INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

UNIVERSITY OF MANNHEIM

The IAB-SMART App

Measurement quality in mobile geolocation sensor data

ESRA 2019

Sebastian Bähr

Georg-Christoph Haas

Florian Keusch

Frauke Kreuter

Mark Trappmann

Background

- Increasing prevalence of smartphones (Pew Research Center 2018)
- Sensors are ubiquitous
- Innovative data source for the social sciences (e.g., Sugie 2016)
- New type of data: Passive sensor data
  - generated without any participation or action from the subject (Onnela and Rauch 2016)
  - unobtrusive, naturalistic observational records
    that reduce the likelihood that participants will behave reactively (Harari et al. 2017)
- Little knowledge about data quality

The “passive” fantasy (Couper 2019)

- Smartphone sensor are data selective
  - General Population > (Android) Smartphone ownership > participation in study > willingness to share passive data > successful data collection
- Sensor measurement ≠ targeted behavior
  - Devices might be turned off or not carried with the (targeted) participant
  - Sensors offer only a limited perspective on behavior (Harari et al. 2017)
  - Interpretation (by the researcher or the participant) is needed
- Passive data are not objective (i.e., error-free)
  - Research app, devices, operating systems, third party apps, and participants can interfere with measurement
- Passive data are noisy data
  - High frequency measurement needs pattern recognition (data preparation)

Couper, M. (2019) “Mobile Data Collection: A Survey Researcher’s Perspective”, Keynote at Mass-Workshop , March, 04th 2019, Mannheim University

IAB-SMART App

An app, that …

… launches surveys.

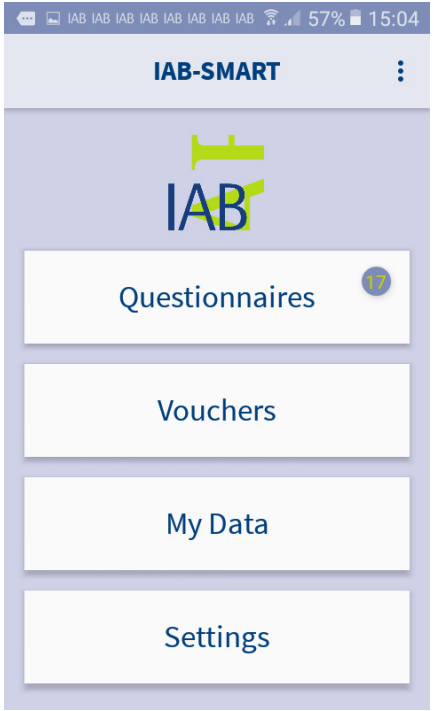
… passively collects smartphone data

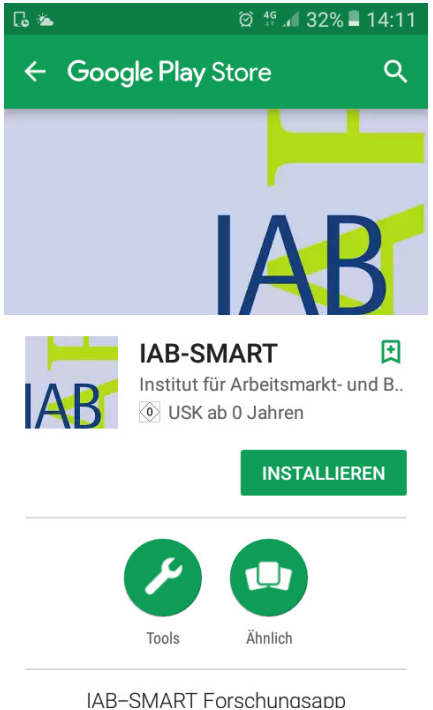
Collected data can be combined with…

… German panel data

… administrative data

Over six months of data collection





Passive Data: Geolocation

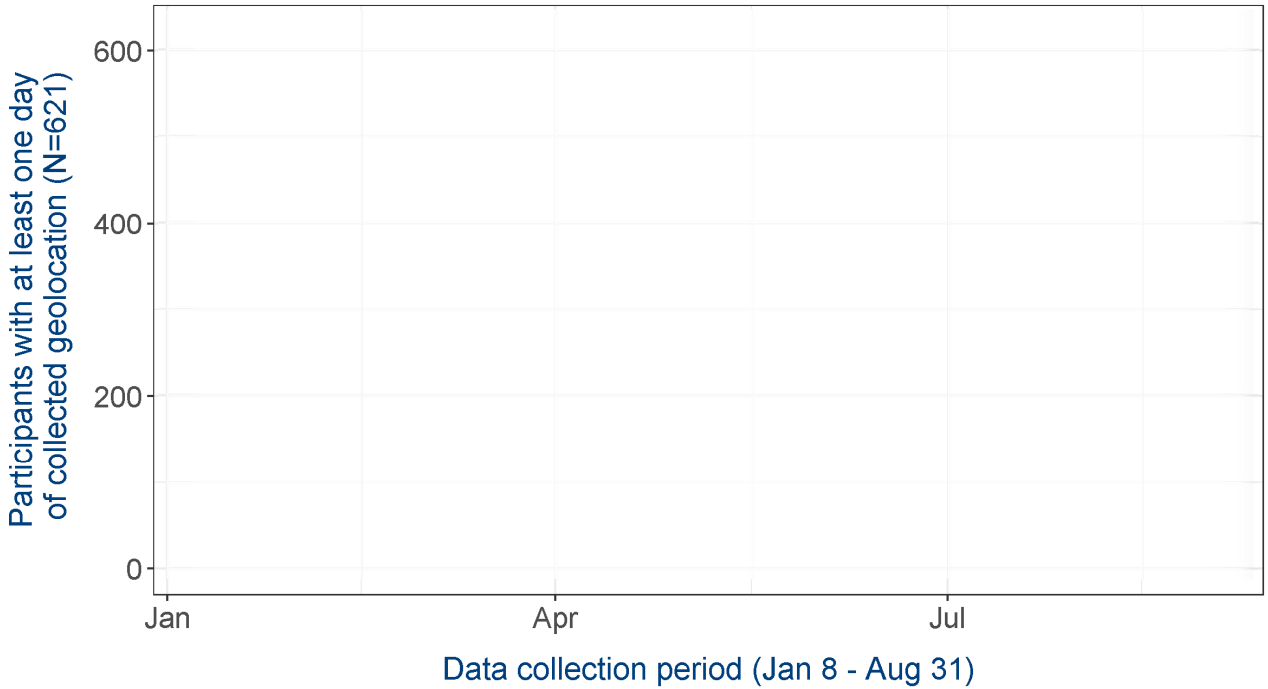
Location sensor data

- Every 30 Minutes
- Geolocation from GPS, mobile carrier network, Wi-Fi (Fused-API)
- Precision (vertically and horizontally) in meters
- Bearing, altitude and speed available
- Precise timestamps for start and end of each measurement

Sample

- 687 (16.7%) installed app
- 621 (90.4%) granted the permission to collect their geolocation
- 483 participants provided geo-data for at least the first 180 days of installation
- Median gap between measurements is 30.7 Minutes, but there are many outliers with far higher gaps (mean 62.3 Minutes)
- Define missing data as gaps > median gap (30.7) minutes

Completeness of data over time

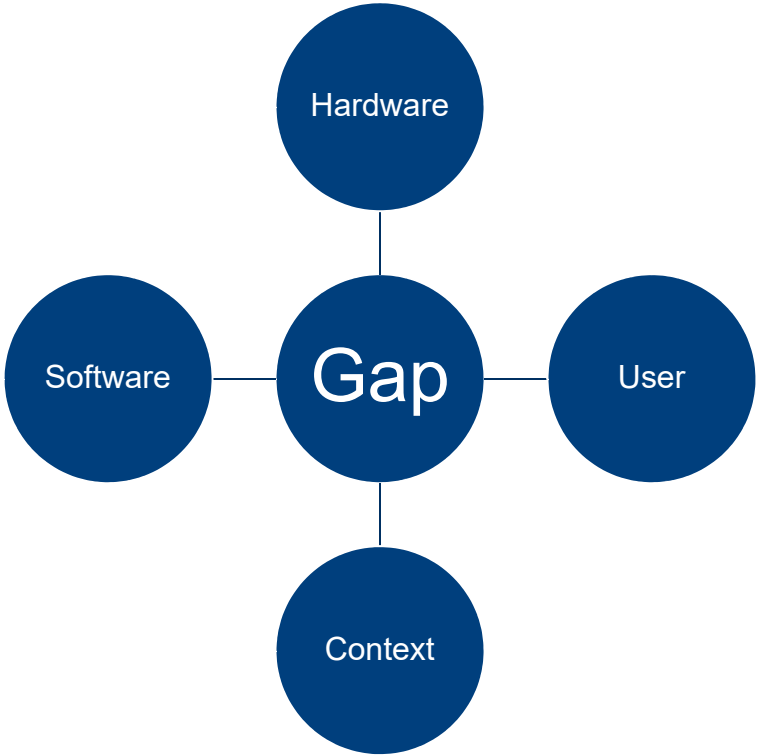


Of all participants who permitted collection of their geolocation:

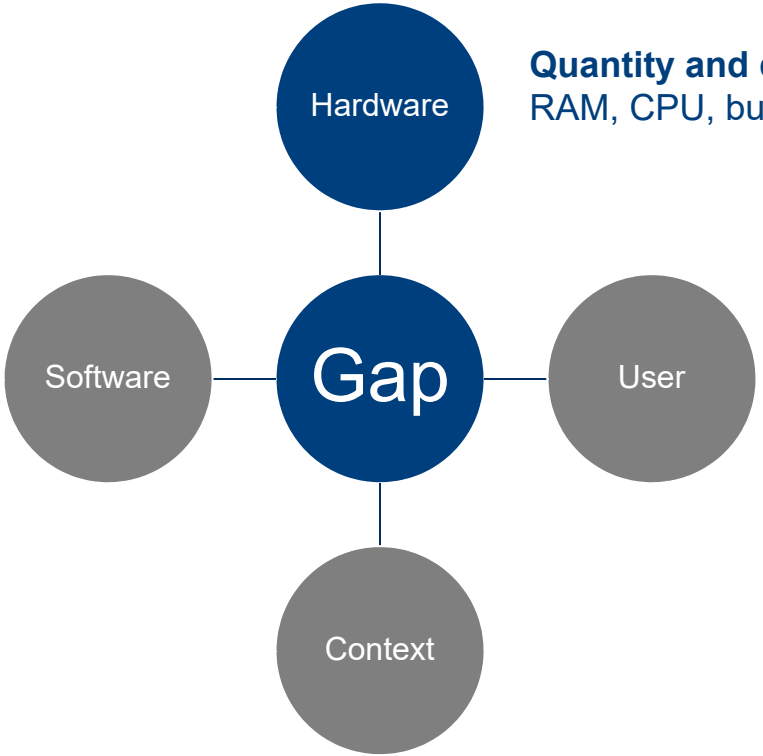
- 73.9% provided at least 180 **cumulative** days of geolocation
- 73,7% provided at least 180 **consecutive** days of geolocation
- Mean Participation: 202 days

Participants sorted by number of days with geolocation measurement

Error sources



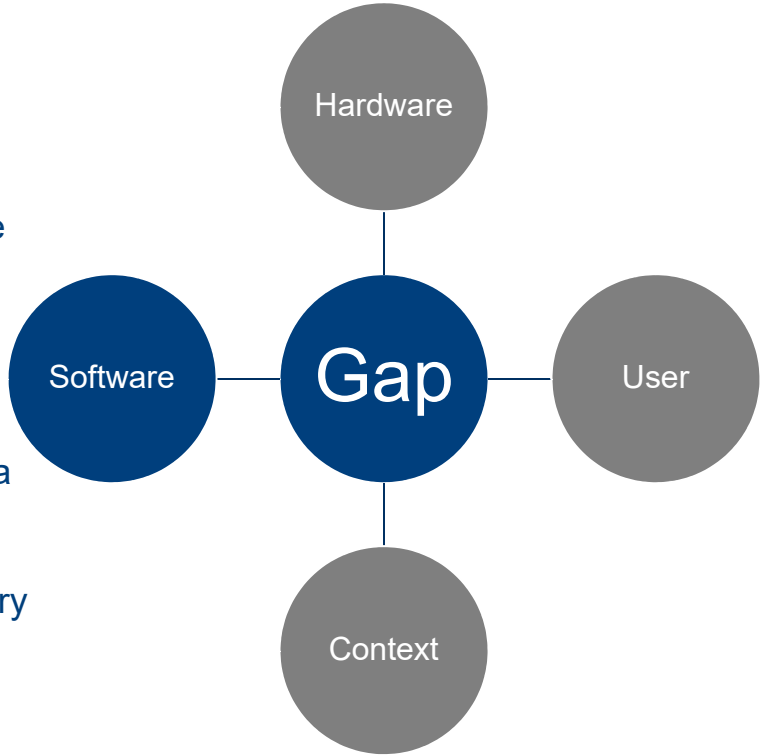
Error sources



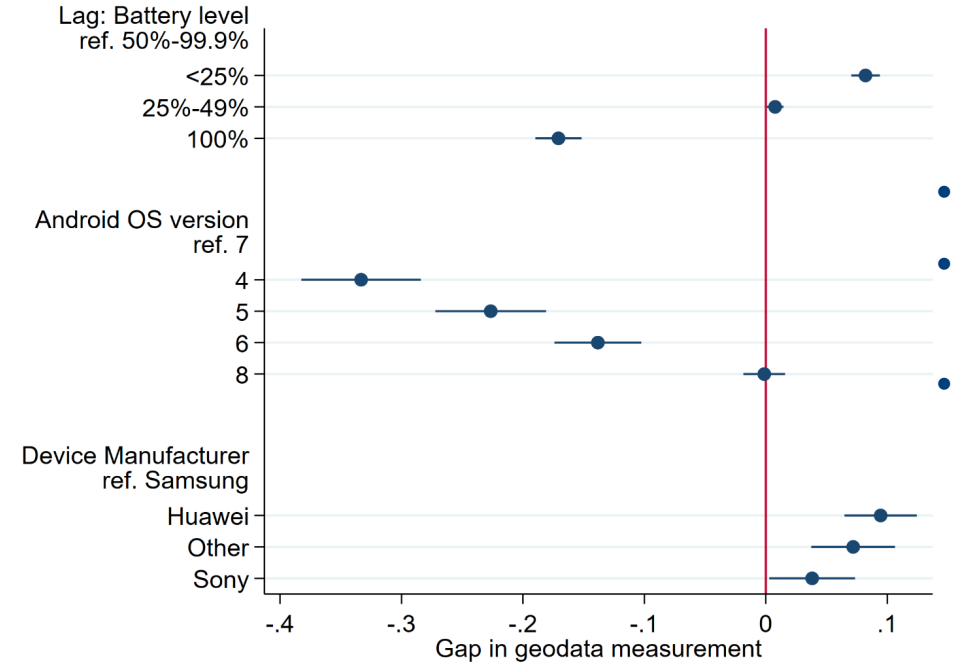
Quantity and quality of components
RAM, CPU, built-in-sensors, battery

Error sources

- **Manufacturer Settings**
  Device specific doze-/battery saving modes inhibit data collection
- **Operating System Settings**
  Data collection may be inhibited by the Operating System (OS)
  OS versions may vary in their rights management
- **Research App Settings**
  How the research app collects the data (what, when, where, for how long, at which interval, from whom)
  Interacts with device / OS / user: battery and RAM/CPU drain
- **Third Party Apps**
  Battery saving apps, Task-killer apps, GPS faker apps



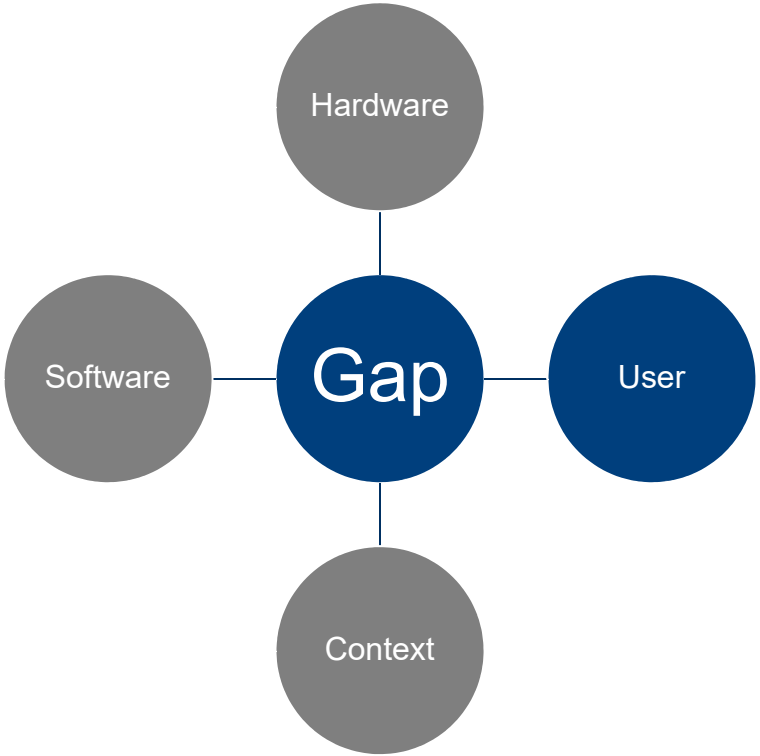
Device-related error sources



AME (with 95% CI) based on binomial probit regression with robust standard errors.

- Low battery endangers data-collection
- Older OS versions seem to be less prone to gaps
- Device specific effects indicate hardware and software issues

Error sources



Participant characteristics
- Technical Competence

Participant behavior
- Fake data, kill / de-install battery-draining apps
- selectively turn off data collection

User-related error sources

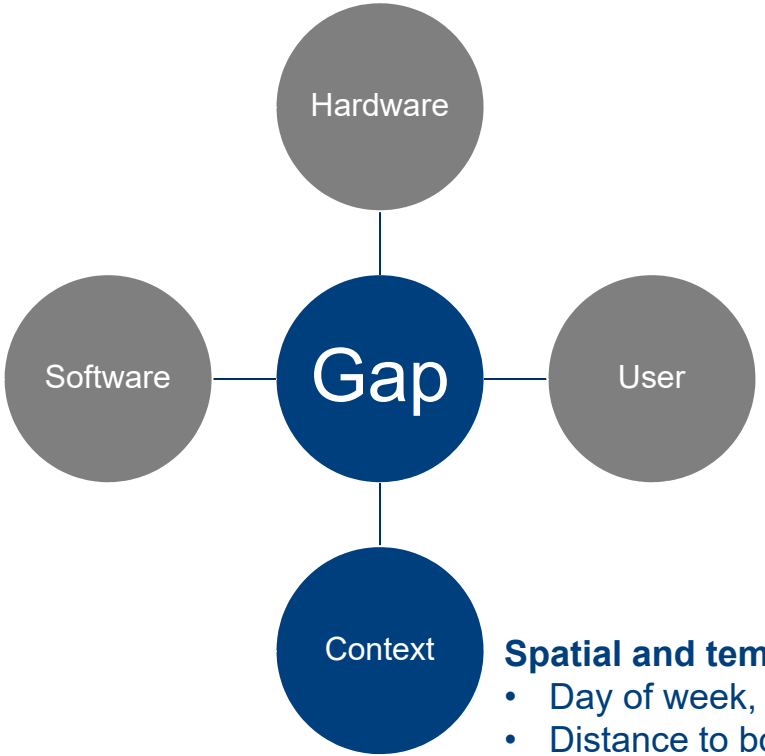
codestring	timestamp	latitude	longit~e	country
dfeh7r4v2v	05aug2018 10:28:48	52.2	8.6	Germany
dfeh7r4v2v	05aug2018 11:43:38	52.2	8.6	Germany
dfeh7r4v2v	05aug2018 12:22:50	8.6	52.2	
dfeh7r4v2v	05aug2018 12:52:49	8.6	52.2	

- Apps falsify geolocation
- Aim: Privacy, access location-specific content
- Validation with app usage data
- 4 / 621 participants had such apps installed

→ Replace false geo-positions with data from immediately before the app use

codestring	AppName	timestamp_start	timestamp_end
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:11:21	05aug2018 12:11:32
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:12:31	05aug2018 12:16:11
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:18:31	05aug2018 12:18:40
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:19:00	05aug2018 12:19:03

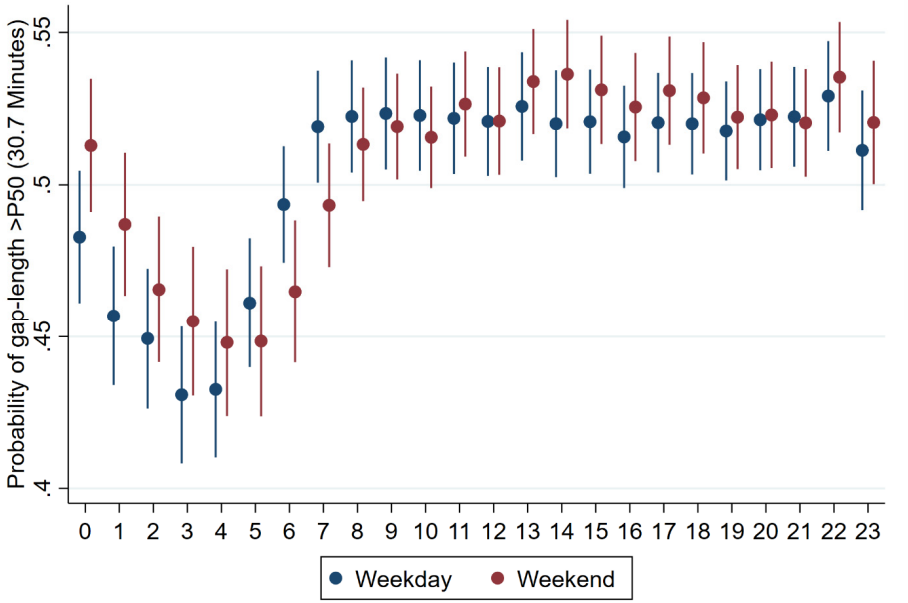
Error sources



Spatial and temporal context

- Day of week, time of day
- Distance to border
- Sparsely populated areas

Context-related error sources



Predicted probability (with 95% CI) based on bivariate probit regression with cluster robust standard errors.

- Time dimension indicates user behavior but also device settings (like doze mode)

Conclusion

- Passive data are not immune to error
- Assessing the quality of passive data necessitates
  - Data specific knowledge
    - How do the sensors work?
    - How are the data collected?
  - A critical stance towards data
    - What checks can we include to assess plausibility and quality
    - Building these checks into the research-app from the beginning
    - Using paradata as control variables in our models
- Future apps might want to
  - Give feedback to users about quality issues (e.g. fake GPS apps)
  - Use the respondents as interpreters of their passive data (is this home/work?)

Thank you! Questions?

Sebastian Bähr sebastian.baehr@iab.de