東京大学　関本研究室 / Sekimoto Lab. IIS, the University of Tokyo.

Pseudo-PFLOW: A Nationwide Synthetic Open Dataset for People Flow based on Limited

Takehiro Kashiyama, Yanbo Pang, Yoshihide Sekimoto

Background

Recently, the rapid increase in human mobility datasets has provided opportunities to uncover the mechanisms of daily movement. However, the release of such data is restricted due to privacy concerns. To overcome this data access barrier, an innovative approach has been developed for generating large-scale daily synthetic human mobility, termed as the "Pseudo-PFLOW" dataset, which maintains stable accuracy across the entire population in Japan. This innovative approach bridges the gap between data accessibility and privacy, contributing to solutions for major societal issues.

Data Overview

This dataset reproduces pseudo-human flows with entire Japan population (130 million) throughout a typical weekday nationwide, provides not only fragmented location information but also details on who is moving, for what purpose, when, by which means of transport, and from where to where. The pseudo-people flow data provided consists of the following four types of datasets:

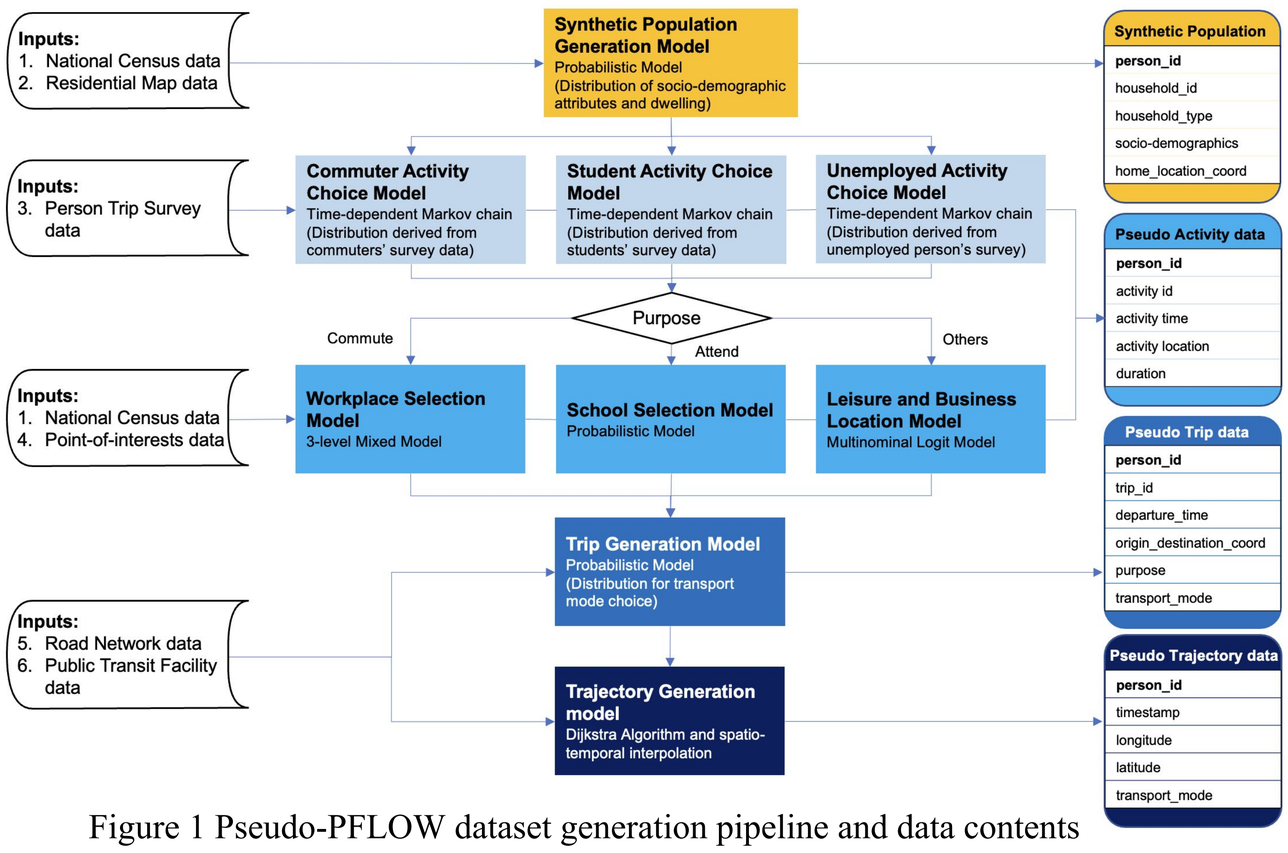

Figure 1 Pseudo-PFLOW dataset generation pipeline and data contents

Population Data represents individual demographic attributes, including household composition, age, gender, employment status, and home address.

Activity Data represents typical daily activities of individuals, capturing when and where these activities.

Trip Data provides insights into human activities, detailing who is moving, when, for what purpose, and from where to where.

Trajectory Data generated based on the trip data, which records the position of movers every few seconds and is treated similarly to GPS data.

Data Release

The Pseudo-PFLOW dataset was made available to researchers worldwide through the Joint Research Access System (JoRAS) of CSIS at the University of Tokyo. Scholars can access the datasets from JoRAS (https://joras.csis.u-tokyo.ac.jp/dataset/list_all) free of charge by submitting a joint research application with the CSIS.

Accuracy of Provided Data

The pseudo-people flow data was evaluated using mobile phone data and urban trip surveys. The dataset faithfully represents

1. Population distribution : R^2=0.81 at 1000×1000m^2 resolution
2. Number of trips: R^2=0.58-0.67 for commuting in metropolitan areas, >0.5 for leisure activities)
3. Trip coverage: around 10% error.

However, due to privacy concerns, some individual-level accuracy in trajectory data, such as path choices during walking or shopping, may be compromised.

For a comprehensive understanding, please refer to the related research paper (https://arxiv.org/abs/2205.00657).

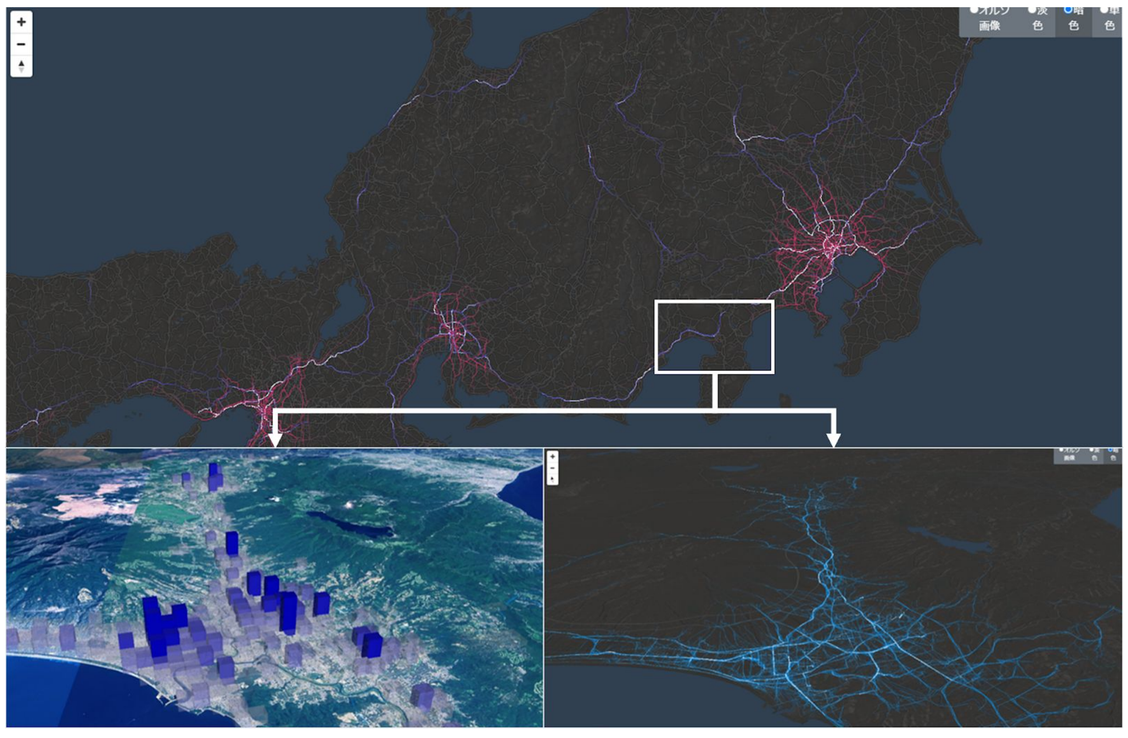

Figure 2 Visualization of aggregated Pseudo-PFLOW data. Up: link traffic volume. Bottom: Hourly Population Distribution.

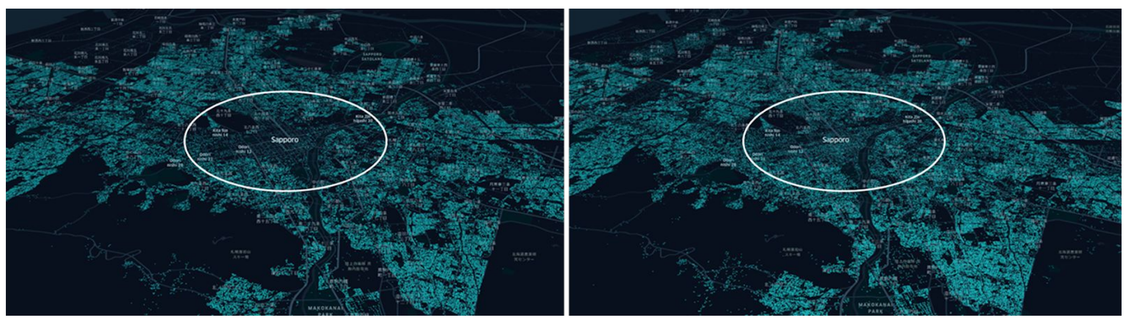

Figure 3 Raw data of Pseudo-PFLOW in Sapporo with 2 million population.