Small area statistics on de jure and de facto populations - population census and operational data of mobile phone network

Naoki Makita [a,*], Masakazu Kimura [a], Masayuki Terada [b], Motonari Kobayashi [b] and Yuki Oyabu [b]

a National Statistics Center b NTT DOCOMO R&D Center

Application of Geostatistic – (session 1)
Tuesday October 8 11:10-12:10
Casa de Montejo 2, Centro Banamex
Mexico City

October 7-9
CIGMA2013
Geography and Environment
Internacional Conference
Geographical and environmental information for decision making

National Statistics Center
Incorporated Administrative Agency

NTT docomo

Contents

1. Introduction

- National statistics office(s)
  - The Statistics Bureau of Japan (SBJ)

    and **the National Statistics Center** (**NSTAC**) (SBJ's affiliate)

- A mobile telecommunication company
  - **NTT DOCOMO**

1. Introduction

- Population Census
  conducted by National Statistics Offices
  (i.e. INEGI, SBJ/NSTAC, US Census Bureau, etc.)

- Small area statistics derived from Population Census
  - Census data by census tract / city block
  - **Grid Square Statistics** (Gridded population)

  :-) Essential and most popular data for geostatistics

  :-(Produced not very often

  Statistics users want data more frequently.

1. Introduction

- A new kind of small area statistics: **Mobile Spatial Statistics (MSS)**
  - developed by NTT DOCOMO, derived from **operational data of mobile phone network**
  - would produce small area population estimates as frequently as **on an hourly basis**.

Is it really reliable?

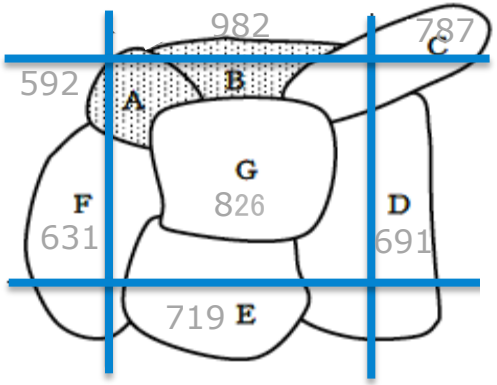
1. Introduction

- NSTAC and DOCOMO have undertook a research to examine the validity of MSS
  - by identifying qualitative differences between GPC and MSS;
  - by calculating 'deviation rates' to quantitatively compare GPC and MSS
- We concluded that **MSS would be plausible for densely populated areas.**

2. Grid Square Statistics of Population Census

- Population Census of Japan
  - taken place every five years
- Grid Square Statistics of Population Census (GPC)
  - compiled from the population counts by Census Enumeration District (CED)

Population counts by CED



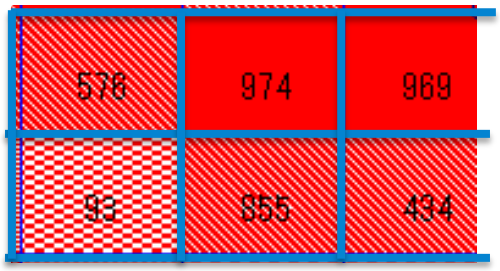
Boundary of CEDs

Grid lines based on latitude and longitude

Assign population counts of CEDs to grid squares

GPC

Population counts by grid square

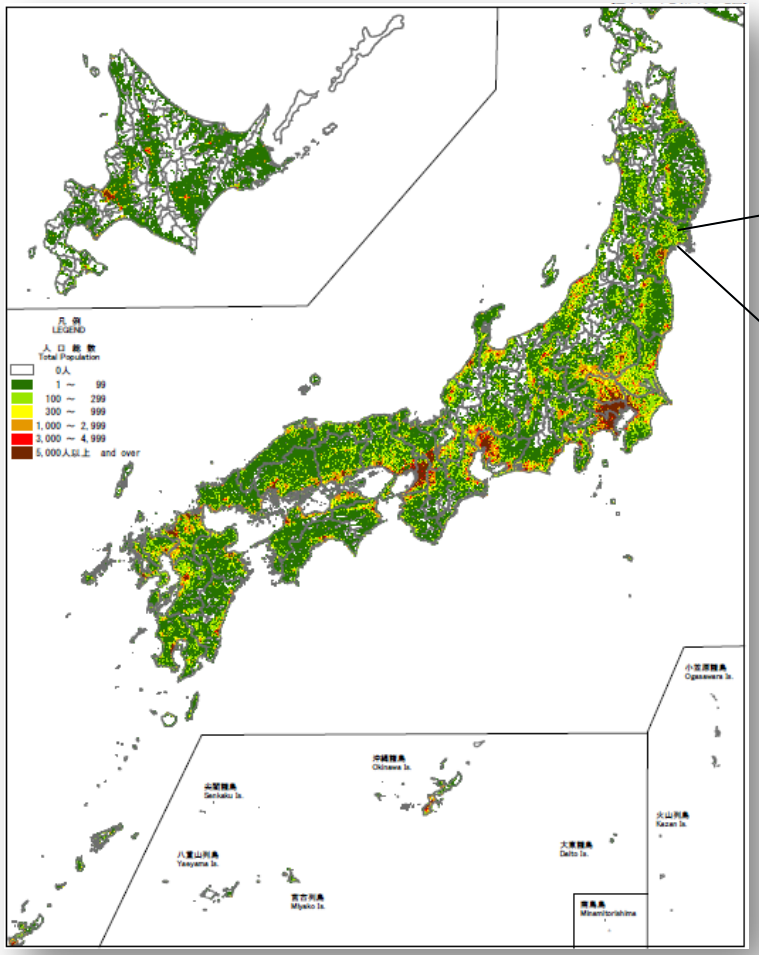


Grid lines based on latitude and longitude

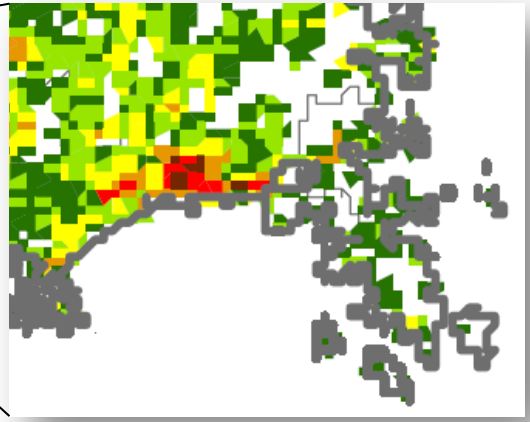
2. Grid Square Statistics of Population Census

Total Population, all Japan at Basic Grid Square level

GPC 2010



A Basis Grid Square is
a square about 1km on a side



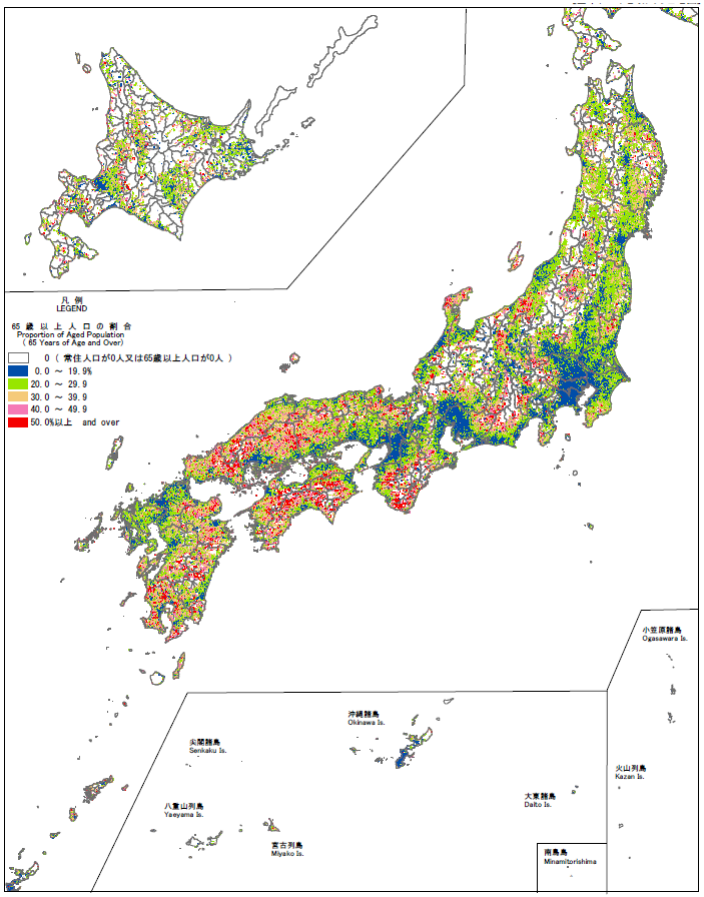
LEGEND

0人	
1 ～	99
100 ～	299
300 ～	999
1,000 ～	2,999
3,000 ～	4,999
5,000人以上	and over

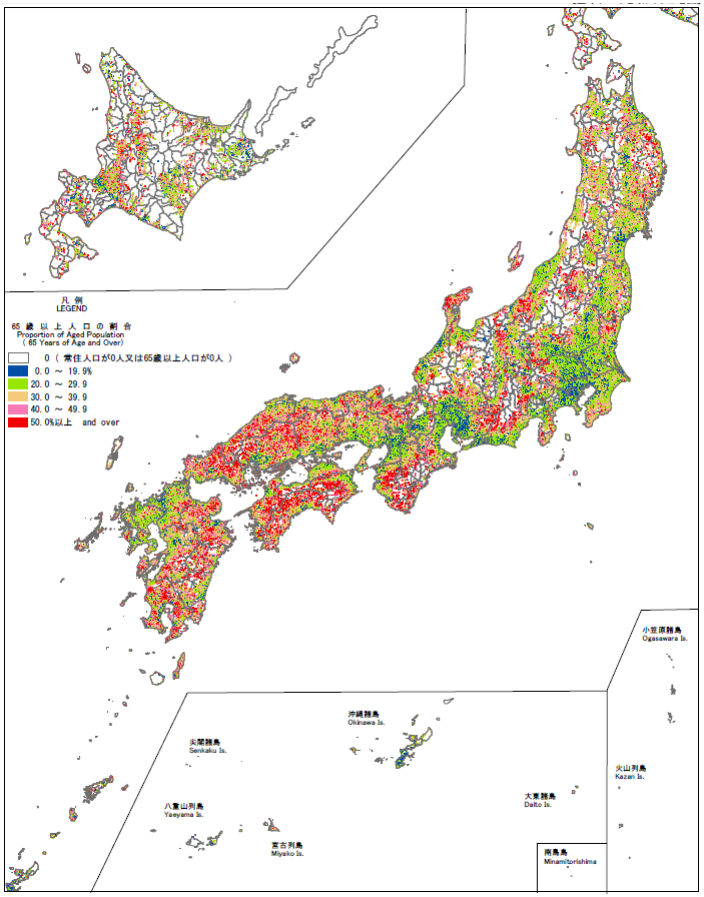
2. Grid Square Statistics of Population Census

Proportion of Aged Population (65 years and older), All Japan
at Basic Grid Square level

GPC 2000



GPC 2010



LEGEND

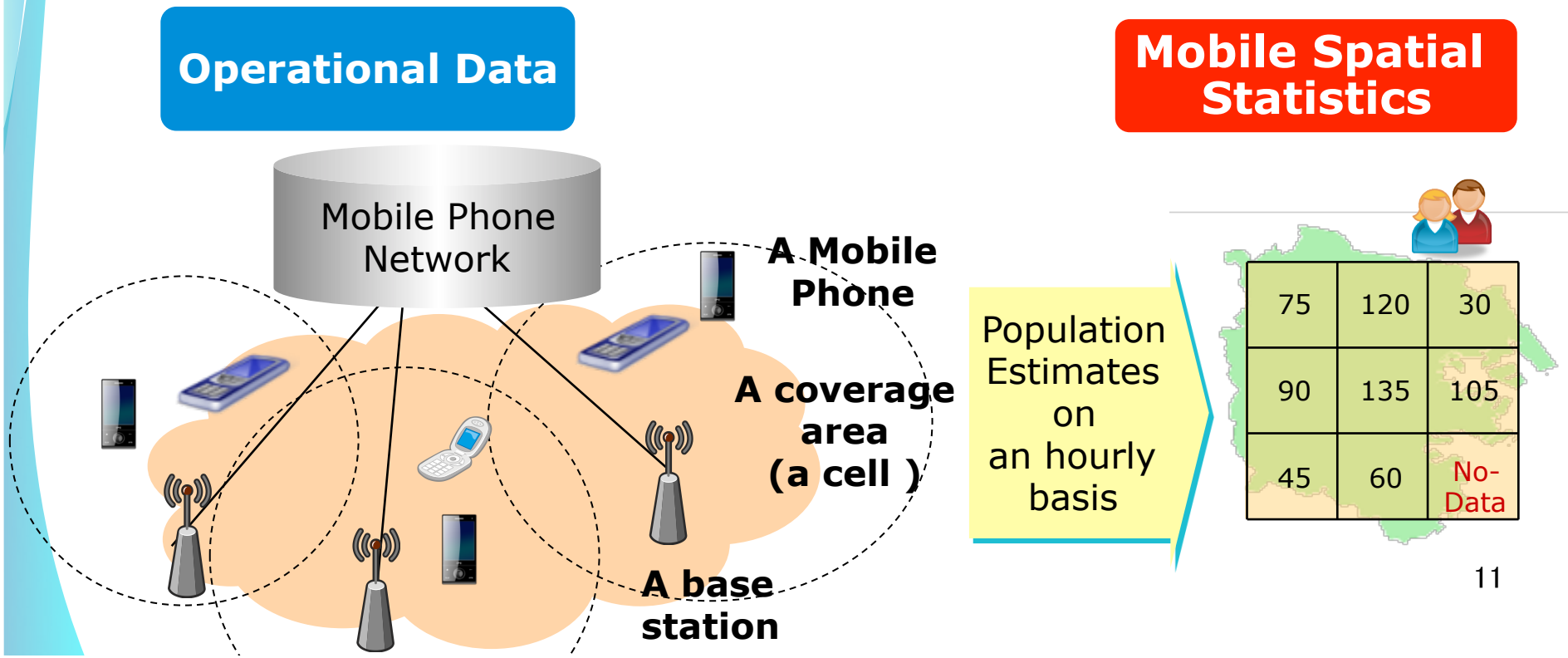
- 0
- 0.0 ～ 19.9%
- 20.0 ～ 29.9
- 30.0 ～ 39.9
- 40.0 ～ 49.9
- 50.0%以上 and over

3. Mobile Spatial Statistics

- NTT DOCOMO
  - The largest service provider of mobile phones in Japan
  - Four out of ten people use DOCOMO's mobile phones.
    - The mobile phone penetration rate ≈ 100%
    - Japanese Population: 128 million
    - DOCOMO has more than 60 million subscribers.

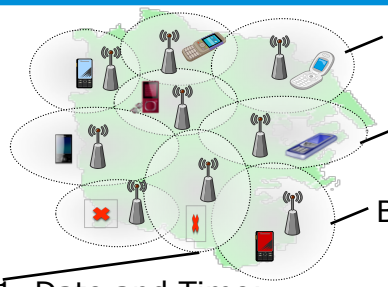
3. Mobile Spatial Statistics

- Mobile Spatial Statistics (MSS)
  - Small area population estimated by DOCOMO derived from the operational data of their mobile network.



3. Mobile Spatial Statistics

Operational Data



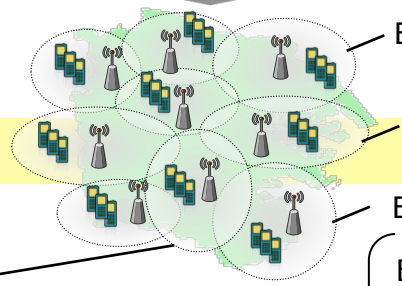
Base station#4:
… …

Base station#3:
… …

Base station#**2**:
… …

Base station#1, Date and Time:
090-XXXX-, ..., Male, Birthday:10/11/1968, 1-2 B Str. Chiyoda Tokyo
090-YYYY-, ..., Female, Birthday: 10/11/1955, 3-4 D Ave. Urayasu Chiba
…

1. De-identification process



BS#4

BS#3

BS#**2**

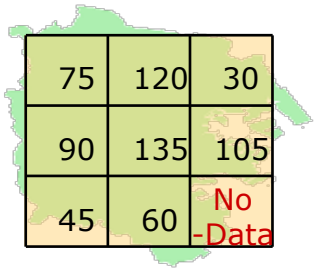
Base station#1, Date and Time:
Male, 40's, A-city Tokyo
Female, 50's, C-city Chiba
…

Mobile phone counts by cell

2. Estimation process

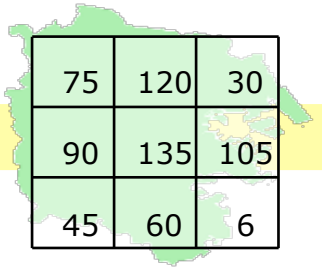
Extrapolate population from Mobile phone counts reflecting the DOCOMO's share in the mobile phone market by sex, age group and district (prefecture)

Mobile Spatial Statistics



By age group and sex, etc.

3. Disclosure Control process



By age group and sex, etc.

Population estimates by grid square

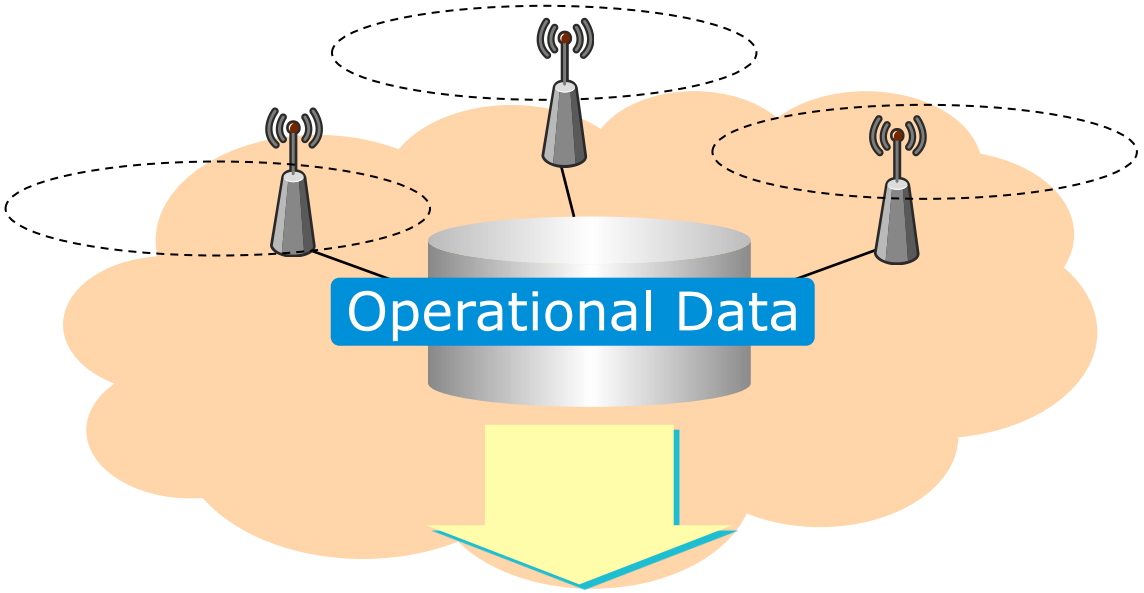
3. Mobile Spatial Statistics

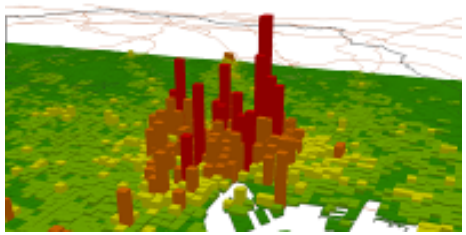
- Privacy protection measures are taken.

  No one can follow a mobile phone
  nor a particular user over time from MSS,
  because:

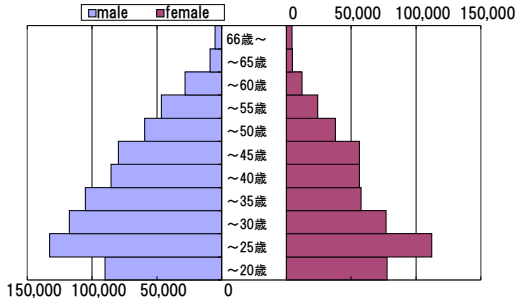
  - Identifiers of mobile phone and personal identification information (name, telephone number, etc.) are removed.
  - Users' dates of birth and residential address are coded in groups.

3. Mobile Spatial Statistics





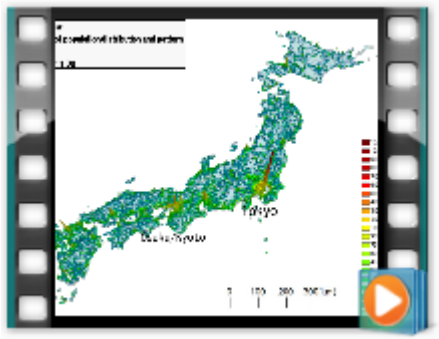
Population Distributions



Population Pyramid

3. Mobile Spatial Statistics

DEMO



3. Mobile Spatial Statistics

- MSS has been in research phase.
  - MSS data has not been publicly available.
  - DOCOMO has formed private partnerships with universities to research the application of MSS, but the validity of MSS had not been examined yet.
  - A joint research project by DOCOMO and NSTAC was launched to compare MSS with GPC.

4. Qualitative difference between GPC and MSS

GPC and MSS look alike,
but there are differences between the two statistics

- 4.1 Characteristics of population
- 4.2 Data source
- 4.3 Coverage of age
- 4.4 Frequency and Timeliness
- 4.5 Cost to collect data

4. Qualitative difference between GPC and MSS

- 4.1 Characteristics of population
  - GPC: *de jure* population
    - An individual's location is recorded to their place of residence on Census day.
  - MSS: *de facto* population
    - An individual's location is recorded to where the person is present at the time of reference

4. Qualitative difference between GPC and MSS

4.1 Characteristics of population (cont.)

GPC
de jure population

People are attributed to their residences on Census day.

Residence

Office, School, etc.

MSS
de facto population

People are attributed to where they are present <u>depending on the time of reference</u>

Residence

Office, School, etc.

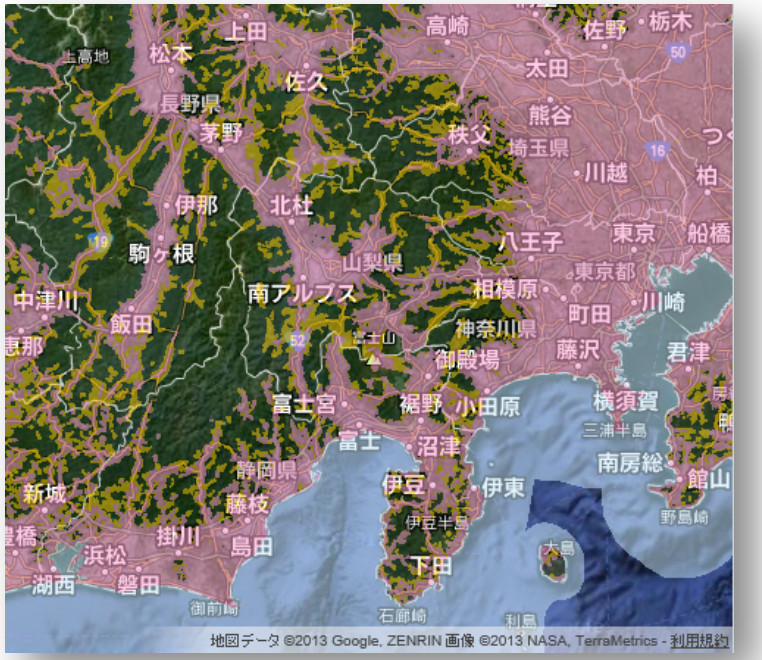
- GPC and MSS are diferrent each other.
- You can not say which one is better than the other.

4. Qualitative difference between GPC and MSS

- 4.2 Data source
  - GPC: Complete enumeration
    - is aggregated from the Population Census that counts all the people all over Japan;
    - can present the population for any given small area.
  - MSS: DOCOMO's mobile phone users
    - based on estimates from DOCOMO users (40% of people in Japan);
    - could be viable only for areas with large enough population;
    - is not applicable for an area with sparse population because the sample of the mobile phone users would be too small to accurately inform estimation.

4. Qualitative difference between GPC and MSS

- 4.2 Data source (conti.)
  - MSS: DOCOMO's mobile phone users
    - The base stations does not completely cover Japan, where more than half of the land area is forest/mountainous zone.



Only red zone is coverage area of the base stations.

4. Qualitative difference between GPC and MSS

- 4.3 Coverage of age
  - GPC: all ages
  - MSS: limited to 15-79 years old
    - Those aged 14 and under and 80 and over are excluded from the estimation because mobile phone penetration rates for these age groups are low, and thus the sample size is too small for MSS to provide accurate estimates.

The population aged 15-79 makes up around 80 percent of the total population (2010 Population Census.)

4. Qualitative difference between GPC and MSS

- 4.4 Frequency and Timeliness
  - GPC is produced every five years
    - It takes around two years for GPC to be released after the Population Census is conducted.
  - MSS could be produced on an hourly basis
    - It would take a few weeks after the reference time to finalize the MSS result ensuring the consistency and confidentiality.

4. Qualitative difference between GPC and MSS

- 4.5 Cost to collect data
  - GPC
    - An enormous budget is required
      for Population Census
      to mobilize enumerators to collect census questionnaires
      and to capture the data
  - MSS
    - The operational data are routinely collected from the mobile phone network

4. Qualitative difference between GPC and MSS

	GPC	MSS
4.1 Characteristics of population	*de jure* population	*de facto* population
4.2 Data source	complete population	DOCOMO users (40% of people)
4.3 Coverage of age	all ages	15-79 years old (generations of active mobile phone users)
4.4 Frequency and Timeliness	not so often and slow	often and quick
4.5 Cost to collect data	enormous	very small (routinely collected from the mobile phone network)

4. Qualitative difference between GPC and MSS

- Lessons
  - You can not easily compare the two data, because GPC and MSS are not equivalent: de jure population and de facto population.
  - MSS is not applicable for an area with sparse population

- Even so, there is a possibility:

  GPC population could be closely aligned to MSS population at night on Census day in densely populated area

4. Qualitative difference between GPC and MSS

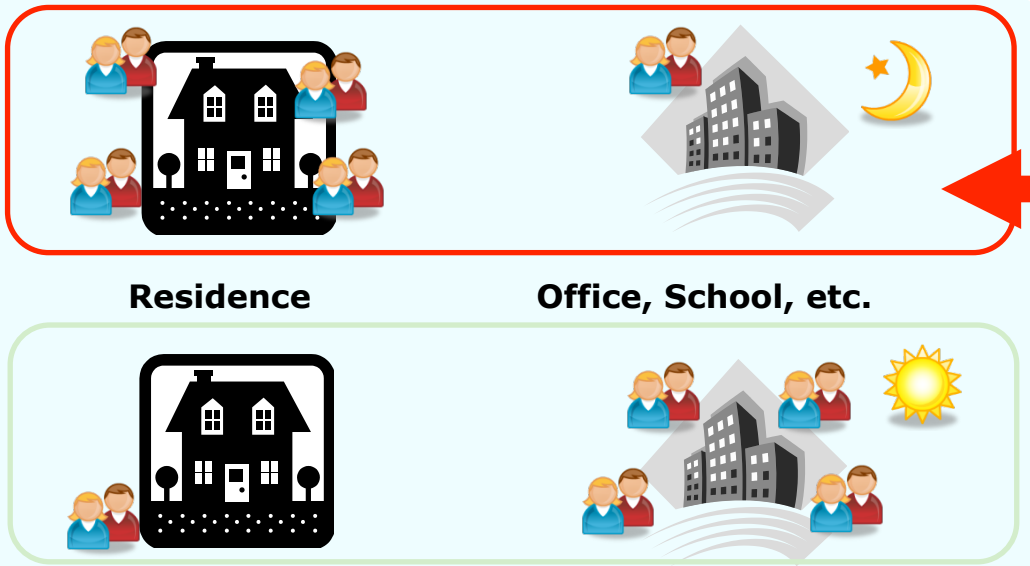
- it would be worth measuring the extent of the closeness of GPC and 'MSS at night.'

GPC
de jure population

People are attributed to their residences on Census day.

MSS
de facto population

People are attributed to where they are present <u>depending on the time of reference</u>



5. Quantitative comparison of GPC and MSS

- To examine the validity of MSS,
  it would be worth measuring the extent of
  the closeness of GPC 2010 population
  and
  MSS population at 4:00 AM on Oct. 1 2010
  (the Census day)
  by calculating '**deviation rate**' $\delta(i)$

5. Quantitative comparison of GPC and MSS

- Definition of '**deviation rate**' $\delta(i)$

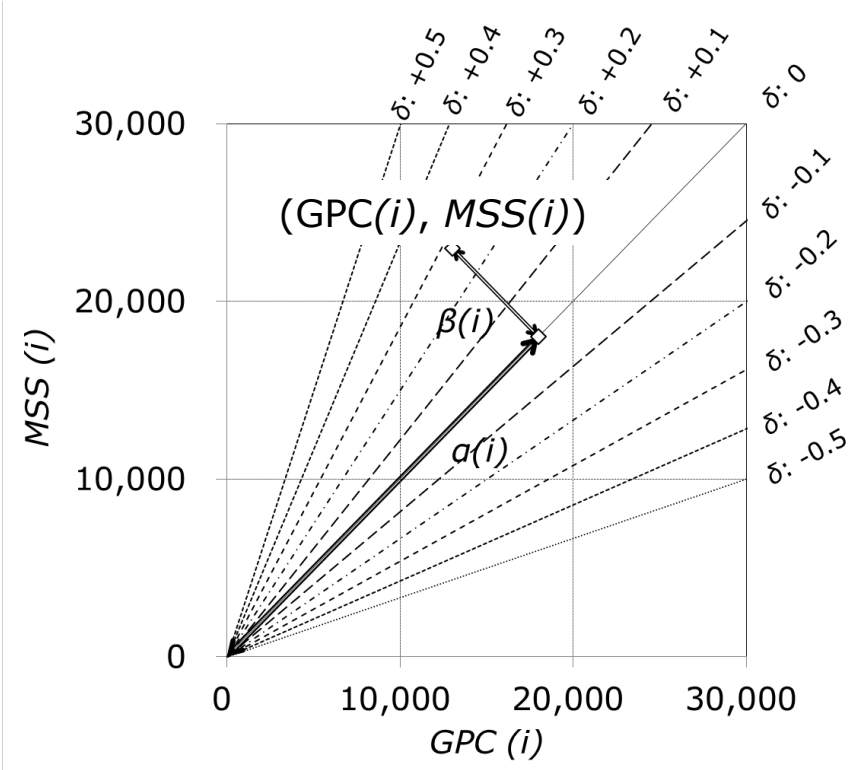
  $MSS(i)$: the MSS population in a grid square i

  $GPC(i)$: the GPC population aged 15-79 in i

  - The deviation rate $\delta(i)$ of the grid square i:
    $\delta(i) = \{MSS(i)-GPC(i)\} / \{MSS(i)+GPC(i)\}$

5. Quantitative comparison of GPC and MSS

- Interval of the deviation rate: $-1 \leq \delta(i) \leq 1$
  - If $\delta(i)$ is close to 0, *MSS(i)* and *GPC(i)* are close to each other, and the point (GPC*(i)*, *MSS(i)*) is close to the 45 degree line.
  - If $\delta(i)>0$, *MSS* > *GPC* and the point (GPC*(i)*, *MSS(i)*) is above the 45 degree line.



$\delta(i) = \{MSS(i) - GPC(i)\} / \{MSS(i)+GPC(i)\} = \beta(i) / a(i)$

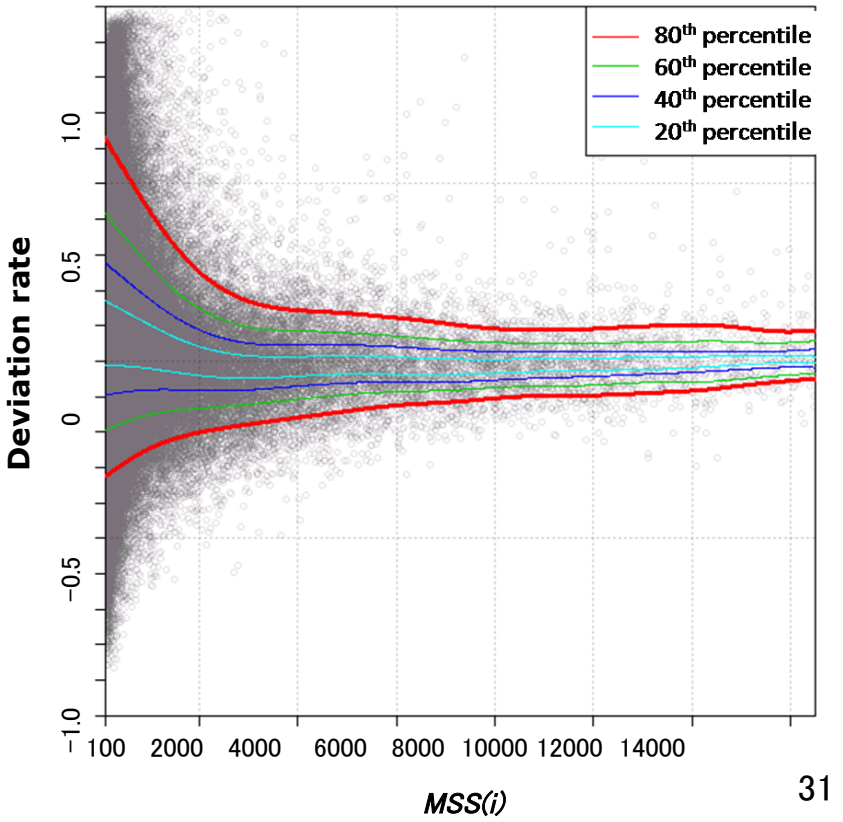
$\beta(i)$: the perpendicular distance from the point (*GPC(i)*, *MSS(i)*) to the 45 degree slope

$a(i)$: the distance between (0,0) and projection of (*GPC(i)*, *MSS(i)*) on to the 45 degree slope.

5. Quantitative comparison of GPC and MSS

- Apprx. 70,000 pairs of *GSS(i)* and *MSS(i)* are inputted to calculate the deviation rates.

A scatter plot of the deviation rates of Basic Grid Squares (apprx. 1km on a side)



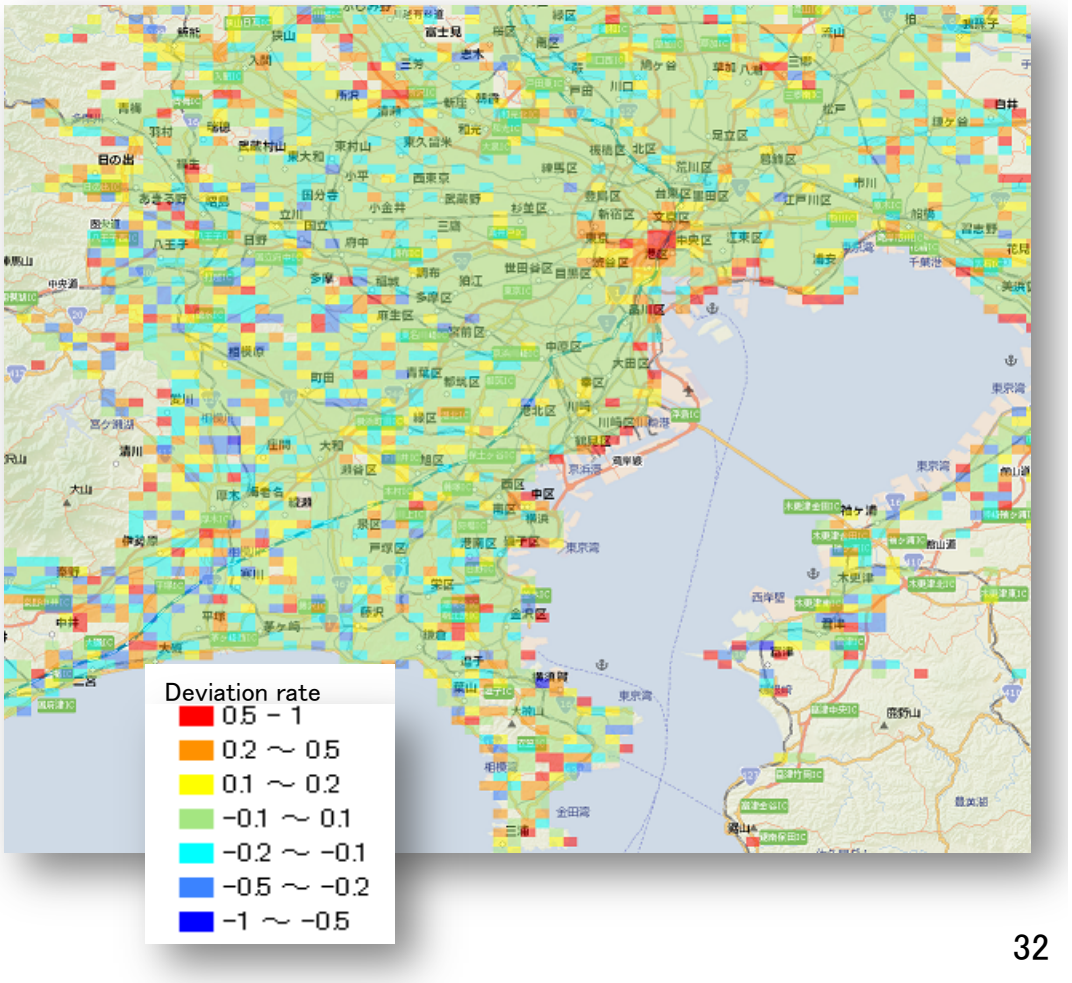
- The fluctuation in deviation rates tend to decrease as the MSS population in the grid square increases

- Above 4,000 of estimated MSS populations, deviation rates become steady.

6. Deviation rate in Tokyo area

- For the most part in Tokyo area, the deviation rate is between +/- 0.1 (in green).
- There remain, however, some areas where the deviation rates exceed this range, mostly beyond +0.1 (in warm color).

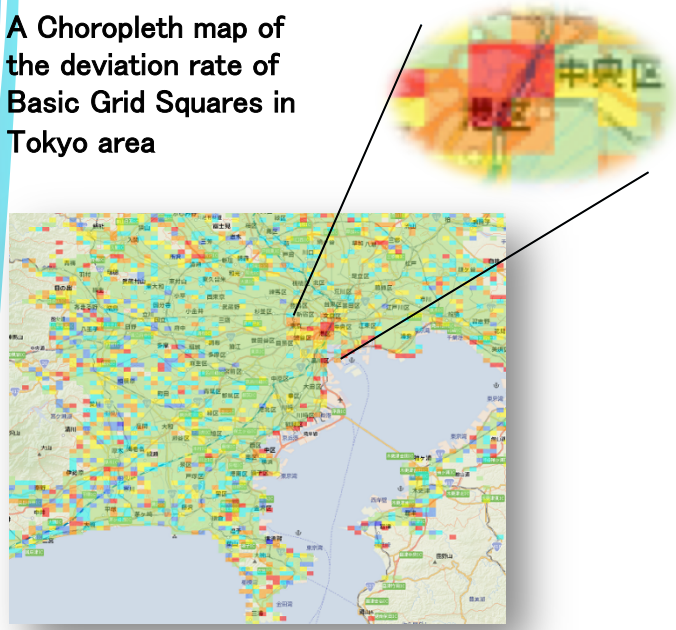
A Choropleth map of the deviation rate of Basic Grid Squares in Tokyo area



6. Deviation rate in Tokyo area

A Choropleth map of the deviation rate of Basic Grid Squares in Tokyo area

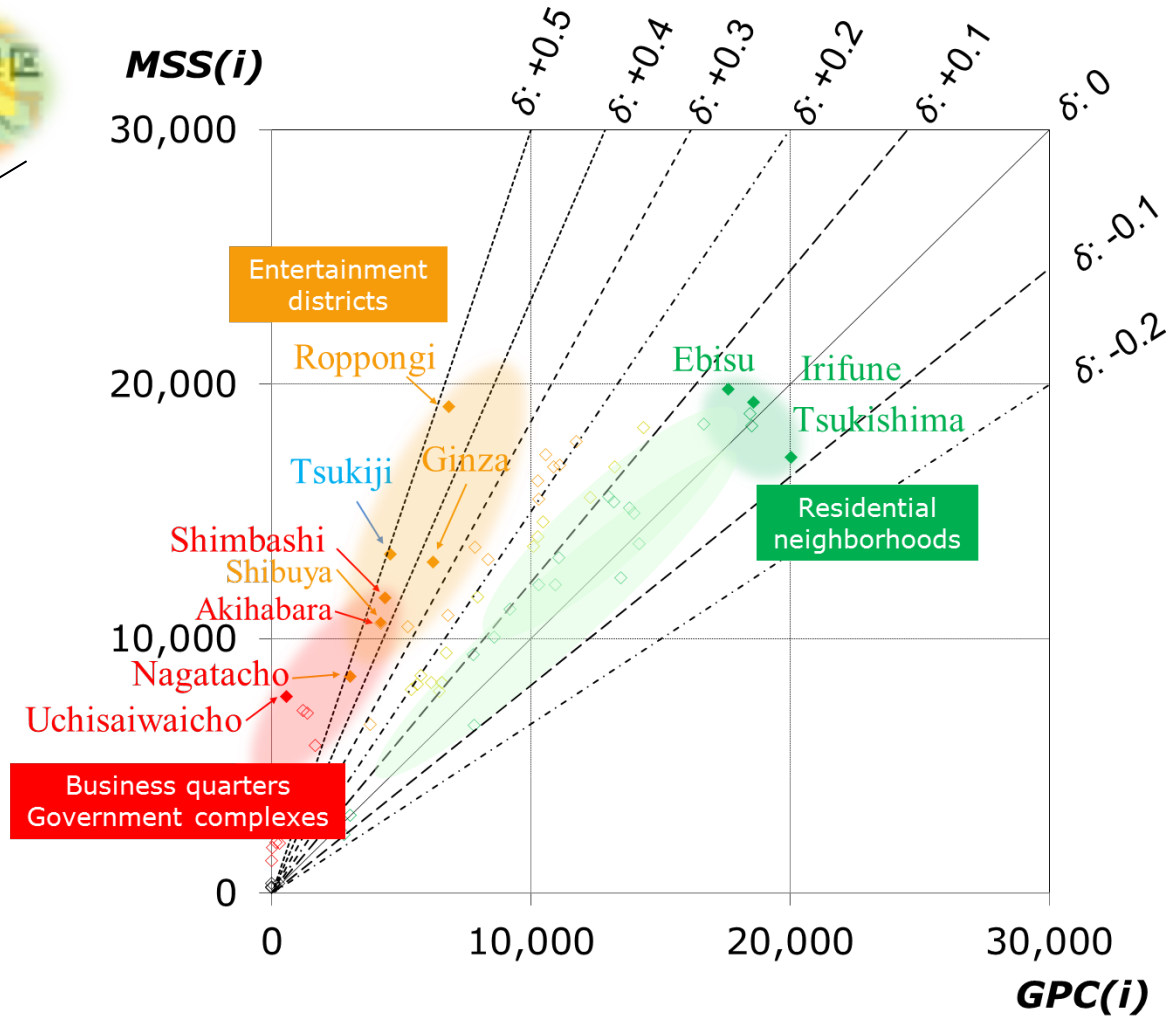
Central Tokyo (Chiyoda, Chuo and Minato Cities)



Deviation rate

- 0.5 – 1
- 0.2 ~ 0.5
- 0.1 ~ 0.2
- -0.1 ~ 0.1
- -0.2 ~ -0.1
- -0.5 ~ -0.2
- -1 ~ -0.5

A scatter plot of Basic Grid Square in central Tokyo



7. Concluding Remarks

- There are differences between MSS and GPC, and limitations in applying MSS to a small area with sparse population.
- Nevertheless, we compared them by calculating the deviation rate in densely populated area, and found that the use of MSS would be plausible to a certain degree.
- MSS was impacted by particular patterns of behavior, as seen in the busy blocks of central Tokyo early in the morning.

Thank you for your attention!

- Views and opinions expressed in this document are those of the authors, and not necessarily those of the organizations which the authors belong to.
- Reference
  - 'Can mobile phone network data be used to estimate small area population?: a comparison from Japan,' Statistical Journal of the IAOS, Volume 29, 2013 (forthcoming)
  - GPC
    - Grid Square Statistics, Statistics Bureau of Japan
      - http://www.stat.go.jp/english/data/mesh/index.htm
  - MSS
    - Technical Journal Number VOL.14 NO.3, NTT DOCOMO
      - http://www.nttdocomo.co.jp/english/corporate/technology/rd/technical_journal/bn/vol14_3/index.html