MobiSurvStd: A tool to standardize French mobility survey datasets

Lucas Javaudin
LVMT, ENPC, Institut Polytechnique de Paris, Univ Gustave Eiffel
THEMA, CY Cergy Paris Université

Outline

Mobility Surveys

- Also called "Household Travel Survey (HTS)" in Eqasim
- Household members are asked about their characteristics and the trips they did the previous day:
  - ▶ Households: home location, housing type, number of cars, etc.
  - ▶ Persons: age, gender, profession, work / study location, driving license, etc.
  - ▶ Trips: origin, destination, main mode, departure and arrival time, etc.
  - ▶ Legs (trip segment): mode, car used, parking type, etc.
  - ▶ Cars: fuel type, age, ownership type, etc.
  - ▶ Motorcycles: fuel type, age, etc.
- Usually 5k / 10k *representative* households are surveyed within the survey area
- Example: EGT (Île-de-France), EMC^2 (CEREMA), EMP 2019 (national)

Issues with Mobility Surveys (in France)

- Diverse formats: CEREMA (EMC^2, EDGT, EDVM, EMD), IDFM (EGT 2010, EGT H2020), SDES (ENTD 2008, EMP 2019), IPR (EMG)
- CSV files (+ GIS files) not always easy to read (which separator, which encoding, which datatype)
- Unclear variable and modality names (e.g., in the EMC^2 format, "P2" is gender with modality 1 for a man and 2 for a woman)
- Complexe join operations between datasets (e.g., merging households with persons or trips)
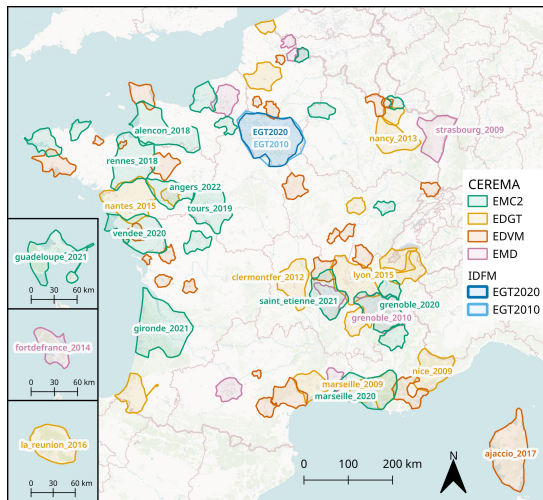
Standardization with MobiSurvStd

- Open-source Python library
- Converts an original survey into its standardized format with a single command
- Parquet format with a clean and complete documentation
- Reads and converts almost all original variables (including zones), for a total of 286 standardized variables
- Many cleaning / validation operations are performed

Check it out: `https://mobisurvstd.github.io/MobiSurvStd/`

Supported Surveys

7 supported survey types:

- 21 EMC^2 surveys (2018–2022)
- 15 EDGT surveys (2009–2017)
- 24 EDVM surveys (2010–2017)
- 7 EMD surveys (2009–2017)
- 2 EGT surveys (2010, 2018–2020)
- 1 EMP survey (2019)

Getting Started

1. Download the EMP 2019 survey (anonymized open-data version), or use any other survey you have access to

2. Install mobisurvstd from PyPI (e.g., `pip install mobisurvstd`)

3. Standardize your survey:

   ```
   python -m mobisurvstd my_survey.zip std_output_dir/
   ```

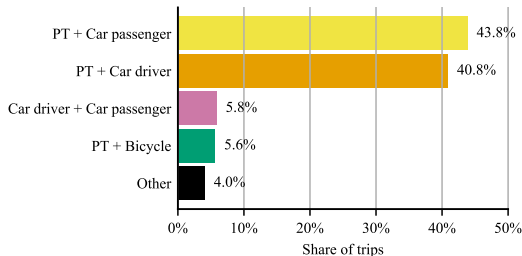4. Analyze the standardized Parquet files with your language of choice

```
std_output_dir
├── cars.parquet
├── detailed_zones.geo.parquet
├── households.parquet
├── legs.parquet
├── metadata.json
├── motorcycles.parquet
├── persons.parquet
├── trips.parquet
```

Extra Features

- Skip GIS files
- Skip cleaning / validation
- Standardize surveys from Python
- Standardize surveys in bulk
- Python helper to read a standardized survey
- Python helper to read many standardized survey

Example Application: Intermodality

- Aggregation of 68 surveys to analyse intermodality behavior
- Re-calibration of sample weights to be representative of France
- 9 stylized facts on intermodality behaviors

Working paper: `https://lucasjavaudin.com/research/intermodality-analysis/javaudin_intermodality.pdf`

Source code: `https://github.com/LucasJavaudin/intermodality-analysis`
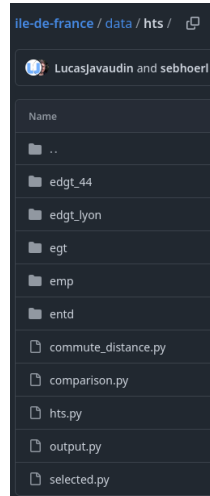
Outline

Use of mobility surveys in Eqasim

1. Statistical matching between synthetic persons and survey persons based on:
   - gender
   - car ownership
   - age (6 classes)
   - socio-professional class
   - department
   - income (if enabled)
   - urban type (if enabled)
2. "Enrich" synthetic persons with values from the matching survey persons:
   - driving license
   - public-transit subscription
   - `is_passenger`
3. Add activity / trip chain from the matching survey persons
4. (4.) Read trip distance distributions from the survey trips

Surveys compatible with Eqasim

- EGT 2010
- ENTD 2008
- EMP 2019
- EDGT Lyon 2015 (2 formats)
- EDGT Nantes 2015
- (PR) Bordeaux 2021
- (old PR) Generic EMC^2

Local vs National Survey

Reasons to prefer recent local surveys when available:

- Better representation of activity chains (especially for post-Covid surveys)
- Better representation of driving license / PT subscription variables

MobiSurvStd Integration

Pull request #346 `https://github.com/eqasim-org/ile-de-france/pull/346`

- Simple two-step process:
  1. The pipeline runs MobiSurvStd to convert the provided survey to the MobiSurvStd format
  2. The MobiSurvStd format is converted in the usual Eqasim format
- All 80 surveys supported by MobiSurvStd are supported by Eqasim (+ any new survey added to MobiSurvStd will be automatically supported by Eqasim)
- Better cleaning process?

How to use?

1. Place the mobility survey (zipped or unzipped) somewhere in the data directory
2. Update the YAML configuration to use MobiSurvStd and locate the survey zip / directory

```
config:
    hts: mobisurvstd
    mobisurvstd:
        path: my_survey.zip
```

Extra Enriched Attributes

- Extra columns can be read in the survey's persons characteristics to further "enrich" the synthetic person, based on the existing statistical matching
- Example of columns (availability depends on the surveys):
  - ▶ education level
  - ▶ professional occupation (full-time or part-time worker, student, retiree, homemaker, etc.)
  - ▶ `"work_only_at_home"`
  - ▶ `"has_bike_sharing_subscription"`
  - ▶ `"has_travel_inconvenience"`
- To use this feature, just update the config:

```
config:
    extra_enriched_attributes:
        - "detailed_education_level"
        - "work_only_at_home"
```

Thank you

Useful links :

- MobiSurvStd documentation: `https://mobisurvstd.github.io/MobiSurvStd`
- Eqasim Pull Request: `https://github.com/eqasim-org/ile-de-france/pull/346`