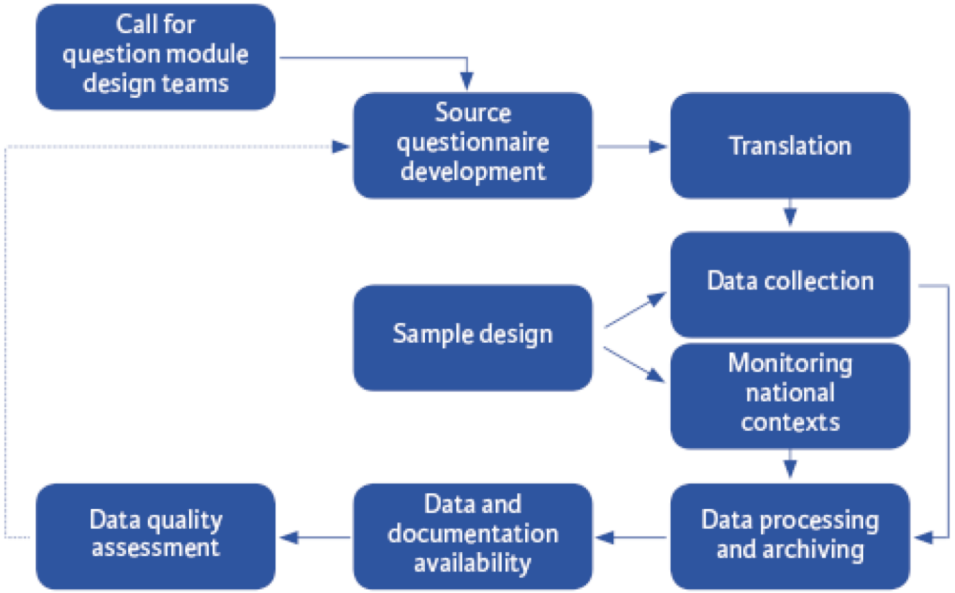
Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



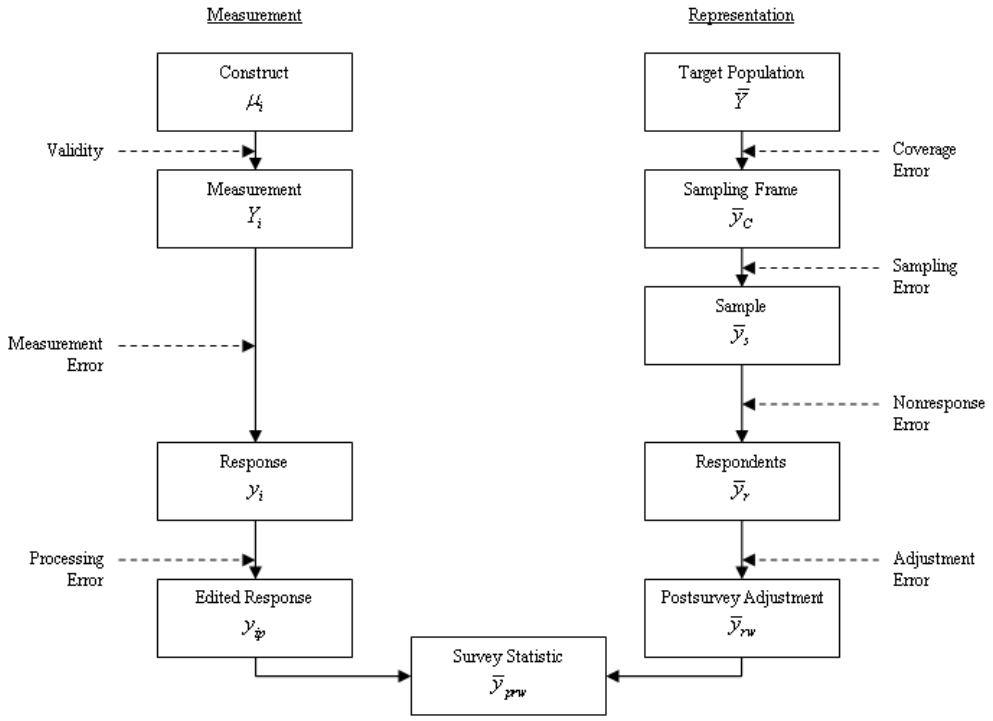
European Social Survey methodology overview

Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

Common definitions,classifications, and methodological standards (often over time)

Teen Internet Activities	
Do you ever…?	*Online Teens (n=886)*
Go to websites about movies, TV shows, music groups, or sports stars	81%
Get information about news and current events	77
Send or receive instant messages (IMs)	68
Watch video sharing site	57
Use an online social networking site like MySpace or Facebook	55
Get information about a college or university you are thinking of attending	55
Play computer or console games online	49
Buy things online, such as books, clothes, and music	38
Look for health, dieting, or physical fitness information	28
Download a podcast	19
Visit chatrooms	18

Source: Pew Internet & American Life Project Survey of Parents and Teens, October-November 2006. . Margin of error for teens is ±4%.

What has changed in the big data landscape?

- More sources
  - Administrative data
  - Survey data
  - Sensor data (phones, IoT)
  - Digital trace data
  - Organic (aka big) data
  - Non-prob surveys
- Each with it's own problems

- Not one methodology for how to do data integration
  - Approach is always statistic-specific

The idea of data integration

Work around a shortcoming of one source with another one

- Credibility
- Comparability
- Interpretability
- Accessibility
- Relevance
- Timeliness
- Completeness
- Accuracy
- Coherence

- Multi-source statistics (de Waal, van Delden, Scholtus, 2020; or Zhang, 2012)

Multi-source statistics

4 dimensions

1. Units, and population
2. Measurement (same, different)
3. Time dimension (same, different)
4. Level of aggregation (micro, or aggregation)

- Can be combined

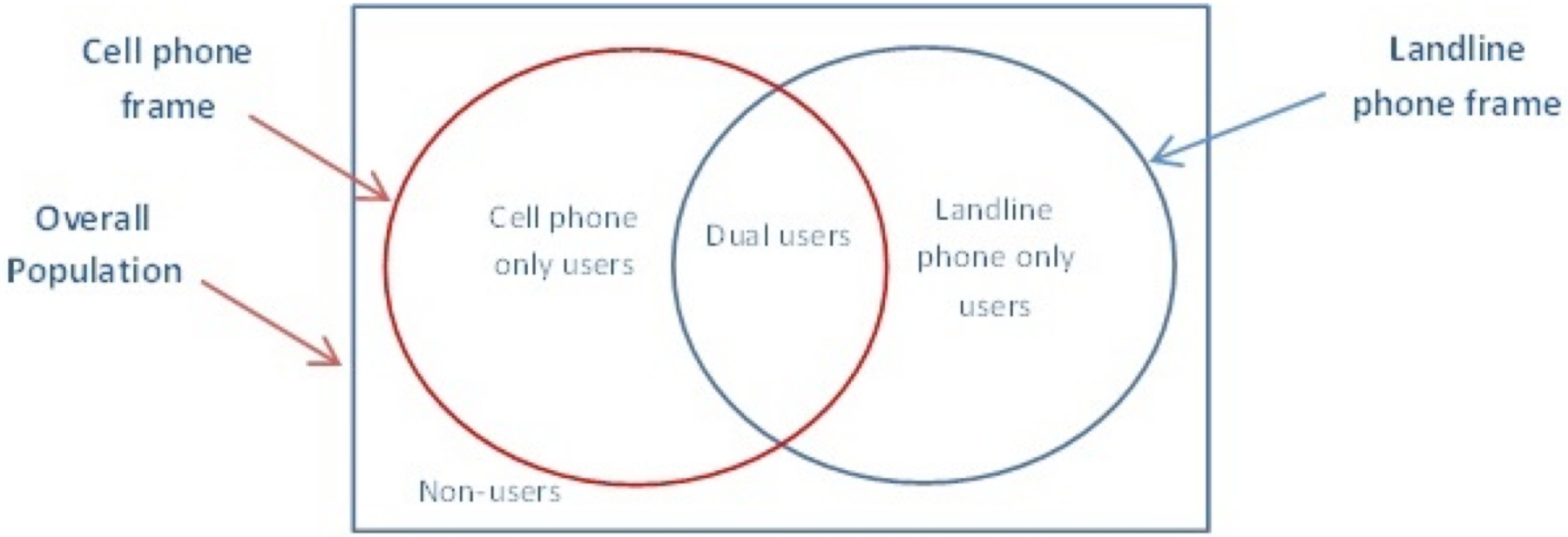
De Waal, T., van Delden, A., & Scholtus, S. (2020). Multi-source statistics: basic situations and methods. *International Statistical Review*, *88*(1), 203-228.

A lot of examples of data integration

Question with every application:

What quality dimension are we trying to improve?

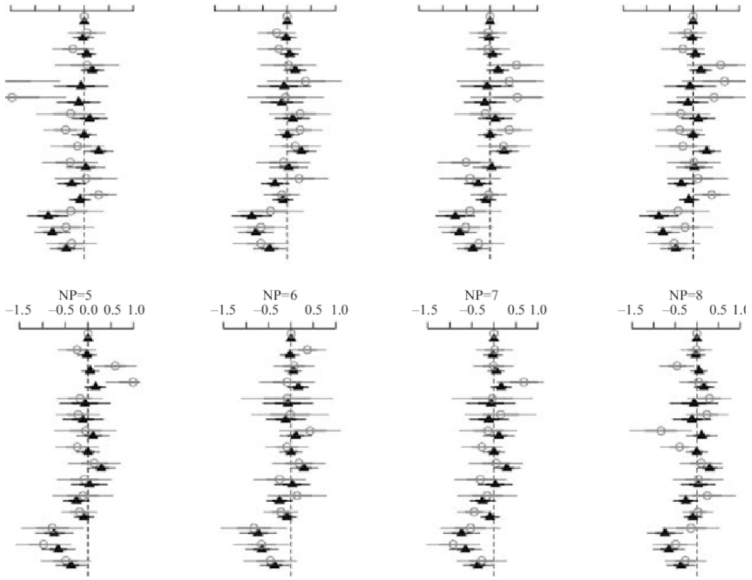
Dimension 1:
Multiple frames (to cover population)



What quality dimension are we trying to improve?

Dimension 1,2:
same measurements, different units

- Integrate smal probability based survey with
- Larger non-probability one
- Later guest lecture by Camilla Salvatore



What quality dimension are we trying to improve?

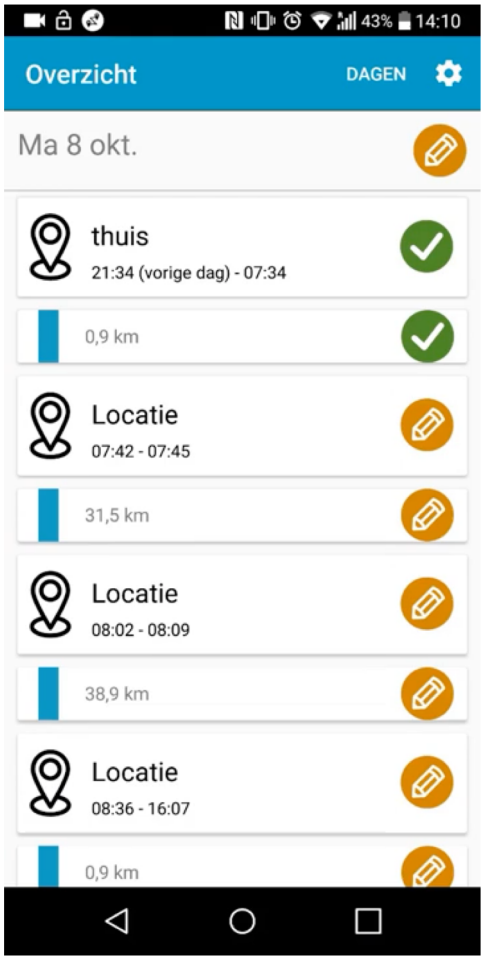
Dimensions 1,2:
Same units, different measurements

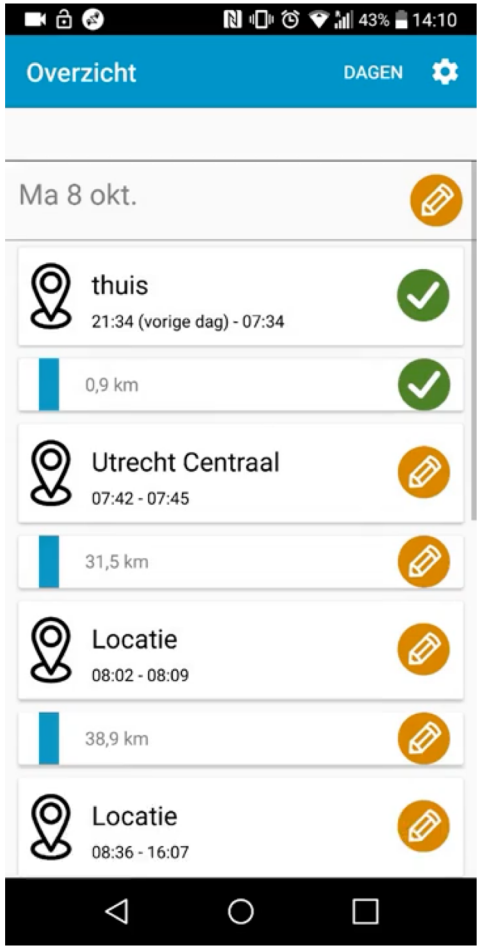
- Link 2 or more microdatasets of same individual
  - Data linkage (admin data)
  - Sampling frame information and survey data
  - Enriching surveys with administrative data

What quality dimension are we trying to improve?

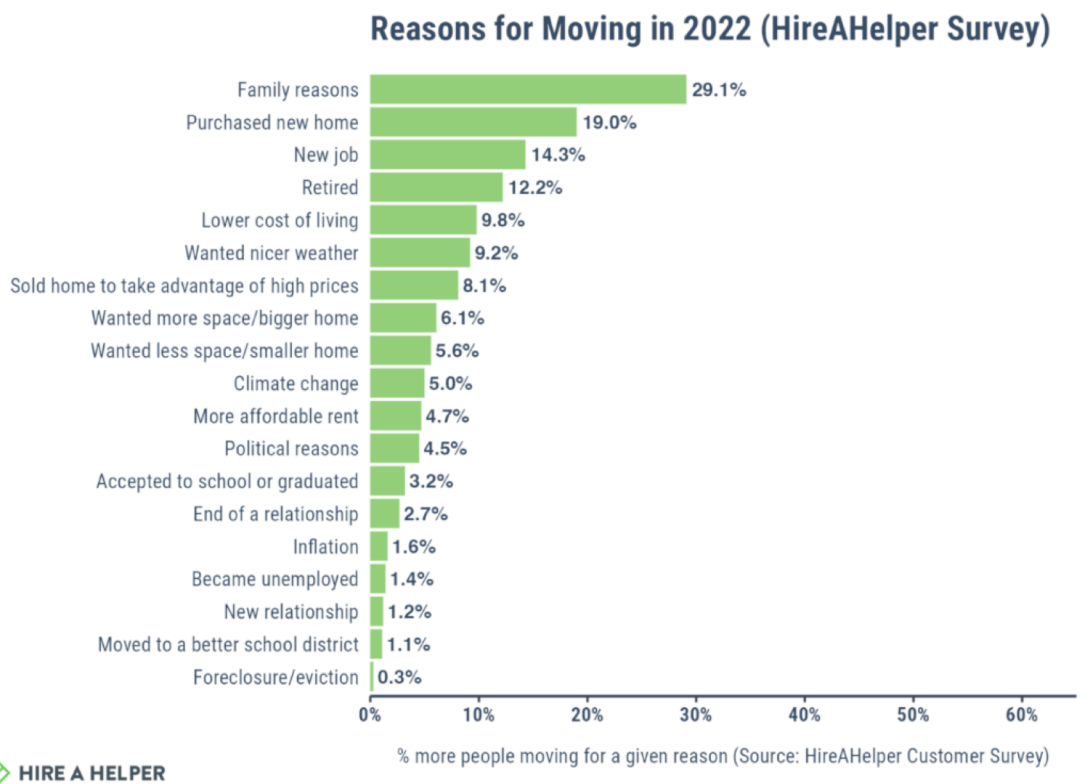
Dimensions 1,2:
Same units, different measurements

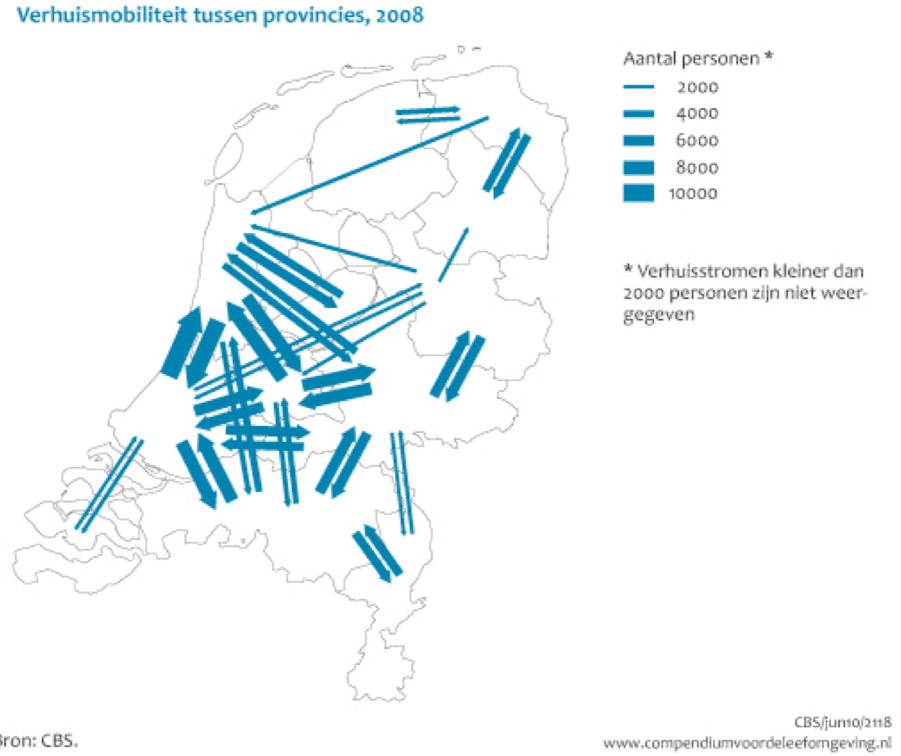
- Link 2 or more microdatasets of same individual
  - Designed big data





Dimension 2,3: measurement, time





Dimensions 2,4:
Same measurement, different aggregation

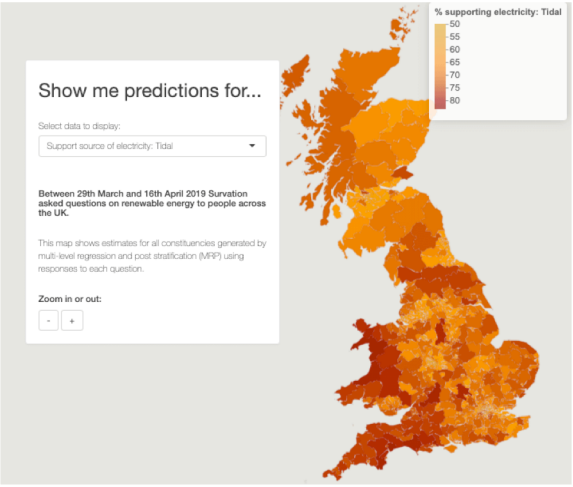
Use microdata + same statistic at aggregate levels

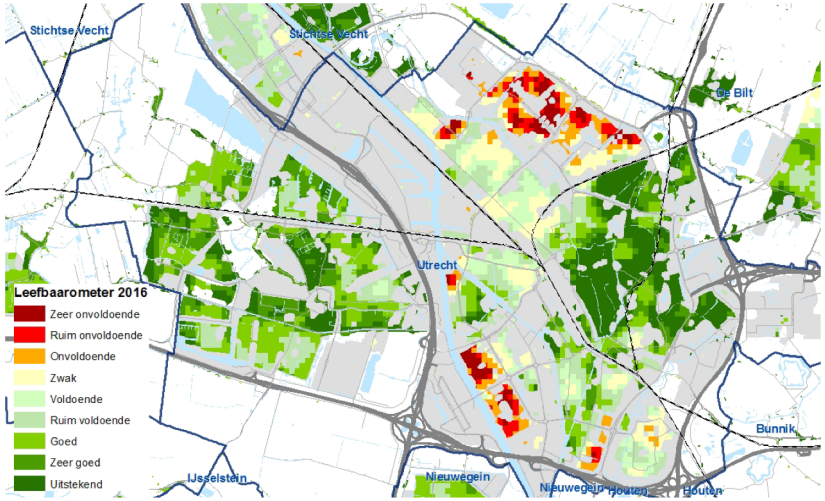
- Use population statistics for weighting/calibration
- Validate and asses accuracy of survey data
  - E.g. Sensitive questions

Dimensions 3,4:
different periods, different levels of aggregation

Use national survey data with local administrative data to make predictions at local level

- Small area estimation





Multilevel logistic regression models

$$\text{Logit}(p_{ij}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + u_j$$

Where β_0 is the 'intercept' and, β_1 to β_p are the coefficients of the p explanatory variables

References

- Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.
- De Waal, T., van Delden, A., & Scholtus, S. (2020). Multi-source statistics: basic situations and methods. *International Statistical Review*, *88*(1), 203-228.
- llic, G., Schouten, J.G., Lugtig, P., Mulder, J., Streefkerk, M., Kumar, and P. Höcük, S. (2022). Pictures instead of survey questions: An experimental investigation of the feasibility of using pictures in a housing survey. JRSS:A.
- McCool, D., Schouten, J.G. & Lugtig, P. (2021). An app-assisted travel survey in official statistics. Possibilities and challenges. *Journal of Official Statistics*
- van Delden, A., Scholtus, S., de Waal, T., & Csorba, I. (2023). Methods for estimating the quality of multisource statistics. *Advances in Business Statistics, Methods and Data Collection*, 781-804.
- Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120-147.