Developing a Typology for Publicly Available Activity Sensing Data

Fiona Draxler, Yannik Peters, Vanessa Lux

UNIVERSITY OF MANNHEIM
School of Social Sciences

GESIS Leibniz Institute for the Social Sciences

MASS Workshop 2025

5th of June, 2025

Working with Sensor Data…

UNIVERSITY OF MANNHEIM
School of Social Sciences

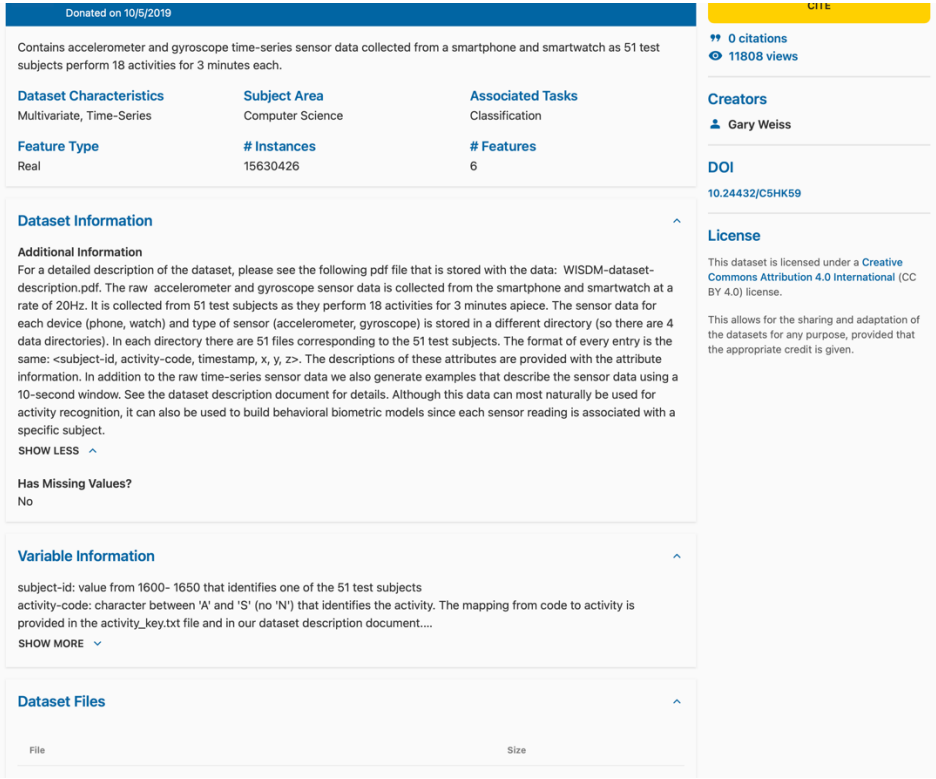
GESIS Leibniz Institute for the Social Sciences

- What sensor datasets are out there already?
  - For secondary analyses
  - For testing algorithms
  - …
- How "good" are they?
  - Extrinsic quality, e.g., variable documentation and clarity about licenses
  - Intrinsic quality, e.g., signal-to-noise ratio, sampling rates

→ We collected and analyzed publicly available datasets on activity.

Example Dataset: WISDM (#21)

- 51 participants performing 18 different activities
- Watch + smartphone (different models)
- Accelerometer + gyroscope
- Developed for a classification scenario



Donated on 10/5/2019

Contains accelerometer and gyroscope time-series sensor data collected from a smartphone and smartwatch as 51 test subjects perform 18 activities for 3 minutes each.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate, Time-Series	Computer Science	Classification
Feature Type	**# Instances**	**# Features**
Real	15630426	6

Dataset Information

Additional Information
For a detailed description of the dataset, please see the following pdf file that is stored with the data: WISDM-dataset-description.pdf. The raw accelerometer and gyroscope sensor data is collected from the smartphone and smartwatch at a rate of 20Hz. It is collected from 51 test subjects as they perform 18 activities for 3 minutes apiece. The sensor data for each device (phone, watch) and type of sensor (accelerometer, gyroscope) is stored in a different directory (so there are 4 data directories). In each directory there are 51 files corresponding to the 51 test subjects. The format of every entry is the same: <subject-id, activity-code, timestamp, x, y, z>. The descriptions of these attributes are provided with the attribute information. In addition to the raw time-series sensor data we also generate examples that describe the sensor data using a 10-second window. See the dataset description document for details. Although this data can most naturally be used for activity recognition, it can also be used to build behavioral biometric models since each sensor reading is associated with a specific subject.

SHOW LESS

Has Missing Values?
No

Variable Information

subject-id: value from 1600- 1650 that identifies one of the 51 test subjects
activity-code: character between 'A' and 'S' (no 'N') that identifies the activity. The mapping from code to activity is provided in the activity_key.txt file and in our dataset description document....

SHOW MORE

Dataset Files

File	Size

0 citations
11808 views

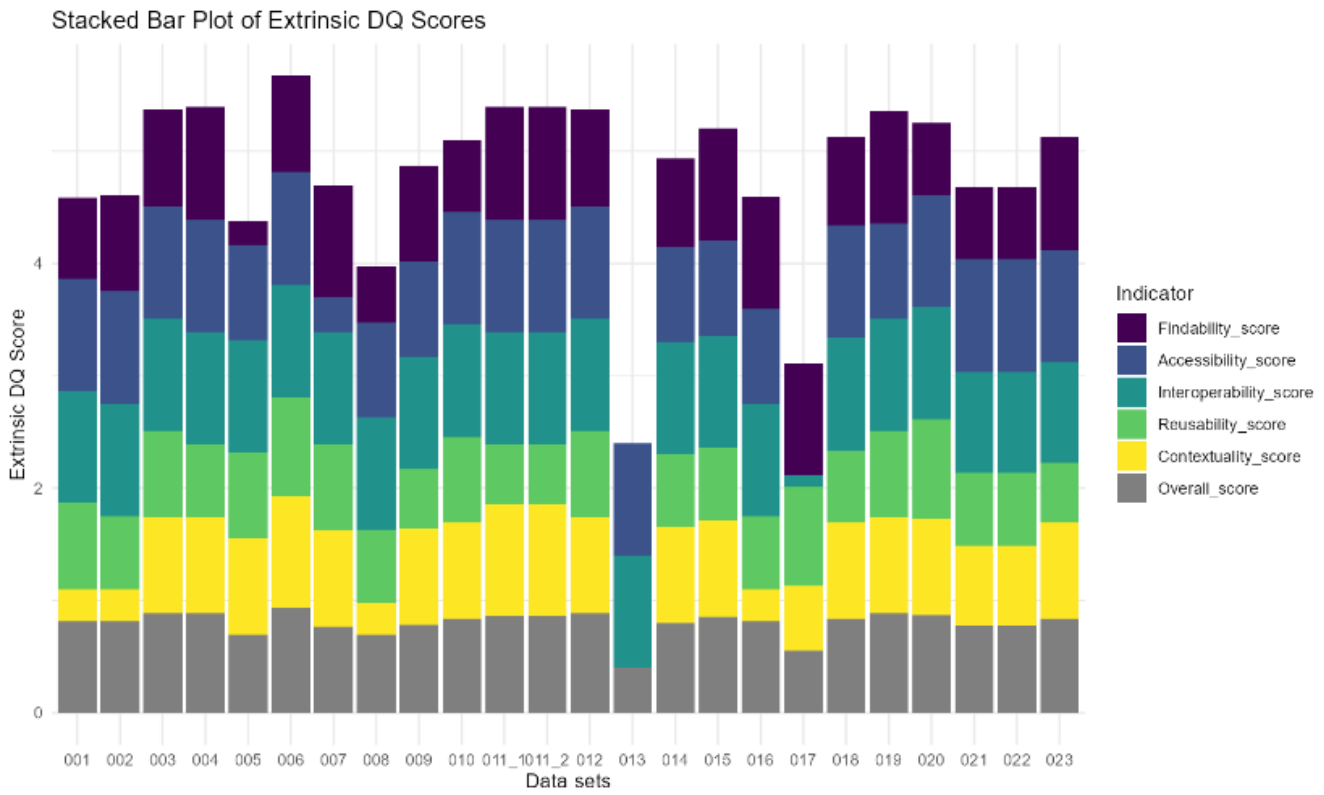
Creators
Gary Weiss

DOI
10.24432/C5HK59

License
This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

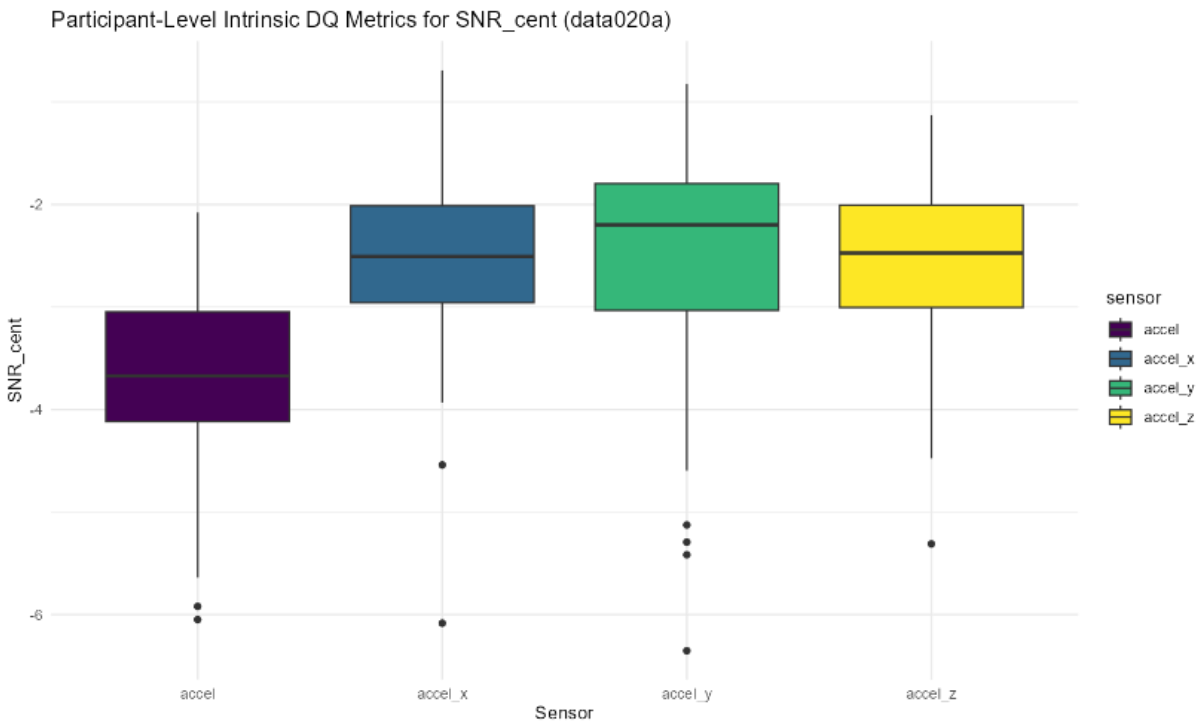
Extrinsic Data Quality



Core questions:

- What are the data?
- What can/may I do with the data?
- First insights into usability

Intrinsic Data Quality



Core aspects:

- Completeness
- Correctness
- Consistency

Across sensors, participants, (datasets ← later)

Outlook: Across-Dataset Classification

UNIVERSITY OF MANNHEIM
School of Social Sciences

GESIS Leibniz Institute for the Social Sciences

- Features derived from sensor axes in datasets 5 and 12
  → Activity prediction (driving, running, sitting, sleeping, standing, walking)
- Random forest model

		true					
		driving	**running**	**sitting**	**sleeping**	**standing**	**walking**
predicted	**driving**	8470	0	144	31	0	22
	running	0	1338	0	0	0	6
	sitting	0	1	5354	3	10	44
	sleeping	0	0	8	5147	0	0
	standing	0	0	17	0	3964	0
	walking	6	56	61	0	70	9128

As researchers interested in sensors…

- How can we make the dataset collection (even) more useful to you?
- What data quality aspects are particularly important?
- What are your favorite datasets that we should include?

Contact us!

Fiona.Draxler@uni-mannheim.de

Vanessa.Lux@gesis.org