gesis Leibniz Institute for the Social Sciences

Linking survey data – state of the art and future directions

Sebastian Ziaja & Pascal Siegers

GESIS Meet-the-Experts Series

November 10 2022

Leibniz Association

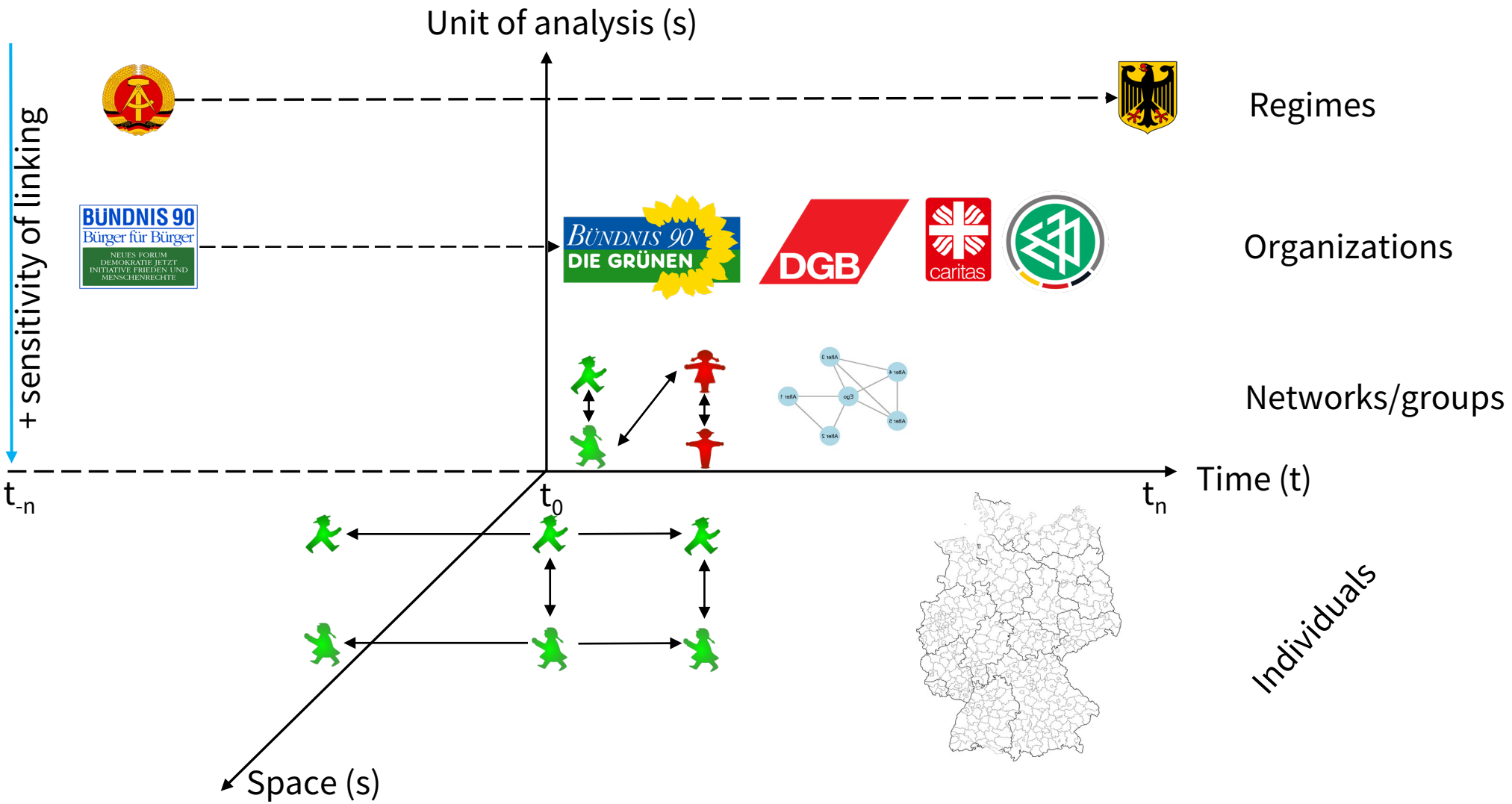
Background: the best of multiple worlds

- Survey data…
  - is necessary to understand social dynamics;
  - is not sufficient to understand social dynamics;
  - due biased recall, spurious perceptions , social desirability, lack of context and behaviour information, etc.
- Other data types can add crucial information (Skaaning 2020):
  - Geospatial data: physical context information
  - Digital trace data: actual online behaviour
  - Expert-coded data: consistent assessments of latent institutional traits
  - Official statistics: consistent information about society
- This is why many social scientists link survey data every day – but challenges and inefficiencies remain.

Definition of data linking

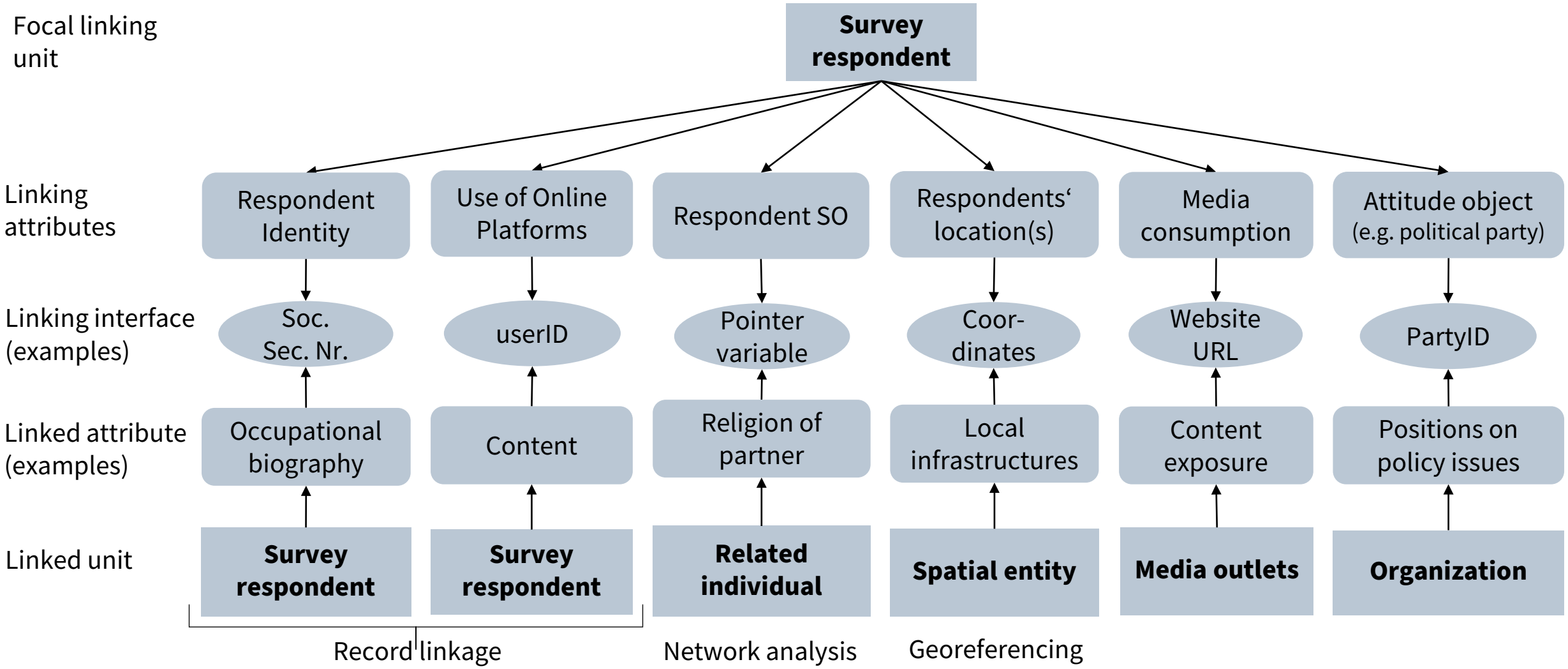
- Linking – in a broad sense – is the "process [of] combining data from multiple sources for joint analyses" (Beuthner et al. 2021).
- We focus on linking where the data linked to surveys stem from a different data generating process (i.e., they are non-survey data).
- Linked data can cover any analytical unit with an interface for linking to individuals or groups in surveys.
- Linking always relates to the attributes of the survey respondents.

Linking data space



gesis Leibniz Institute for the Social Sciences

Linking in research practice



Four challenges of survey-based linking

1. Obtaining consent for linking survey data
2. Identifying respondents' treatment status in experiments
3. Choosing the appropriate level of spatial aggregation
4. Aligning temporal units

1. Obtaining consent and cooperation for linkage

- Scholars use record linkage to add detailed attributes from administrative records to survey data (Antoni and Schnell 2017)
  - Recent example is linked data from GSOEP to public pension records (Lüthen et al., 2021)

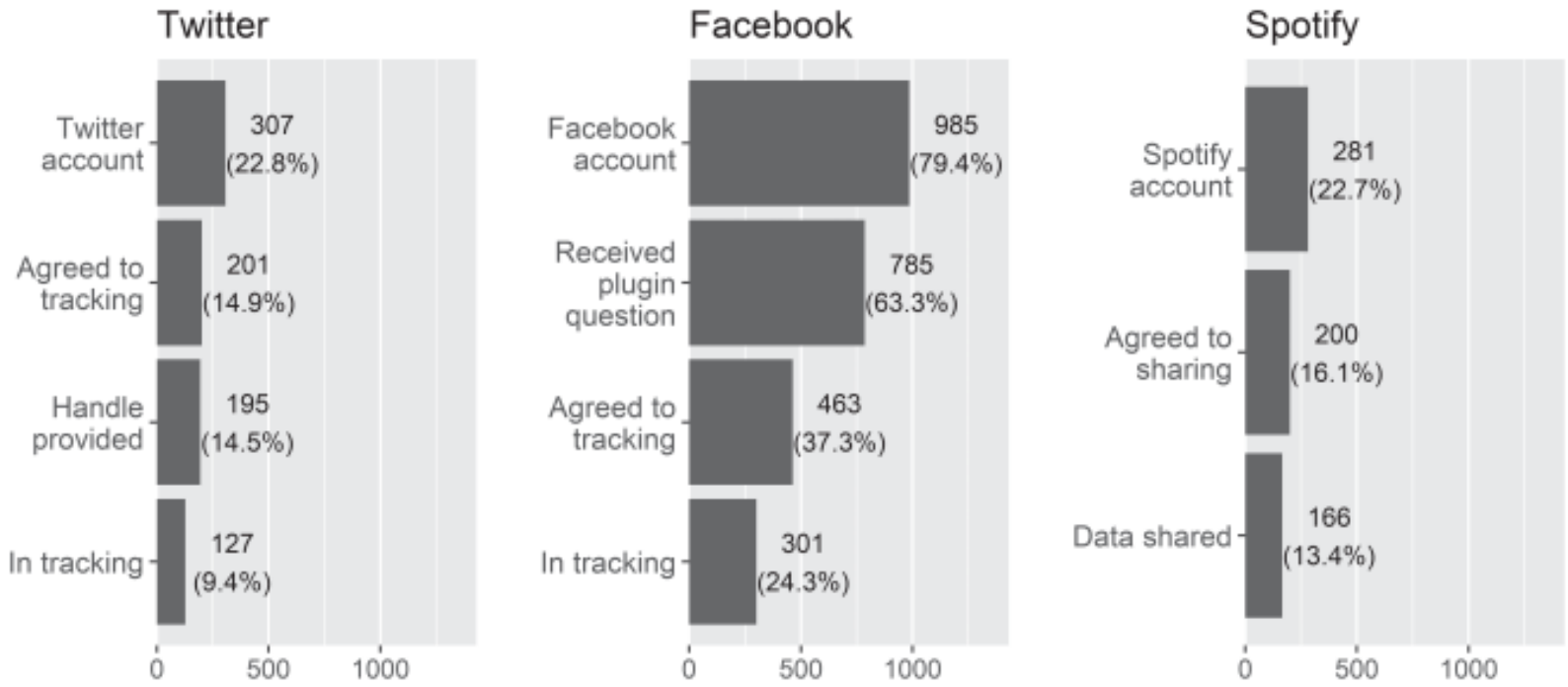
- Rise of digital behavioural data creates new opportunities for record linkage (Stier et al. 2020)
  - Profiles from Social Media platforms, data from mobile devices, user trackings etc.

- Any form of record linkage requires participants consent and (most often) cooperation (Breuer et al. 2021, Sloan et al. 2019)
  - revealing ID information, installing tools for data collection

1. Obtaining consent and cooperation for linkage

- Each step in linking leads to a loss of respondents
  1. Platform penetration
  2. Consent to linkage
  3. Cooperation to linkag
  4. Data ingest

- Each step is prone to selectivity bias

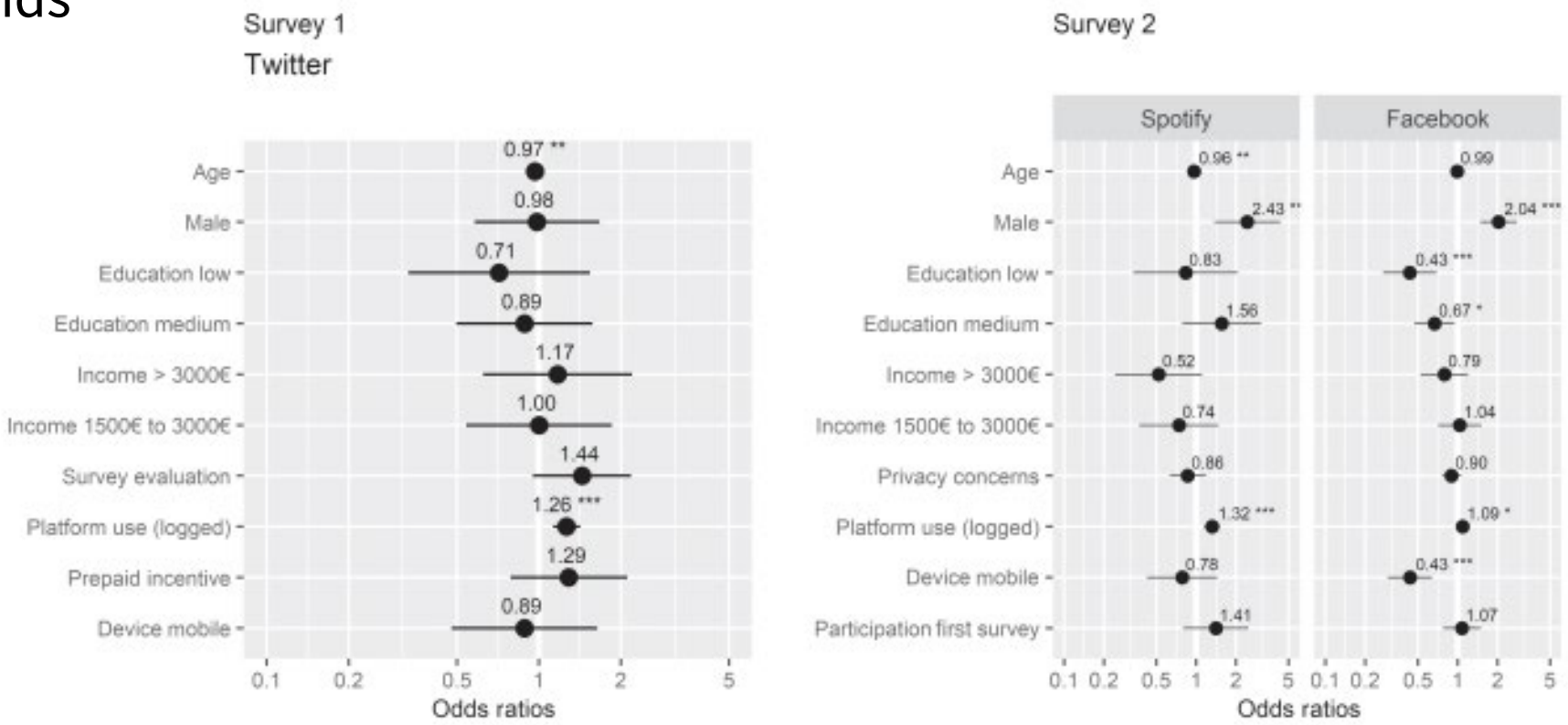


Source: Silber et al. 2022.

1. Obtaining consent and cooperation for linkage

- Consent to linkage depends on age, sex, intensity of platform use

- Linkage needs…
  - small but significant incentives
  - minimum burden for respondents
  - compliance with legal and ethical standards



Source: Silber et al. 2022.

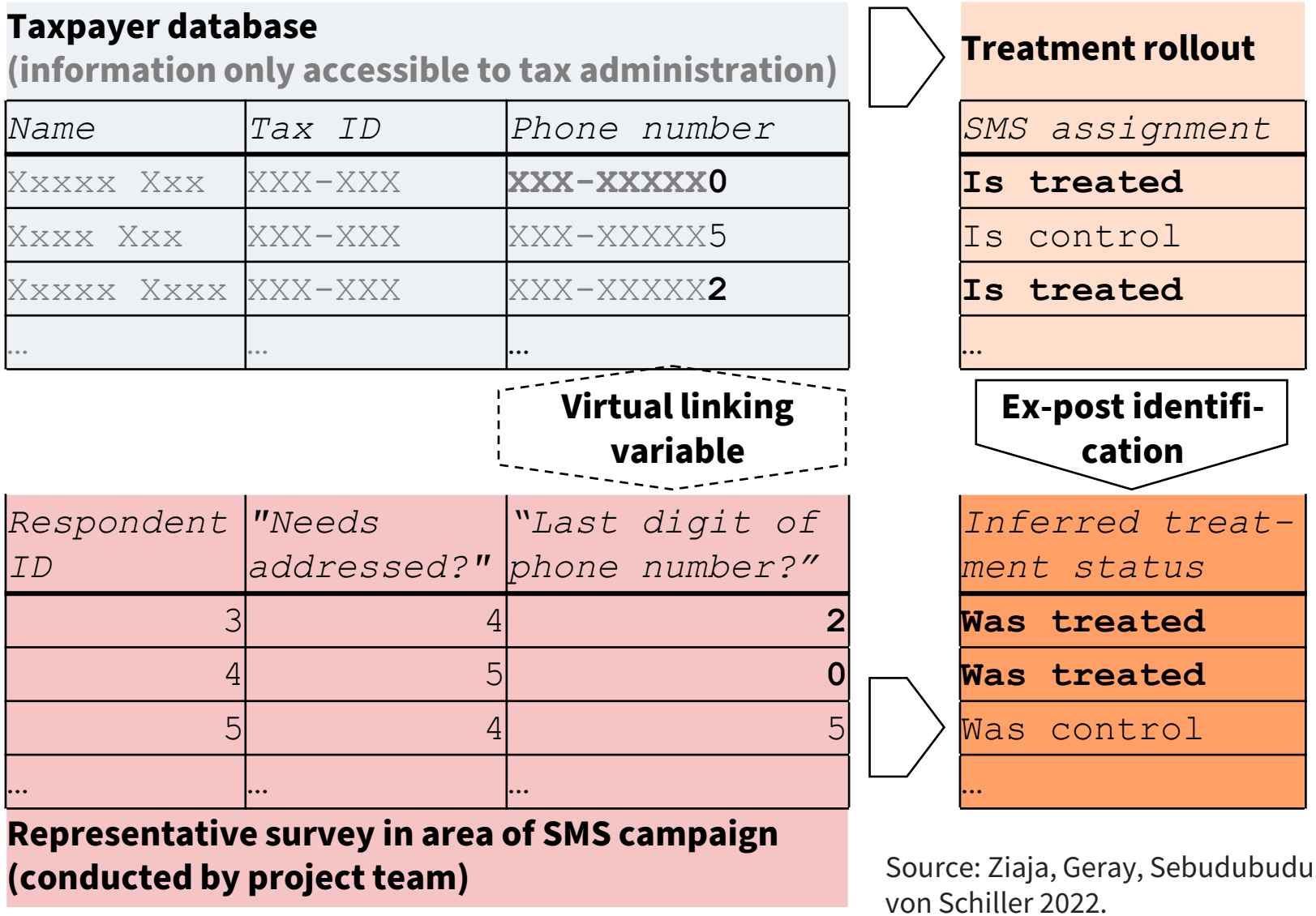
2. Identifying respondents' treatment status

- Experiments provide high levels of confidence in causality.
- In the social sciences, survey experiments prevail; real-world experiments are difficult to implement due to concerns over ethics and data privacy.
- Especially in cooperation with (non-)government organisations, treatments can be ethical, but data sharing about the treated remains legally barred.
- Self-reporting is unreliable for socially desirable traits (Munzert & Selb 2020; Hansen, Larsen & Gundersen 2021).
- *Solution:* Pseudo-randomised treatment, i. e., assigning with a deterministic (= fully replicable) rule that results in quasi-random distribution of treatment, allows for anonymous identification.

gesis Leibniz Institute for the Social Sciences

2. Identifying respondents' treatment status

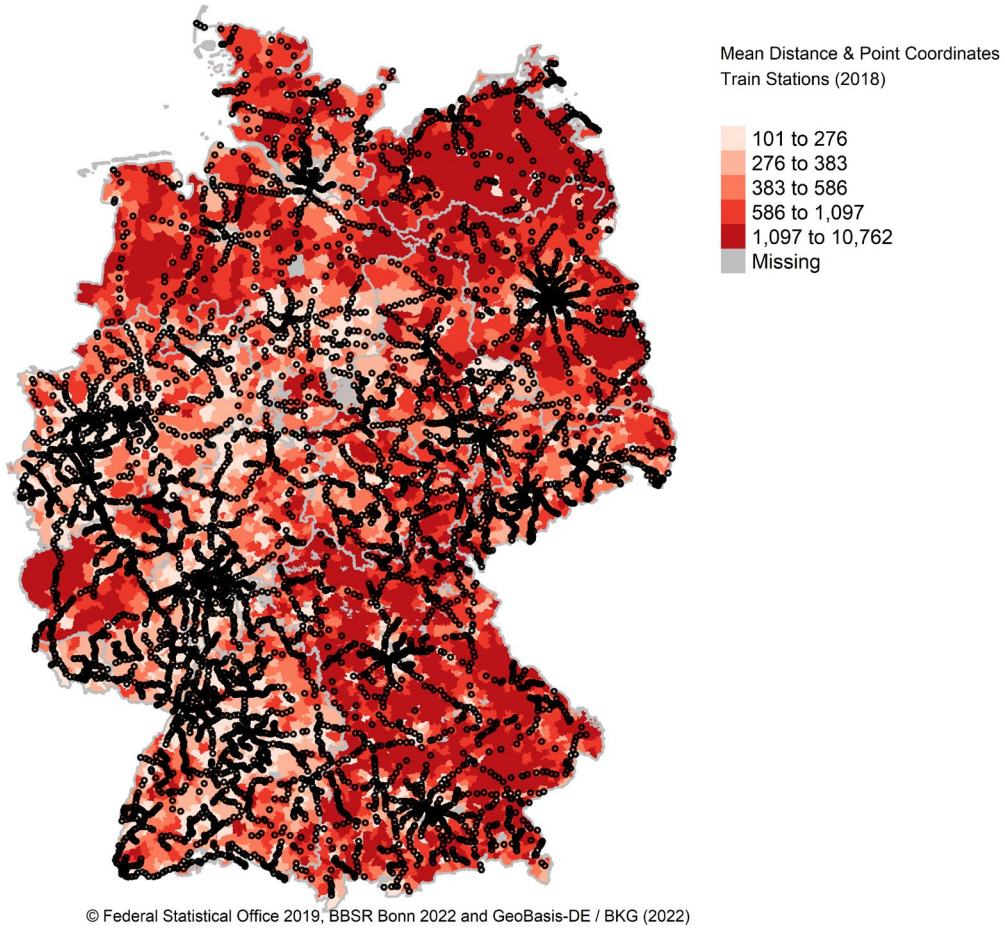
- Example of pseudo-randomised treatment assignment by last digit of phone number (0-4 = treated).
- SMS sent to 67,000 inhabitants of Gaborone, Botswana.
- Random survey of 2,048 inhabitants.
- Captured 470 potential recipients; 232 treated.



Source: Ziaja, Geray, Sebudubudu, von Schiller 2022.

3. Choosing the appropriate level of spatial aggregation

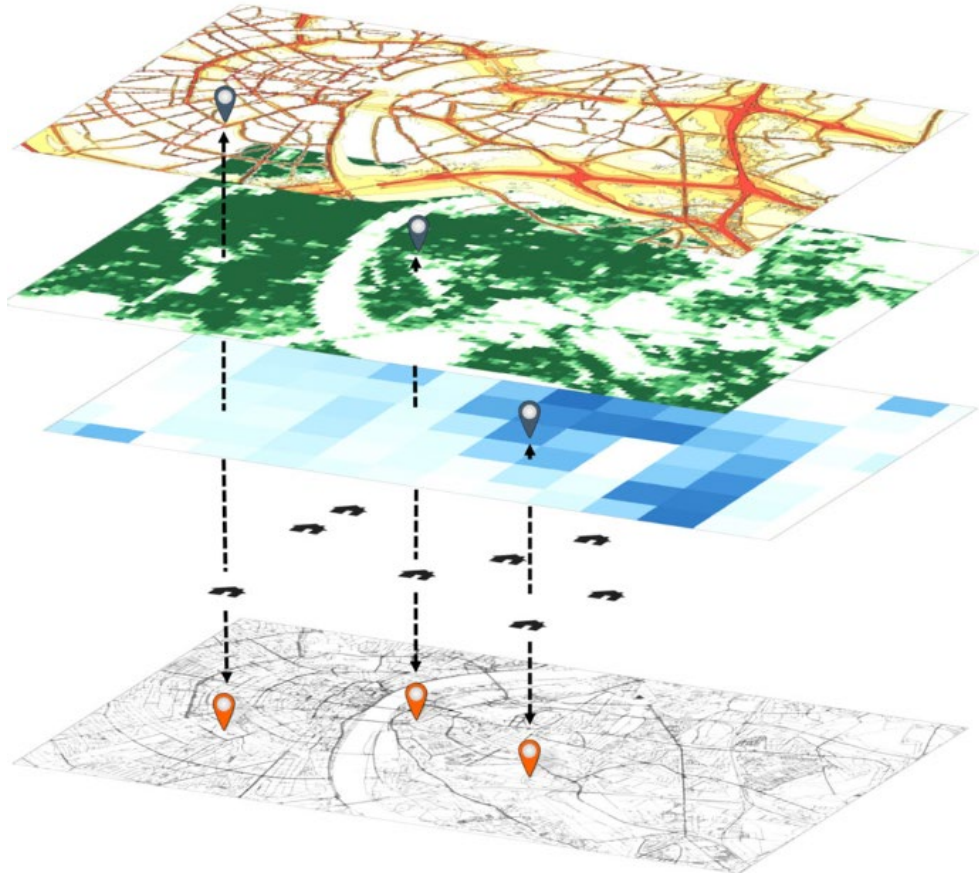
- Using spatial data to model individuals‘ living conditions
- Indirect spatial references are transformed into direct spatial references
- Projection into a geographic coordinate space for linking to spatial attributes
  - Space usage, infrastructure, population composition



Source: Stroppe 2022.

3. Choosing the appropriate level of spatial aggregation

- Modifiable Area Unit Problem (MAUP)
  - Depending on how researchers choose their spatial units for analysis, the results of the analysis change because spatial units are arbitrary
  - Scale Effect: variation of statistical results when spatial units are aggregated into larger units
  - Zoning Effect: variation of statistical results depending on different methods for aggregating the spatial units

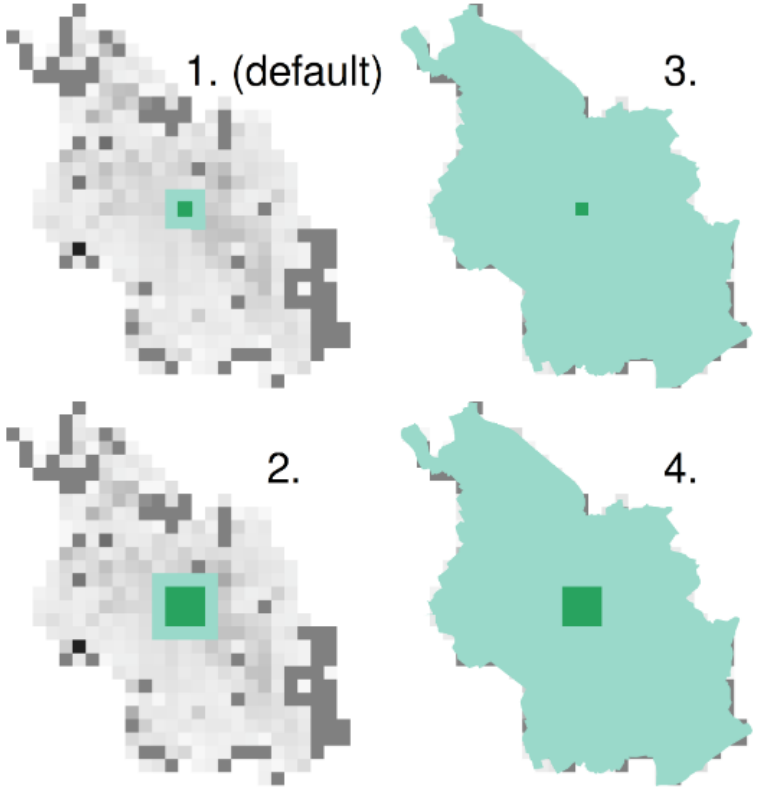


Source: Siegers et al. 2019.

3. Choosing the appropriate level of spatial aggregation

- Jünger (2019) uses different ways of modelling the “halo” of neighborhoods

- Share of migrant population living around the residence of survey respondents



Data Sources: Statistical Offices of the Federation and the Länder (2016) and Federal Agency for Cartography and Geodesy (2018)

Source: Jünger 2019.

3. Choosing the appropriate level of spatial aggregation

- Hillmert et al. (2017) show that the relationship between migration and income depends on where the migrants live
- Effect only visible if spatial dependencies are modelled

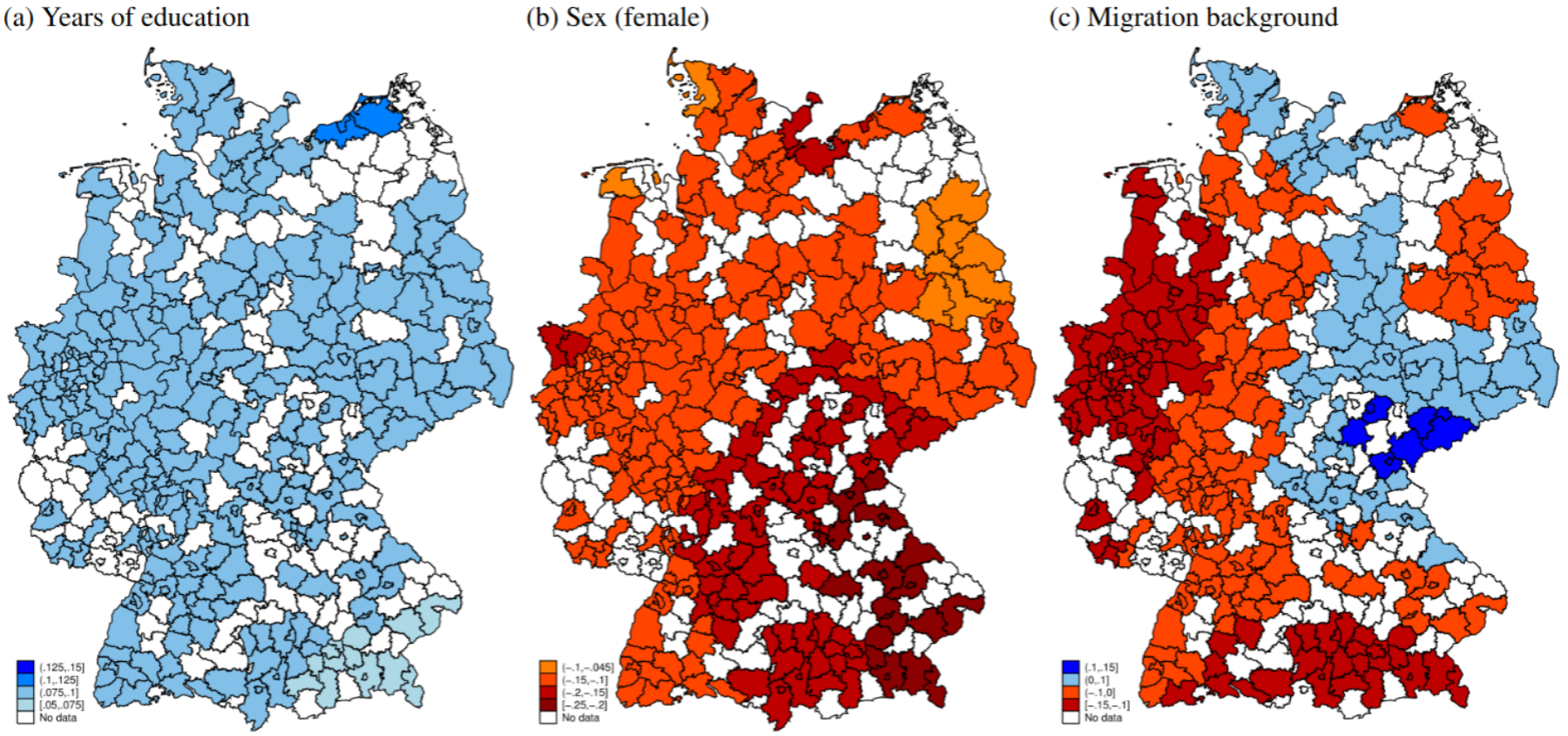


Figure 7. Geographically varying effects of education (a), sex (b) and migration (c) background on individual earnings: Results from GWR. Regional classification: admin. districts (NUTS-3). Data: GSOEP; BKG, 2016 (shapefile), own calculations.

Source: Hillmert et al. 2017.

gesis Leibniz Institute for the Social Sciences

4. Aligning temporal units

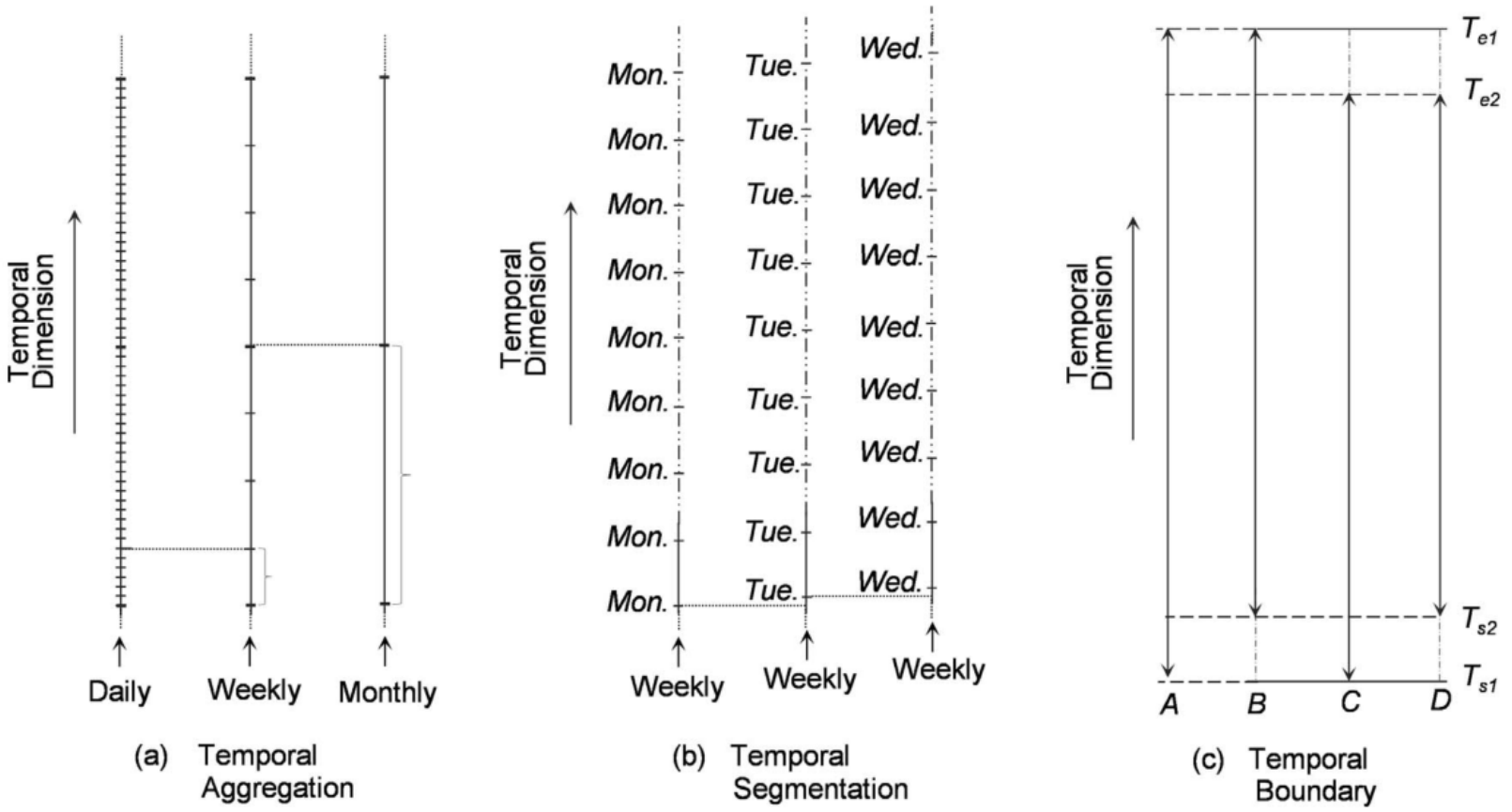


Figure 1. Modifiable Temporal Unit Problem (MTUP) (a) Temporal aggregation (b) Temporal Segmentation (c) Temporal boundary.
doi:10.1371/journal.pone.0100465.g001

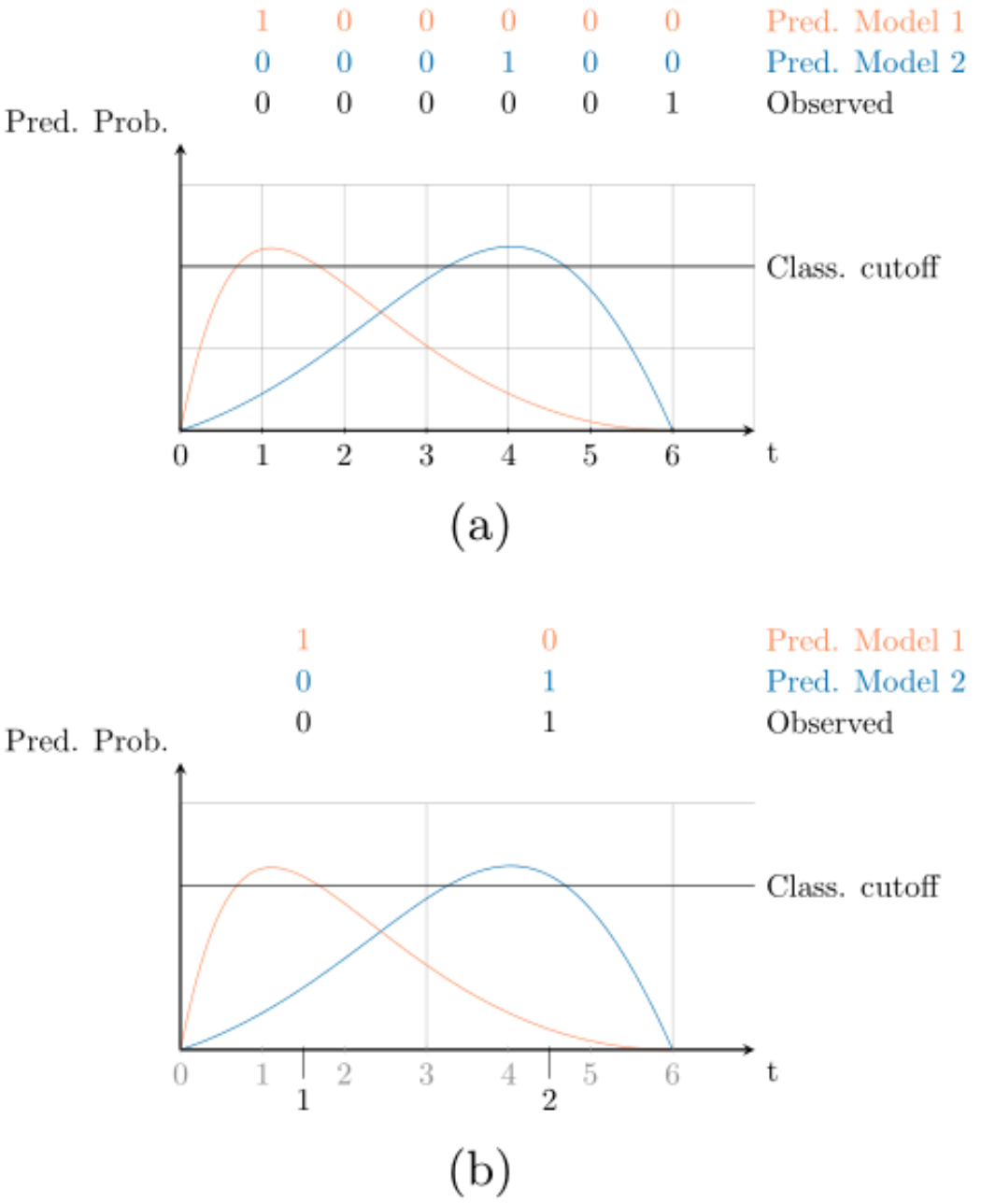
Source: Cheng and Adepeju 2014.

- There is no natural temporal unit for survey data – nor for other data types.
- The appropriate unit depends – surprise – on the research question and the underlying theory.
- Not only length of the temporal interval matters, but also where it starts and the overall boundaries of the time period under investigation.

gesis Leibniz Institute for the Social Sciences

4. Aligning temporal units

- The temporal unit matters in theory.
- Temporal residual problem [panel (a)]:
  - Small units may lead to low predictive power of models, as time between prediction and outcome is usually not considered in evaluating models.
- Modifiable temporal unit problem (MTUP) [panel (b)]:
  - Adjusting the temporal unit can change predictive power.
- But is model (a) better than model (b)?

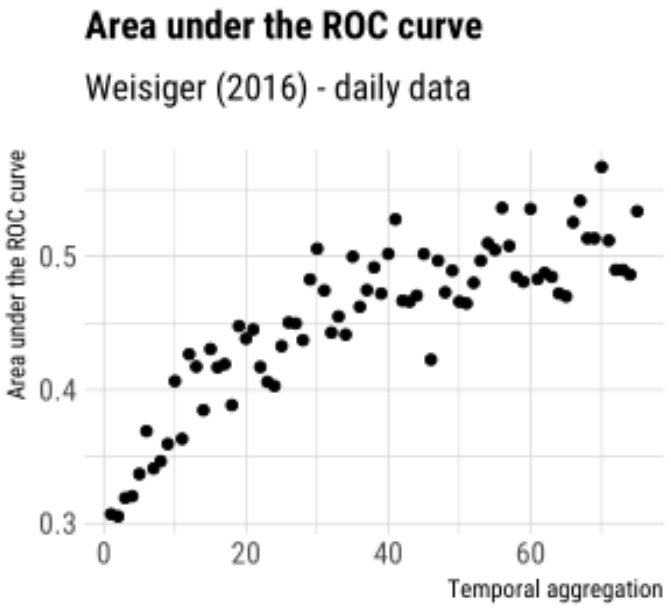


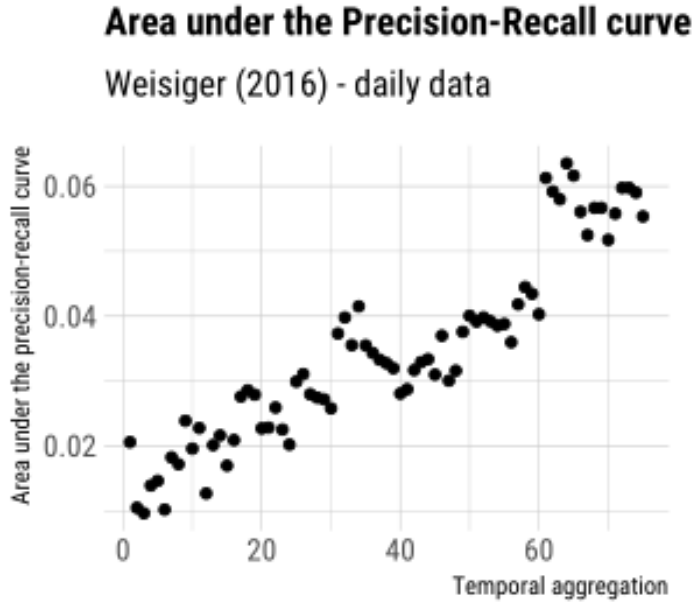
Source: Çiflikli, Metternich, Weber & Rickard 2020: 2.

gesis Leibniz Institute for the Social Sciences

4. Aligning temporal units

- The temporal unit matters in practice!
- Example: defection from wartime coalitions.
- Precision of estimates can increase both ways (Bae et al. 2021):
  - Larger temporal units -> more variation between observations -> smaller SEs
  - Smaller temporal units -> inflation of observations -> smaller SEs
- *Solution:* Need to carefully theorise time and know temporal properties of all involved data sources.





Source: Çiflikli, Metternich, Weber & Rickard 2020: 10.

Future outlook for survey linking

- Linking potential for survey data has grown exponentially with the digital revolution.
- Augmenting survey data provides opportunities, but potential pitfalls have grown with opportunities.
- Carefully theorising space, time, and actors before engaging in data linking is imperative.
- Knowing the data generating processes of all data sources in detail helps make right linking decisions.
- Many data linking applications are scalable; community-built open source solutions are the way towards generating synergies.

Upcoming Meet-the-experts programme

- 08.12.2022, Dr. Marlene Mauk: Linking surveys with electoral integrity assessments to explain political trust
- 11.01.2023, Anne-Kathrin Stroppe: The Geocoded German Longitudinal Election Study (GLES): Analyzing place-based effects on the 2021 German Federal Election
- 09.02.2023, Dr. Boris Heizmann: Meet the Eurobarometer
- 09.03.2023, Dr. Sonja Schulz: Meet the ALLBUS cumulation (in German)
- 13.04.2023, Dr. Stefan Jünger: bkggeocoder: a geocoding tool for survey data

Pascal Siegers | @pascalsiegers1 | pascal.siegers@gesis.org
Sebastian Ziaja | @sziaja | sebastian.ziaja@gesis.org

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

gesis Leibniz Institute for the Social Sciences

References

Antoni, M., & Schnell, R. (2019). The past, present and future of the German Record Linkage Center (GRLC). Jahrbücher für Nationalökonomie und Statistik, 239(2), 319-331.

Bae, B., Lee, C., Pak, T.-Y., & Lee, S. (2021). Identifying Temporal Aggregation Effect on Crash-Frequency Modeling. *Sustainability*, 13(11), Article 11. https://doi.org/10.3390/su13116214

Çiflikli, G., Metternich, N. W., Weber, S., & Rickard, K. (2020). Taking time seriously when evaluating predictions in Binary-Time-Series-Cross-Section-Data. *SocArXiv*. https://doi.org/10.31235/osf.io/tvshu

Breuer, J., Al Baghal, T., Sloan, L., Bishop, L., Kondyli, D., & Linardis, A. (2021). Informed consent for linking survey and social media data - Differences between platforms and data types. IASSIST Quarterly, 45(1). https://doi.org/10.29173/iq988

Hansen, P. G., Larsen, E. G., & Gundersen, C. D. (2022). Reporting on one's behavior: A survey experiment on the nonvalidity of self-reported COVID-19 hygiene-relevant routine behaviors. *Behavioural Public Policy*, 6(1), 34–51. https://doi.org/10.1017/bpp.2021.13

HILLMERT, Steffen; HARTUNG, Andreas; WEßLING, Katarina (2017): Dealing with Space and Place in Standard Survey Data. Survey Research Methods [S.l.] 11(3), 267-287. https://doi.org/10.18148/srm/2017.v11i3.6729

Jünger, S. (2019). Using georeferenced data in social science survey research: The method of spatial linking and its application with the german general social survey and the GESIS panel (Vol. 24, p. 208).

Lüthen, H., Schröder, C., Grabka, M. M., Goebel, J., Mika, T., Brüggmann, D., ... & Penz, H. (2022). SOEP-RV: Linking German Socio-Economic Panel Data to Pension Records. Jahrbücher für Nationalökonomie und Statistik, 242(2), 291-307.

Munzert, S., & Selb, P. (2020). Can we directly survey adherence to non-pharmaceutical interventions? *Survey Research Methods*, 14(2). https://doi.org/10.18148/srm/2020.v14i2.7759

Siegers, Pascal, Stefan Müller, and Julia Klinger. 2019. "Regionalisierung durch Georeferenzierung in der Sozialforschung." In *Regionale Standards: Ausgabe 2019*, edited by Arbeitsgruppe Regionale Standards, GESIS-Schriftenreihe 23, 78-93. Köln: GESIS. doi: https://doi.org/10.21241/ssoar.62343.

Silber,H.,Breuer,J.,Beuthner,C.,Gummer,T.,Keusch,F., Siegers,P.etal.(2022)Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*,1–21.Availablefrom:https://doi.org/10.1111/rssa.12954

Skaaning, S.-E. (2018). Different Types of Data and the Validity of Democracy Measures. *Politics and Governance*, 6(1), 105–116. https://doi.org/10.17645/pag.v6i1.1183

Sloan L, Jessop C, Al Baghal T, Williams M. Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics.* 2020;15(1-2):63-76. https://doi.org/10.1177/1556264619853447

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5), 503-516.

Stroppe, A. 2022. Left Behind in Public Services Wasteland? On the Accessibility of Public Services and Political Trust. Working Paper.

Weisiger, A. (2016). Exiting the coalition: When do states abandon coalition partners during war? In *International Studies Quarterly*, 60(4):753–765.

Ziaja, Sebastian, Markus Geray, David Sebudubudu, and Armin von Schiller (2022): E-government as a state-building tool? A field experiment from Botswana on perceptions of government responsiveness, *under review*