Contribution ID: **65** Type: **Oral Presentation**

Integrative Pipelines for Preprocessing Mobile Sensing Data

Thursday 24 July 2025 09:30 (15 minutes)

Oral presentation

Integrative Pipelines for Preprocessing Mobile Sensing Data

Author

Larissa Sust

Affiliation

Ludwig-Maximilians-Universität München

Abstract

Research in the social sciences has traditionally emphasized questionnaire-based assessments, often overlooking the study of real-world behavior. However, with the proliferation of smartphones and digital platforms, researchers now have access to vast amounts of behavioral data generated in people's everyday lives. This shift offers unprecedented opportunities to model diverse phenomena from psychology and beyond, but it also introduces challenges as digital behavioral data are often high-dimensional and low-density, requiring sophisticated preprocessing techniques to extract meaningful variables for formal analysis. Addressing these challenges is essential to ensure the integration of such data into behavioral research in a replicable and sustainable manner.

In this presentation, we introduce a conceptual framework for systematically and transparently reporting preprocessing strategies for mobile-sensing data. Drawing from extensive analyses of smartphone-generated data, including, for example, high-resolution app usage events or GPS logs, our framework focuses on two key dimensions of preprocessing. The first dimension, data enrichment, involves transforming raw data into meaningful variables by adding context, such as integrating multiple data sources or creating new labels. The second dimension, data aggregation, refers to summarizing data at various levels of complexity, ranging from basic descriptive statistics to advanced machine learning models.

To illustrate the application of this framework, we present several preprocessing cases. In the simplest scenario, raw logging data are meaningful enough to be aggregated directly. For instance, app usage events from a specific app, such as TikTok, can be grouped into sessions to derive behavioral indicators like daily usage duration using straightforward algorithms. However, most raw data require enrichment before aggregation. This can involve manually categorizing app usage events into broader categories, such as social media apps, to generate more general variables. Advanced enrichment methods, such as natural language processing, can also be applied, for example, by automatically creating app labels based on their commercial descriptions. Further complexity arises when integrating multiple data sources to provide richer context. For instance, app usage data can be combined with GPS logs to explore patterns, such as how social media usage differs when being at home versus elsewhere. Clustering algorithms, like DBSCAN, can reduce raw GPS coordinates into location categories. Once enriched, these data can be aggregated using statistical models to extract variables that capture relationships across sources. For example, integrating app usage data with ecological momentary assessments (EMAs) can reveal person-level parameters, such as how smartphone-mediated communication relates to social experiences. Again, more complex approaches like machine learning models may also be applied to establish association between data, which may, in turn serve for formal statistical modeling afterwards.

These cases highlight how the complexity of preprocessing affects both the computational demands and the

interpretability of the resulting variables. By systematically addressing these challenges, the proposed framework aims to enhance the replicability and interdisciplinary integration of mobile-sensing research. Ultimately, this approach provides practical guidance for leveraging high-dimensional digital data to explore behavior more effectively.

Keywords

digital behavioral data; data preprocessing

Primary authors: SUST, Larissa (LMU Munich); Prof. SCHOEDEL, Ramona (Charlotte Fresenius University); STERNER, Philipp (Ruhr University Bochum); GORETZKO, David; Prof. BÜHNER, Markus (LMU Munich)

Presenters: SUST, Larissa (LMU Munich); Prof. SCHOEDEL, Ramona (Charlotte Fresenius University); STERNER, Philipp (Ruhr University Bochum); GORETZKO, David; Prof. BÜHNER, Markus (LMU Munich)

Session Classification: Session 19 : "Advanced statistical models and trust in Science"

Track Classification: Design/Research methods: Design/Research methods