# Development of Dynamic Census:

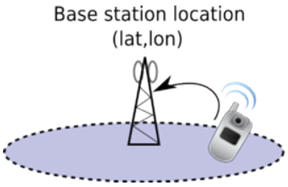
## Estimating demographics and trajectories of actual populations in Bangladesh using CDR data

**University of Tokyo**
**Shibasaki & Sekimoto Lab.**
**Dynamic Census Development Team**
Ayumi Arai*
Apichon Witayangkurn
Hiroshi Kanasugi
Zipei Fan
Ryosuke Shibasaki

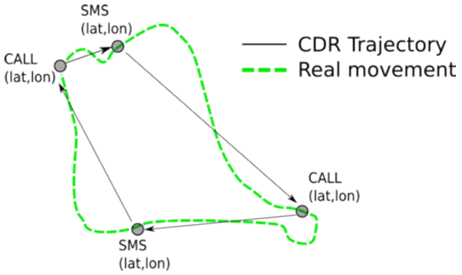
# What is CDR data?

**Call Detail Records = CDR data**

**Localization**



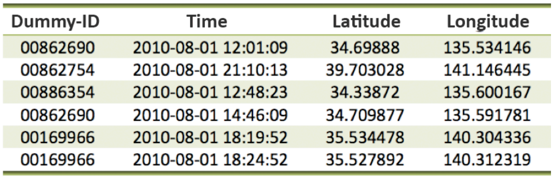
**Trajectory**



**How the data look like**

Starting time of calls

Location of antenna



| Dummy-ID | Time | Latitude | Longitude |
|---|---|---|---|
| 00862690 | 2010-08-01 12:01:09 | 34.69888 | 135.534146 |
| 00862754 | 2010-08-01 21:10:13 | 39.703028 | 141.146445 |
| 00886354 | 2010-08-01 12:48:23 | 34.33872 | 135.600167 |
| 00862690 | 2010-08-01 14:46:09 | 34.709877 | 135.591781 |
| 00169966 | 2010-08-01 18:19:52 | 35.534478 | 140.304336 |
| 00169966 | 2010-08-01 18:24:52 | 35.527892 | 140.312319 |

**CDR data can provide partial views of large-scale human mobility and distribution**

## Motivation

- Population statistics are important for activities both in private and public sectors. But are these enough for understanding human activity?
- CDR data are useful for understanding human mobility. But ..
  - Interpretation of analysis results may be misleading if CDRs can represent limited part of society (James & Versteeg, 2007; Tatem & Smith, 2010)
  - Difficult to examine the impact of representative bias without knowing which part of society CDRs depict (Wesolowski *et al*., 2013).

Can we develop human trajectory data, which are labeled with demographic attributes and represent actual populations using CDR ?

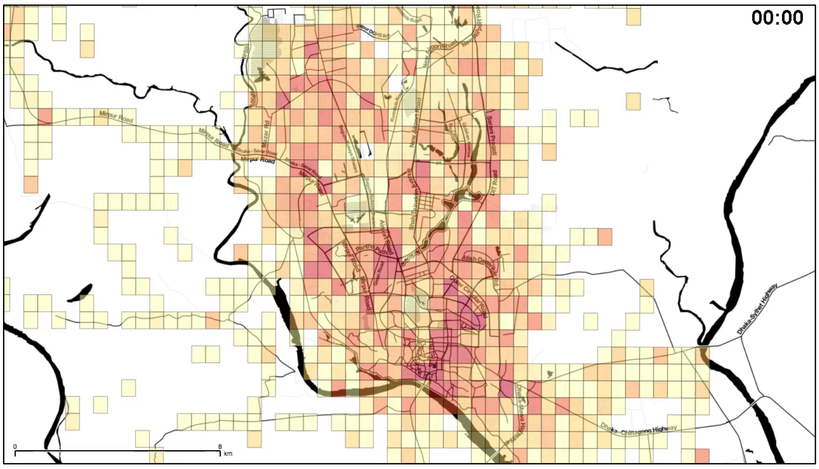
## Advantages and challenges of CDR data

- **Advantages**
  1. Potentially high population coverage
  2. Near real-time human mobility
  3. Routinely collected by the mobile network operator (MNO)
- **Challenges**
  1. Recorded at irregular intervals
  2. Spatial resolution depending on cell antenna locations
  3. Anonymized
  4. Representativeness bias

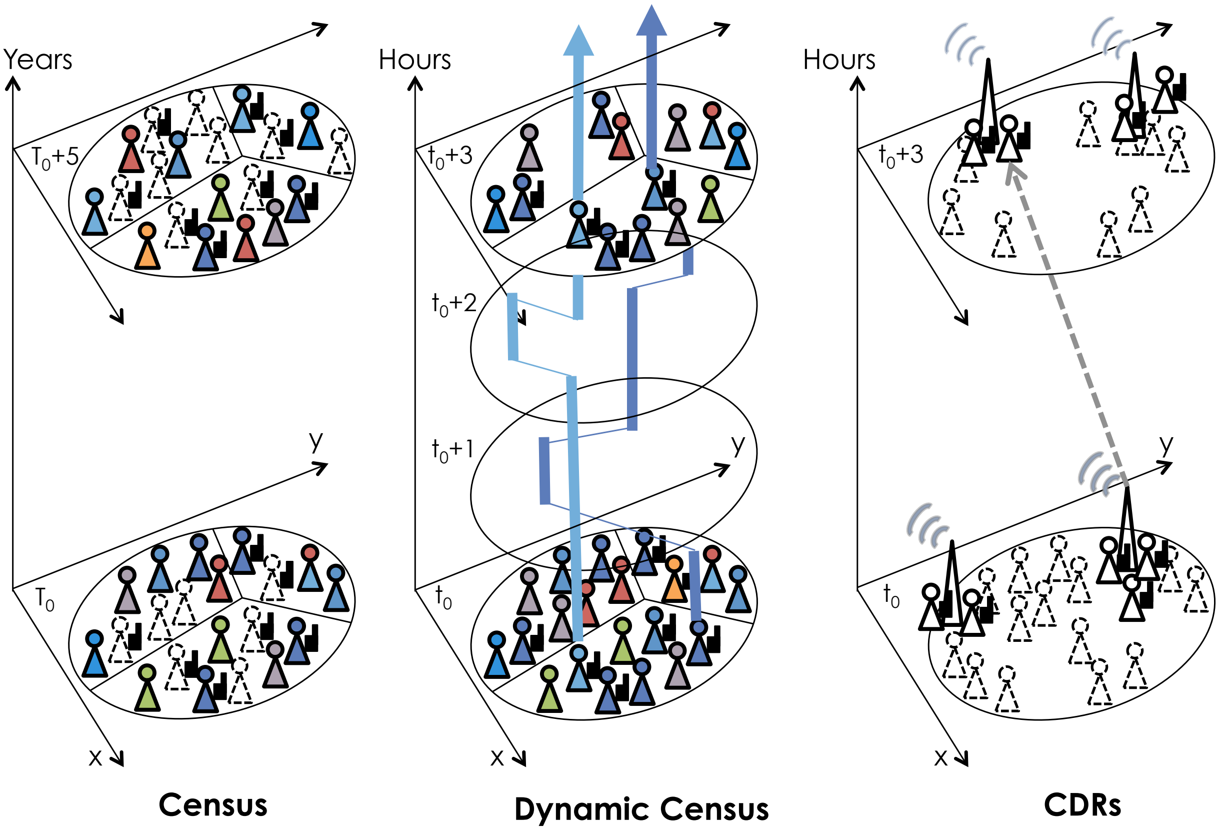
**A novel data set "Dynamic Census" is developed by addressing these challenges**

# What is "Dynamic Census"?

- **Trajectories** and **demographics** of **actual population** in areas covered by CDR data
- **Gridded population statistics** on key demographic attributes at hourly basis, e.g. working male, housewife, student, and other



**Working males**' 24 hour population distribution in Dhaka



**Census**

**Dynamic Census**

**CDRs**

## Impacts and Uniqueness of Dynamic Census

**95%**
World's cellular network coverage

Applicable anywhere covered with cellular networks

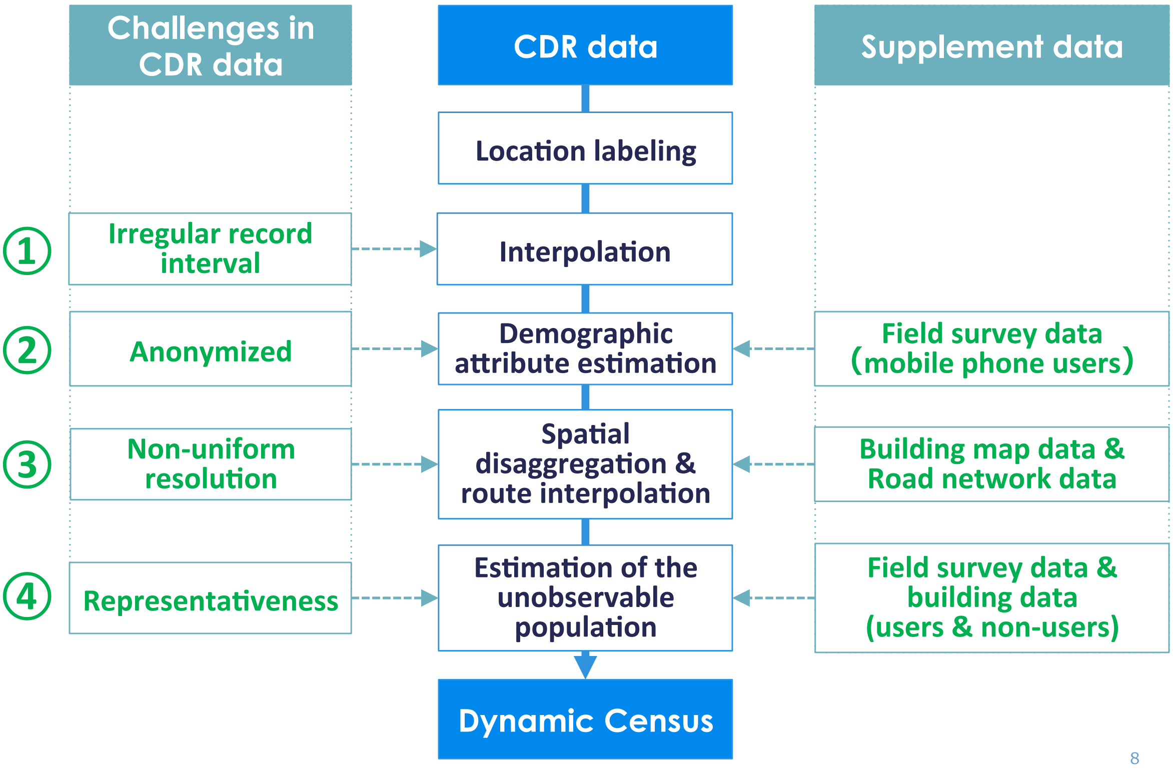
**40%**
Those who belong to Base of Pyramid

Can capture BOP which has non-marginal impacts on economy. Difficult-to-reach population for field survey can be also captured.

**<1 %**
Cost necessary for developing Dynamic Census

Time and financial costs are much lower than conducting conventional census

## How to address challenges in CDR data



Challenges in CDR data

CDR data

Supplement data

Location labeling

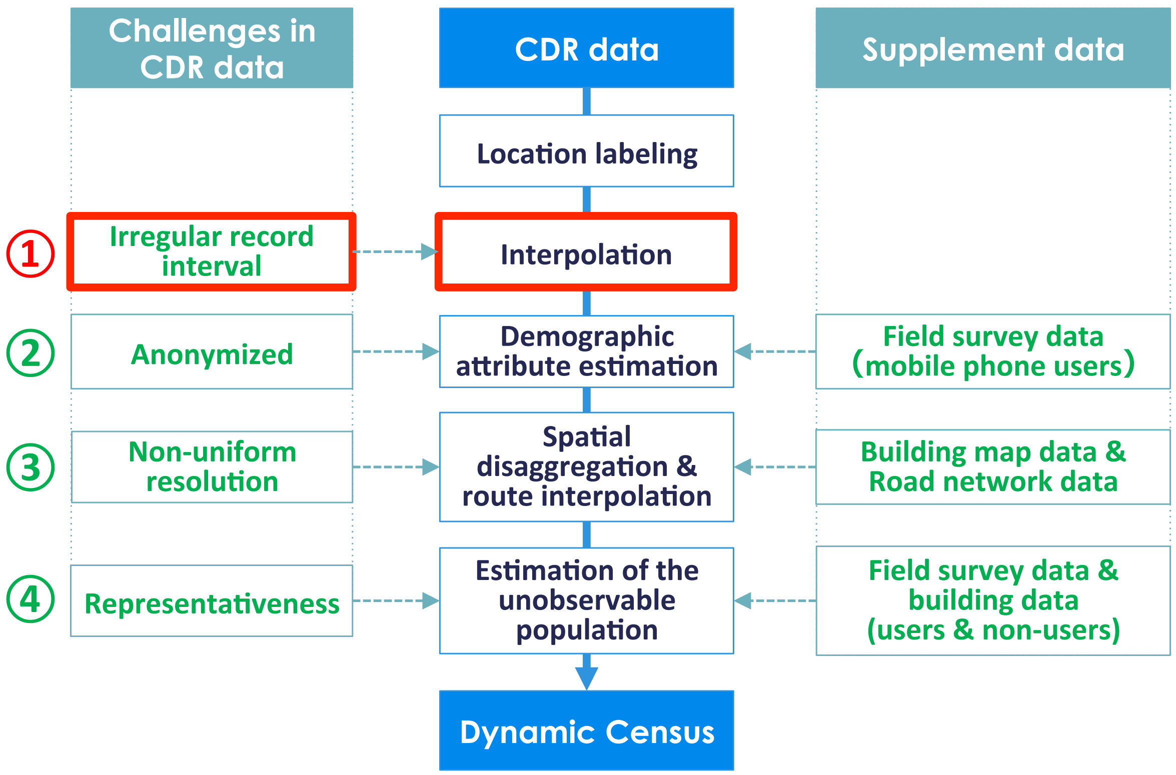
① Irregular record interval → Interpolation

② Anonymized → Demographic attribute estimation ← Field survey data (mobile phone users)

③ Non-uniform resolution → Spatial disaggregation & route interpolation ← Building map data & Road network data

④ Representativeness → Estimation of the unobservable population ← Field survey data & building data (users & non-users)

Dynamic Census



Challenges in CDR data

CDR data

Supplement data

Location labeling

① Irregular record interval → Interpolation

② Anonymized → Demographic attribute estimation ← Field survey data (mobile phone users)

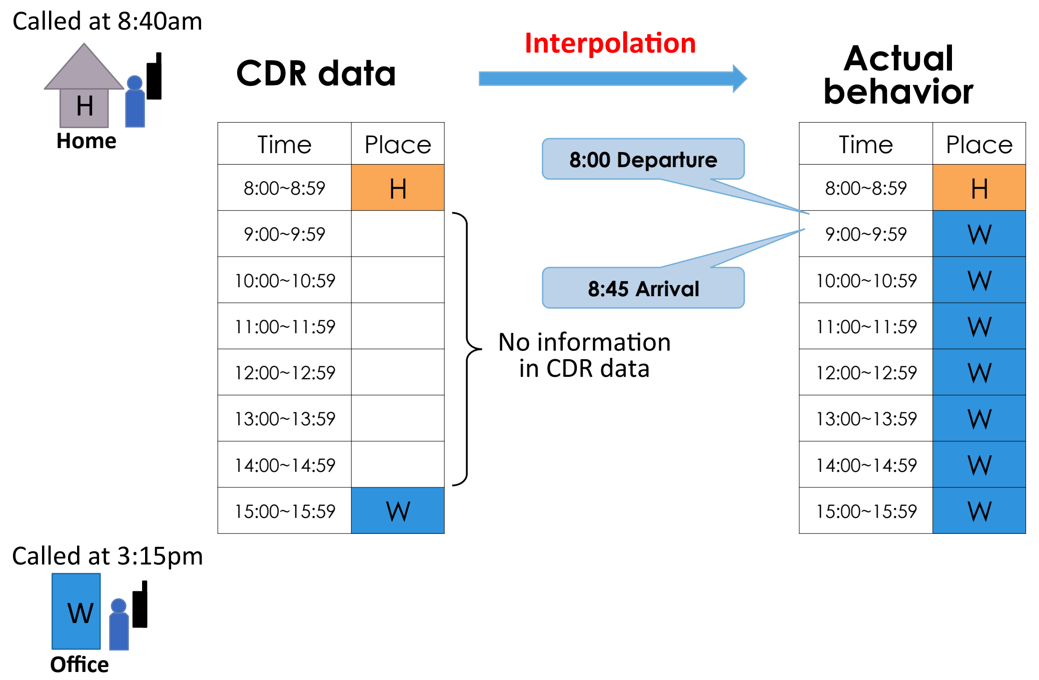
③ Non-uniform resolution → Spatial disaggregation & route interpolation ← Building map data & Road network data

④ Representativeness → Estimation of the unobservable population ← Field survey data & building data (users & non-users)

Dynamic Census

# 1. Irregular record interval

**Interpolate CDRs based on the routine observed from longer-term data**



**CDR data**

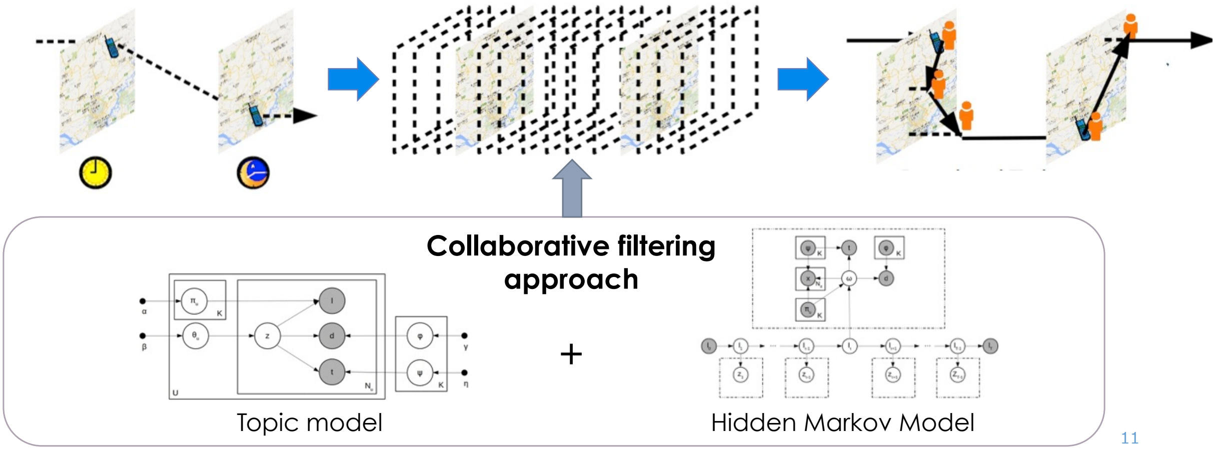
| Time | Place |
|---|---|
| 8:00~8:59 | H |
| 9:00~9:59 | |
| 10:00~10:59 | |
| 11:00~11:59 | |
| 12:00~12:59 | |
| 13:00~13:59 | |
| 14:00~14:59 | |
| 15:00~15:59 | W |

**Actual behavior**

| Time | Place |
|---|---|
| 8:00~8:59 | H |
| 9:00~9:59 | W |
| 10:00~10:59 | W |
| 11:00~11:59 | W |
| 12:00~12:59 | W |
| 13:00~13:59 | W |
| 14:00~14:59 | W |
| 15:00~15:59 | W |

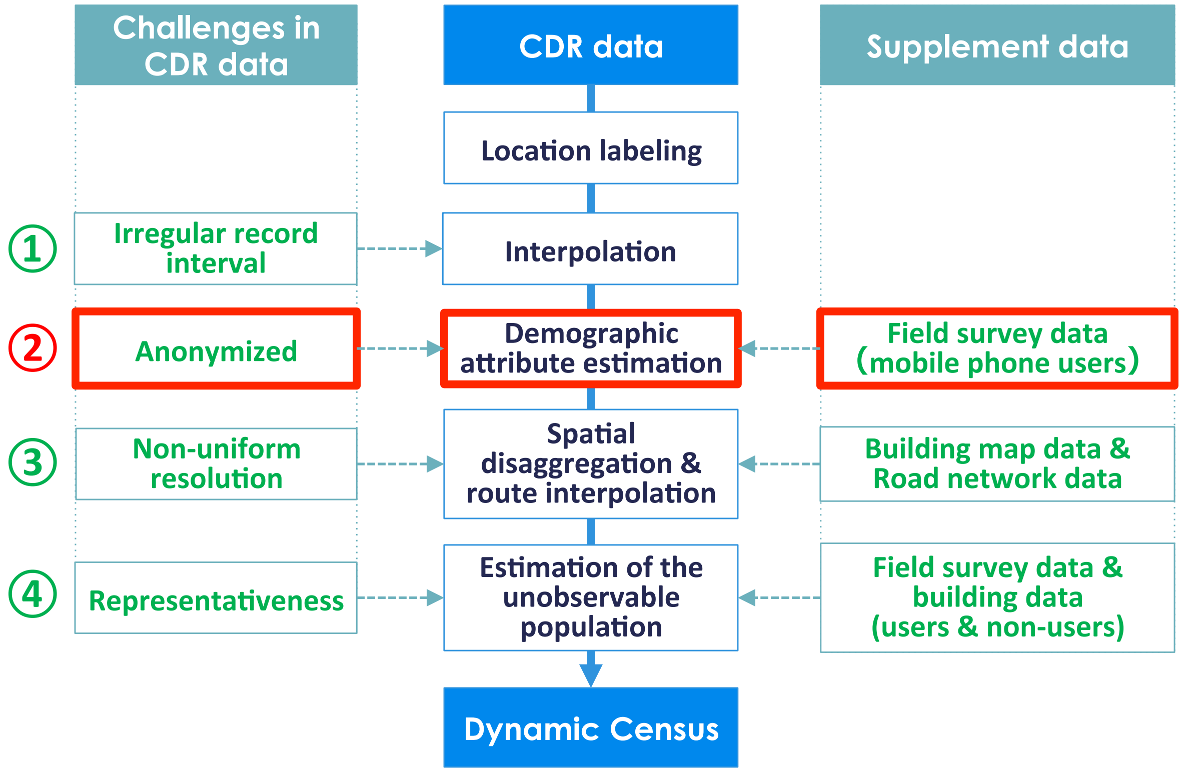
# Interpolation

- **Extracting routine patterns**
  - A topic model is employed
  - Routine pattern is expressed as the probability distribution of key locations (Home, Work, and Other)
- **Spatiotemporal interpolation**
  - Hidden Markov Model is employed (timing of transition is identified)



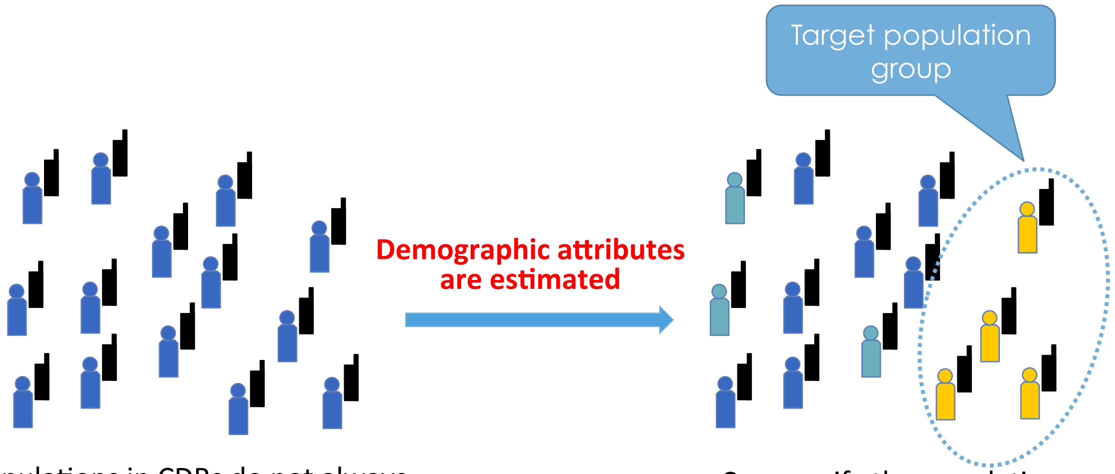
**Collaborative filtering approach**

Topic model + Hidden Markov Model



| Challenges in CDR data | CDR data | Supplement data |
|---|---|---|
| | Location labeling | |
| ① Irregular record interval | Interpolation | |
| ② Anonymized | Demographic attribute estimation | Field survey data (mobile phone users) |
| ③ Non-uniform resolution | Spatial disaggregation & route interpolation | Building map data & Road network data |
| ④ Representativeness | Estimation of the unobservable population | Field survey data & building data (users & non-users) |
| | Dynamic Census | |

# 2. No demographic attribute info



Target population group

Demographic attributes are estimated

Populations in CDRs do not always represent the population under study

Can specify the population under study

# Demographic attribute estimation

- **Approach**
  - Random Forest is employed for building an estimation model
  - One-month-call-records from 58 volunteers are used as training data
  - One-day-call-records from 922 mobile phone users are used for examining relationship between calling behavior and demographic attributes
- **Estimated features**
  - Working male, housewife, student, and other
  - Income level (individual) and Age group (-20/21-35/36-60/61-) ←**Results to be improved**

Estimation results

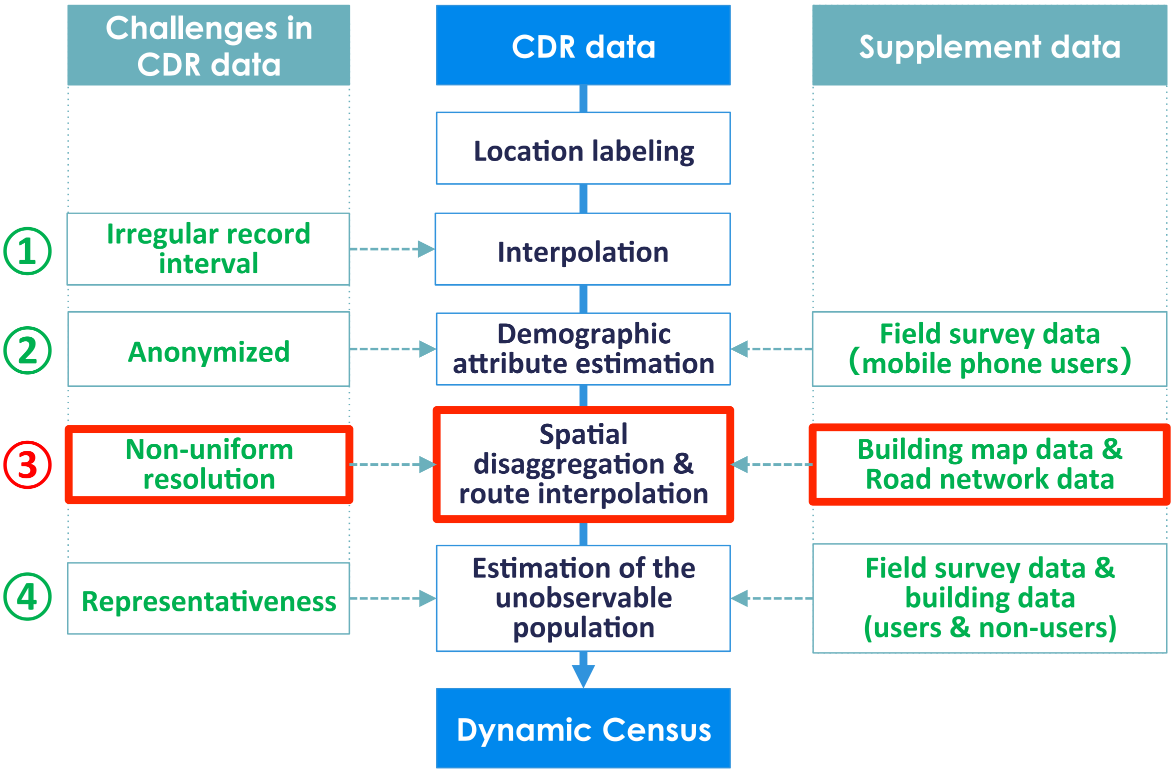
| Class | Accuracy | Precision | Recall |
|---|---|---|---|
| Working male | 0.79 | 0.63 | 0.70 |
| Housewife | 0.67 | 0.47 | 0.82 |
| Student | 0.89 | 0.40 | 0.22 |
| Other | 0.63 | 0.20 | 0.07 |

# Calling behavior survey to relate demographic attributes and CDR data

- **Purpose**
  - Relate calling behavior (call records) and demographic attributes
- **Surveyed area and population**
  - 15 Wards are chosen based on land use. For each Ward, 18 HHs each are chosen from 3 income groups in Greater Dhaka (Two-stage stratified sampling)
  - All members are interviewed
  - Interviewed on demographic attribute, travel-activity, and mobile phone use
- **Key of this survey**
  - Income level is determined based on the type of buildings

Interview at a slum household

Interview at a high income household



Challenges in CDR data

CDR data

Supplement data

Location labeling

① Irregular record interval → Interpolation

② Anonymized → Demographic attribute estimation ← Field survey data (mobile phone users)

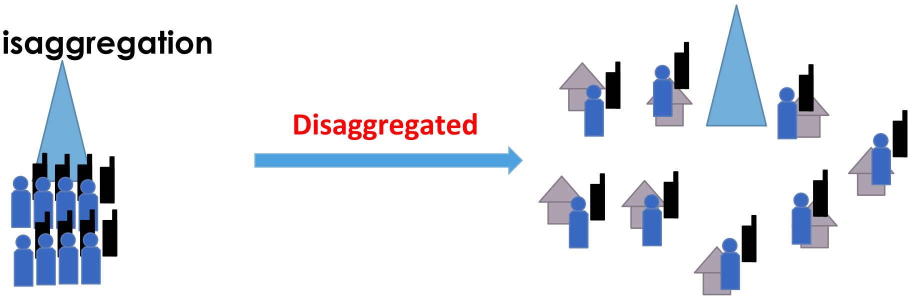
③ Non-uniform resolution → Spatial disaggregation & route interpolation ← Building map data & Road network data

④ Representativeness → Estimation of the unobservable population ← Field survey data & building data (users & non-users)

Dynamic Census

# 3. Non-uniform spatial resolution

**Spatial disaggregation**

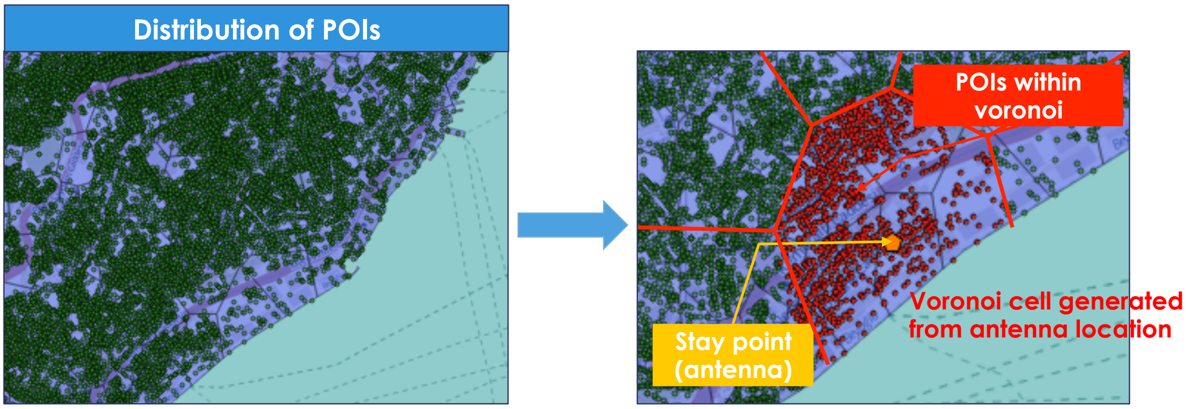


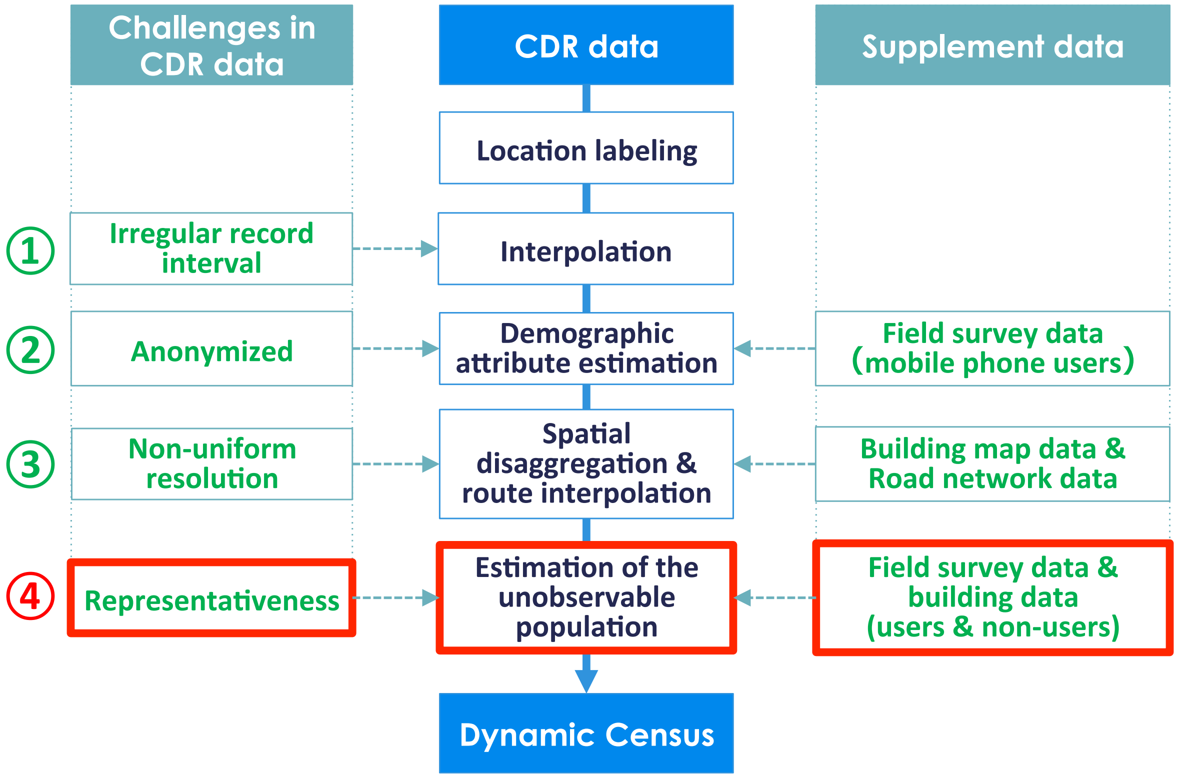
Location in CDR data is at antenna level

Disaggregated based on the distribution of buildings

# Stay point reallocation

- **Modifying spatial resolution**
  - **Stay points are reallocated to building POIs**
    - Antenna basis locations are reallocated to building POIs within voronoi
    - Each voronoi cell is considered to be an area covered by an antenna
  - **Allocation probability is based on the area size of building**
  - **Types of buildings are used as the proxy of the income level**





Challenges in CDR data

CDR data

Supplement data

Location labeling

① Irregular record interval → Interpolation

② Anonymized → Demographic attribute estimation ← Field survey data (mobile phone users)

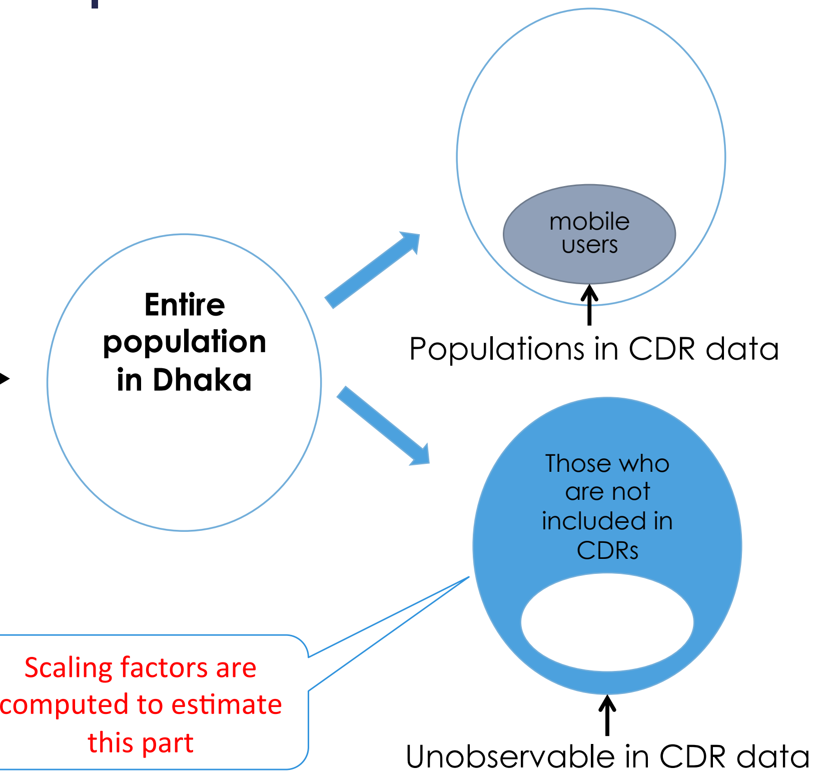
③ Non-uniform resolution → Spatial disaggregation & route interpolation ← Building map data & Road network data

④ Representativeness → Estimation of the unobservable population ← Field survey data & building data (users & non-users)

Dynamic Census

# 4. Representativeness

Suppose you have CDRs from Dhaka …



Entire population in Dhaka
mobile users

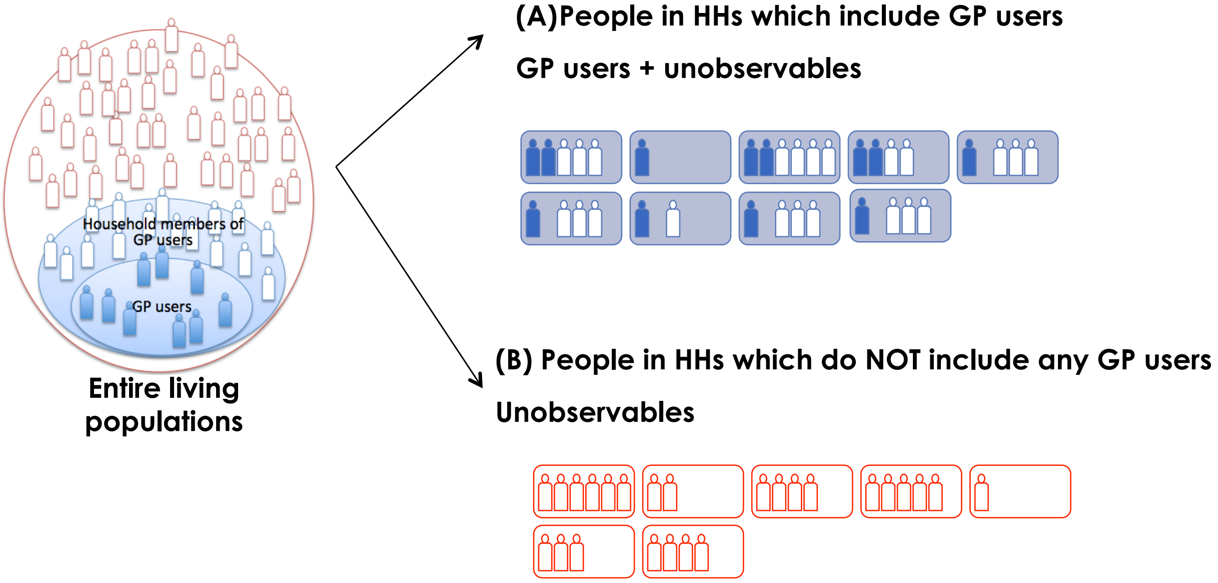
Populations in CDR data

Those who are not included in CDRs

Unobservable in CDR data

Scaling factors are computed to estimate this part

# Understanding population covered by CDRs on household basis



(A)People in HHs which include GP users

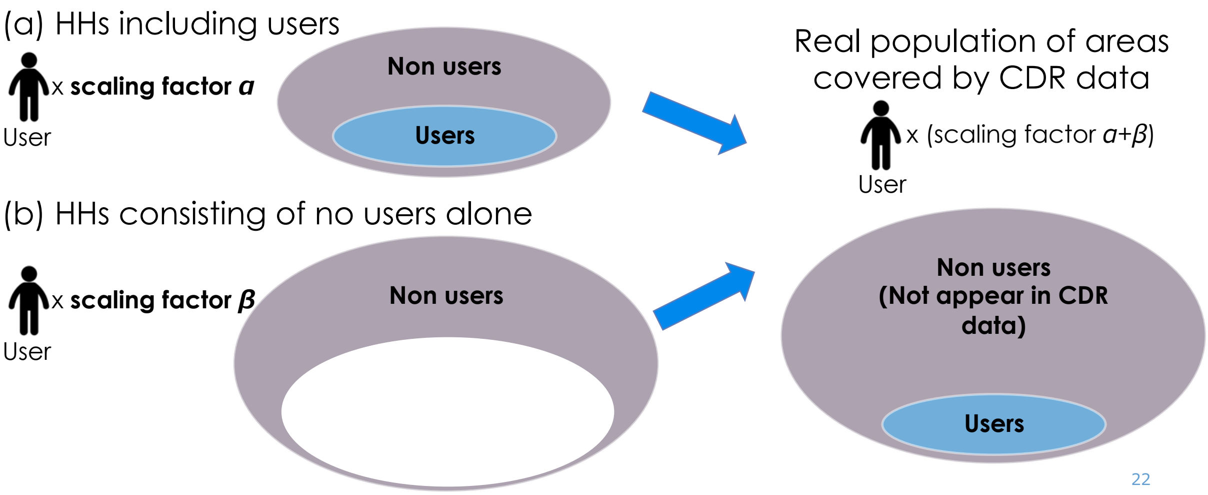
GP users + unobservables

(B) People in HHs which do NOT include any GP users

Unobservables

# Estimation of the unobservable

- **Scaling factor**
  - Approx. household number is calculated based on the number of buildings
  - Scaling factor is computed from the typical HH structure, obtained through field survey



(a) HHs including users

x scaling factor $\alpha$
User
Non users
Users

(b) HHs consisting of no users alone

x scaling factor $\beta$
User
Non users

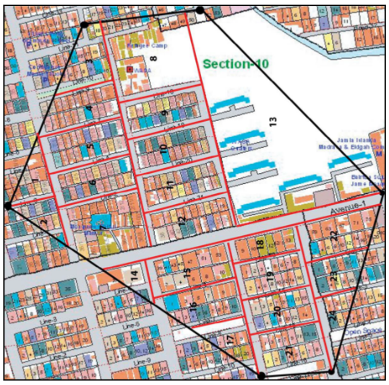
Real population of areas covered by CDR data

x (scaling factor $\alpha+\beta$)
User
Non users (Not appear in CDR data)
Users

# Small-scale census survey (SSC) to see population structure for each income level

## Purpose

- Investigate the population structure for each income level
- Obtain data to calculate scaling factors to compute the number of populations from the distribution of building by income level



Surveyed Voronoi area

## Surveyed area and population

- Entire populations in a Voronoi cell were surveyed in December 2014
- 2,839 HHs consisting of 11,521 people from 366 buildings (out of 367 buildings)

## Key of SCC

- Income level is determined based on the type of buildings

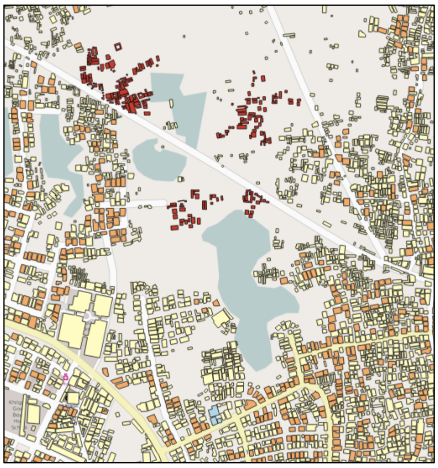
# Average HH structures obtained from survey

| | | HHs including GP users | | | HHs not including GP users | | |
|---|---|---|---|---|---|---|---|
| | | GP user | People not using GP | HH size | GP user | People not using GP | HH size |
| High | Workmale | 0.93 | 0.38 | 4.74 | N/A | 1.07 | 4.00 |
| | Housewife | 0.73 | 0.38 | | N/A | 0.96 | |
| | Student | 0.10 | 1.22 | | N/A | 1.23 | |
| | Other | 0.12 | 0.88 | | N/A | 0.74 | |
| Middle | Workmale | 0.89 | 0.41 | 4.77 | N/A | 1.12 | 4.37 |
| | Housewife | 0.53 | 0.51 | | N/A | 0.97 | |
| | Student | 0.11 | 1.24 | | N/A | 1.35 | |
| | Other | 0.19 | 0.89 | | N/A | 0.94 | |
| Low | Workmale | 0.82 | 0.42 | 4.80 | N/A | 1.29 | 4.47 |
| | Housewife | 0.44 | 0.74 | | N/A | 0.87 | |
| | Student | 0.08 | 1.02 | | N/A | 1.08 | |
| | Other | 0.20 | 1.08 | | N/A | 1.23 | |
| Slum | Workmale | 0.83 | 0.51 | 4.92 | N/A | 1.26 | 4.59 |
| | Housewife | 0.13 | 0.72 | | N/A | 0.68 | |
| | Student | 0.04 | 1.10 | | N/A | 0.94 | |
| | Other | 0.20 | 1.39 | | N/A | 1.71 | |

# Type of building and income level

### Contents of the map data

- Approx. 650,000 buildings (with the type of buildings)
- Residential buildings are classified into four groups by the height of buildings



Sample of the map

### Criteria of the type of buildings

- High (Seven or more stories)
- Middle (More than two stories)
- Low (One to two stories)
- Slum (One story)

### Legend of the type of building

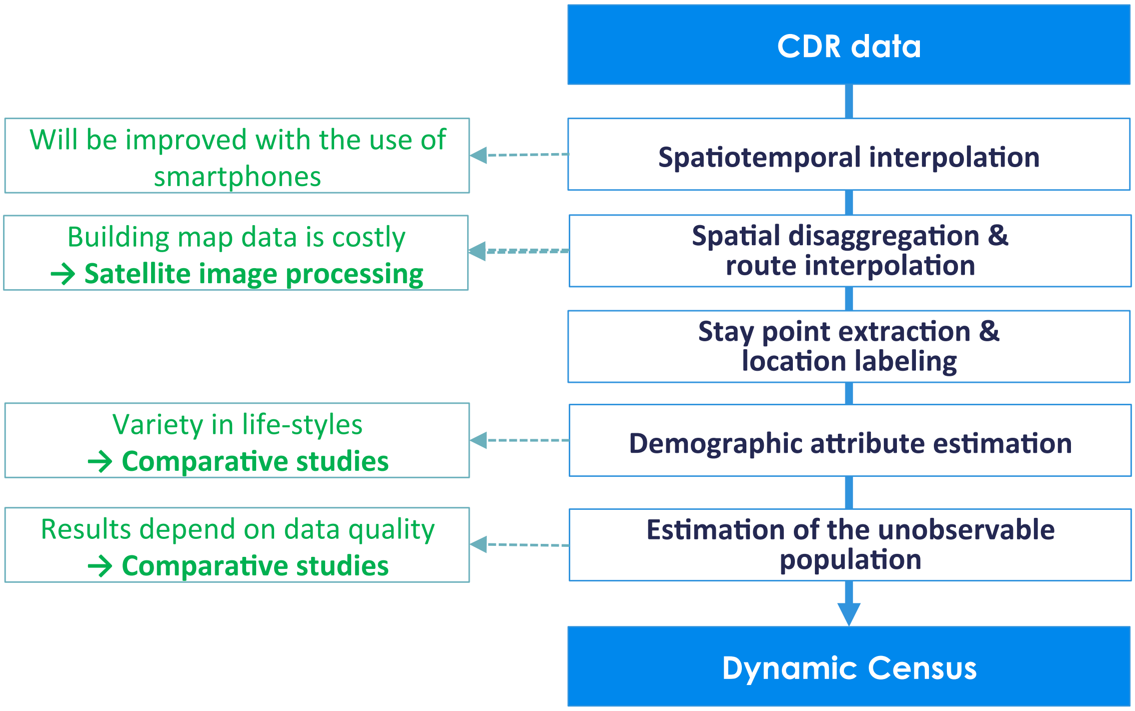
🟦 : High

🟨 : Middle

⬜ : Low

🟥 : Slum

# Future work



CDR data

Spatiotemporal interpolation

Spatial disaggregation & route interpolation

Stay point extraction & location labeling

Demographic attribute estimation

Estimation of the unobservable population

Dynamic Census

Will be improved with the use of smartphones

Building map data is costly
**→ Satellite image processing**

Variety in life-styles
**→ Comparative studies**

Results depend on data quality
**→ Comparative studies**

Any questions/suggestions are welcome!

arai@csis.u-tokyo.ac.jp