MSS Population Estimation Statistical Processing

Technology Reports

“Mobile Spatial Statistics” Supporting Development of Society and Industry
—Population Estimation Technology Using Mobile Network Statistical Data and Applications—

Population Estimation Technology for Mobile Spatial Statistics

We give an explanation of the approach to estimation and the estimation methods used for MSS, which is used to make estimations of population using the operations data from a mobile terminal network. When estimating populations from operational data, biases in the estimated values due to characteristics of the mobile terminal network must be taken into consideration. It is also difficult to use estimated population values output for each base station area as is. Accordingly, in tabulating MSS, we eliminate biases in the estimated populations by taking characteristics of the mobile phone network into consideration, and we provide a process to convert output population values to other geographic units, such as grid units, that are easier to use in application fields.

Research Laboratories

Masayuki Terada
Tomohiro Nagata
Motonari Kobayashi

1. Introduction

Mobile Spatial Statistics (MSS) are statistics of the actual population for all of Japan that are generated continuously from mobile terminal network operational data.

MSS are created through a three-step process of (1) de-identification, (2) estimation, and (3) disclosure limitation. In this article, we give an overview of the approach and methods used for estimation, which estimate populations for each geographic area (grid section, administrative division, etc.) from the de-identified operational data.

2. Issues

Intuitively, MSS are created from the operational data in the following manner:

(1) Aggregate the number of mobile terminals present in each of the base station areas (or cells), based on the operational data.

(2) Estimate the total number of mobile terminals in use by extrapolation using the adoption rates of NTT DOCOMO mobile terminals.

(3) The Re-aggregate the per-cell populations obtained from the

previous process into each grid section, municipality or other geographic area.

However, since the mobile terminal network was not designed to generate population statistics, MSS cannot actually be generated from operations data as simply as described above.

For example, the mobile terminal base stations periodically (approximately hourly) check which mobile terminals are present within their communications range (the cell) to enable terminals to be paged. The data gathered from this checking process are referred as to location data. However, the timing for checking the presence of mobile terminals is different for each terminal, and a terminal might move to another cell during the checking period. A terminal may move from one base station area to another as the user moves, but the mobile terminal network cannot know which area the terminal was actually in between one checking time and the next. Further, in a real mobile terminal network, the intervals between checking are not fixed and can fluctuate significantly according to the movements and usage of the terminal. Consequently, the mobile terminal network cannot directly know how many mobile terminals are within a cell at a given time, so this value must be estimated from the results of each base station checking the presence of mobile terminals by an appropriate statistical means.

Also, even if the number of mobile terminals within a cell (the phones present) could be estimated exactly, it would, of course, be different from the population value, which includes people not using NTT DOCOMO. Superficially, if the adoption rate of NTT DOCOMO telephones nationally is known, simply dividing the number of terminals present by the adoption rate can give an estimation of the population. For example, suppose the population of Japan is 120 million, and there are 60 million contracts for NTT DOCOMO mobile terminal services. Then the adoption rate would be one terminal for every two people, or 0.5, and dividing the number of terminals present in a base station area by 0.5 (multiplying by two) would calculate a population value for that area. However, the actual adoption rates for NTT DOCOMO telephones vary with region, age-group and gender, so this sort of simple calculation produces large biases according to region, age-group and gender, and is not suitable for population statistics. Further, mobile terminals are not always turned on, and the mobile terminal network cannot know the location of such terminals. Thus, it is necessary to take these factors into consideration in order to extrapolate appropriately from the number of terminals present to an estimation of the population.

Finally, even after these issues have been resolved and the population within each cell can be estimated accurately, it is difficult to use these population statistics in this form. Even knowing that there are Y people within cell X, there is no way to use this population data without knowing where cell X is and how large it is; the cell area covered by each base station varies depending on the station type, which frequencies it uses, surrounding geographic configuration and other factors. Accordingly, a process is needed for converting the estimated per-cell population values into appropriate geographical units, such as grids or administrative-area (cities, towns, etc.).

Summarizing the above discussion, estimation is composed of the following three steps.

1) Estimation of Terminals Present

Estimates the number of mobile terminals in each cell based on the de-identified operational data, taking the variability of operational data into consideration.

2) Extrapolation of Populations

Extrapolates the population from the per-cell numbers of mobile terminals resulting from the previous step, taking into consideration factors such as biases in the adoption rates of NTT DOCOMO mobile terminals and the effects of mobile terminals that are not present in the mobile network (e.g.: turned off or out of service area).

3) Area Conversion

Converts per-cell population estimated in the previous step into appropriate geographic units that are easy to use in application fields, such as grid sections or administrative regions, considering the coverage area of each cell.

3. Estimation

We now describe the concepts and estimation methods for each of the steps: estimating terminals present, extrapolated population estimations and area conversion.

3.1 Estimating Terminals Present

The purpose of estimating terminals present is to estimate the number of mobile terminals within each cell over a set period of time, based on the de-identified operational data.

As described earlier, the mobile terminal network does not accurately know how many mobile terminals are present for each base station at a given time. Thus, based on the location data[*1] gathered during each set time period, such as one hour, (hereinafter referred to as "the observation period, T"), the average number of terminals present for that time period is calculated. The average terminals present for cell C during the observation period, T, is designated m_C^T.

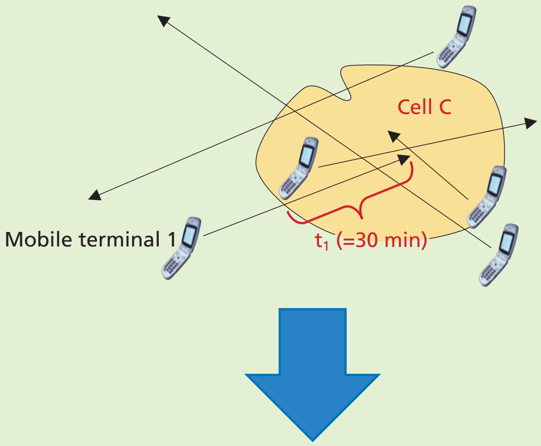
The number of terminals present, m_C^T, can be calculated as follows, where the amount of time that mobile terminal i is within base station area C during observation period T is given by t_{iC}^T (**Figure 1**).

$$m_C^T = \Sigma_i \, t_{iC}^T \ / \ |T| \qquad (1)$$

Here $|T|$ represents the length of observation period T.

However, as mentioned earlier, the mobile terminal network does not know where each mobile terminal is all of the time, so we cannot know the times that each mobile terminal is in the area, t_{iC}^T, directly. Thus, these have to be estimat-

- Suppose that five mobile terminals pass through or stay within cell C in a given hour.
- And t_1 to t_5 are the times that mobile terminals 1 to 5 spend within cell C during this hour, respectively.
- Specifically, mobile terminal 1 was within C for t_1=30 minutes.



The number of terminals located in C is estimated as follows:
During this 1 hr, each mobile terminal i was in the area t_i minutes.
⇒At a point in time within the above mentioned hour, the probability that mobile terminal i was within area C is $t_i/60$.
(i.e. for mobile terminal 1, the probability is 30/60 = 1/2).
⇒Thus, the expected value for the number of mobile terminals within area C is
$t_1/60+t_2/60+\ldots\ t_5/60 = \Sigma\, t_i/60$.

Figure 1 Estimating the number of terminals in an area

[*1] **Collected location data**: More accurately, this refers de-identified operational data, which is location data combined with corresponding attribute data, with information such as names and phone numbers that can identify individuals removed, and attributes such as addresses and birthdays rounded to age groups and residential area codes by the de-identication process. To simplify the explanation, we refer to it as location data hereinafter.

ed statistically from the location data.

Location data is generated when checking the presence of mobile terminals as mentioned before. If the interval of the checks is fixed to some time period, w, then as the sum of $t_{i\,\mathrm{C}}^{\mathrm{T}}$ increases by w, the number of location data observed in C is probabilistically expected to increase by one. Thus, $\Sigma_i t_{i\,\mathrm{C}}^{\mathrm{T}}$ can be estimated by Equation (2), where $L_{\mathrm{C}}^{\mathrm{T}}$ is the set of location data observed in cell C during observation period T, and $|L_{\mathrm{C}}^{\mathrm{T}}|$ denotes the number of element of $L_{\mathrm{C}}^{\mathrm{T}}$ (i.e., the number of location data observed in C during T).

$$\mathrm{E}\,(\Sigma_i t_{i\mathrm{C}}^{\mathrm{T}}) = w\,|L_{\mathrm{C}}^{\mathrm{T}}| \qquad (2)$$

However, as discussed in chapter 2, the presence checking interval on a real mobile phone network fluctuates with various factors such as the usage and movement of the mobile phone, so w changes with each location data (the density of location data varies in time). To take this variance in the time density of location data into consideration, we apply a feature value (a weighting) to each location data point, according to its surrounding signal intervals. The feature value, w_j, for location data point, j, is typically the average of the intervals for the previous and next location data point. Consequently, $\Sigma_i\, t_{i\,\mathrm{C}}^{\mathrm{T}}$ can be estimated using w_j by Equation (3), regardless the variance of the intervals of presence checking.

$$\mathrm{E}\,(\Sigma_i t_{i\mathrm{C}}^{\mathrm{T}}) = \Sigma_{j \in L_{\mathrm{C}}^{\mathrm{T}}} w_j \qquad (3)$$

Thus, $m_{\mathrm{C}}^{\mathrm{T}}$, the number of mobile phones located in cell C during observation period T, can be estimated from $L_{\mathrm{C}}^{\mathrm{T}}$, the set of location data points observed in cell C during observation period T and their feature values, by Equation (4).

$$\mathrm{E}\,(m_{\mathrm{C}}^{\mathrm{T}}) = \Sigma_{j \in L_{\mathrm{C}}^{\mathrm{T}}} w_j \;/\; |\mathrm{T}| \qquad (4)$$

In the discussion so far, we have not referred to estimating populations for attributes (age groups, gender, residential area), which are needed for estimations of population composition. However, simple modifications of the above method can give the number of phones present for each of these attributes.

Let some combination of attribute conditions (age group, gender, residential area) be called A. As an example, for males in their 30's living in Tokyo, A = (30's, male, Tokyo). Then, the number of mobile terminals whose owner's attributes satisfy attribute condition A, $m_{\mathrm{C,A}}^{\mathrm{T}}$, is similarly estimated by Equation (5), where $L_{\mathrm{C,A}}^{\mathrm{T}}$ denotes a subset of $L_{\mathrm{C}}^{\mathrm{T}}$, the set of location data corresponding to the mobile terminals whose owner's attributes satisfy A.

$$\mathrm{E}\,(m_{\mathrm{C,A}}^{\mathrm{T}}) = \Sigma_{j \in L_{\mathrm{C,A}}^{\mathrm{T}}} w_j \;/\; |\mathrm{T}| \qquad (5)$$

3.2 Extrapolating Populations

This step extrapolates populations from the number of terminals estimated in the previous step and residential populations such as those from a census.

As described earlier, if the adoption of NTT DOCOMO mobile terminals in the population was uniform regardless age-group or gender, and all mobile terminals were turned on and were actually present within the area of one of the base stations, the population could be estimated by simply multiplying the number of terminals by the inverse of the adoption rate.

However, the adoption rate of NTT DOCOMO terminals actually varies significantly with age-group, gender and residential area. If these variations in the adoption rates are not applied correctly, intolerable biases in the estimated populations can result. For example, the estimated population would be too high in areas where the actual adoption rate was higher than assumed, and would be too low where the actual adoption rate was lower.

Also, if people turn off their mobile terminals, when they go to sleep for example, no interaction can be performed between the mobile terminal and the network during that time, so those terminals do not appear as terminals present in any cell. Thus, if the extrapolation multiplier (expansion rate) is determined simply, based on the number of contracts, the estimated populations will be lower than the actual populations, particularly during times when the rate of presence is low (e.g., when there are many terminals turned off).

In order to reflect the adoption rates correctly as they vary with attributes including gender, age-group and residence area, extrapolated estimates must be handled for each attribute. Also, when calculating extrapolation rates, the adoption rates cannot simply be used as is, but must be adjusted using a presence rate that reflects the state of presence with a mobile terminal. The process for extrapolating estimations is illustrated in **Figure 2**.

Specifically, in order to reflect biases in adoption rates correctly for each attribute, public statistics such as the national census and the basic residency register are used to collect residential populations for each attribute. Let r_A be the residential population for attributes A. At the time of writing, partitioning by attributes was done in five-year age groups, by gender, and by prefecture-level geographic boundaries. Next, to eliminate the effects of terminals that are turned off or otherwise not-present, the estimated total number of terminals present during the observation time T was calculated for each attribute (collected from all base station areas throughout Japan). The total terminals present nationally during observation period T, $m^T_{*,A}$, can be estimated by Equation (6).

$$E(m^T_{*,A}) = \Sigma_C E(m^T_{C,A}) \qquad (6)$$

Then, the number of present mobile terminals per person with attributes A during the observation time period T, in other words, presence rate k^T_A, is given by Equation (7).

$$k^T_A = E(m^T_{*,A}) / r_A \qquad (7)$$

And the population with attributes A within cell C during observation time period T, or $p^T_{C,A}$, can be estimated by Equation (8).

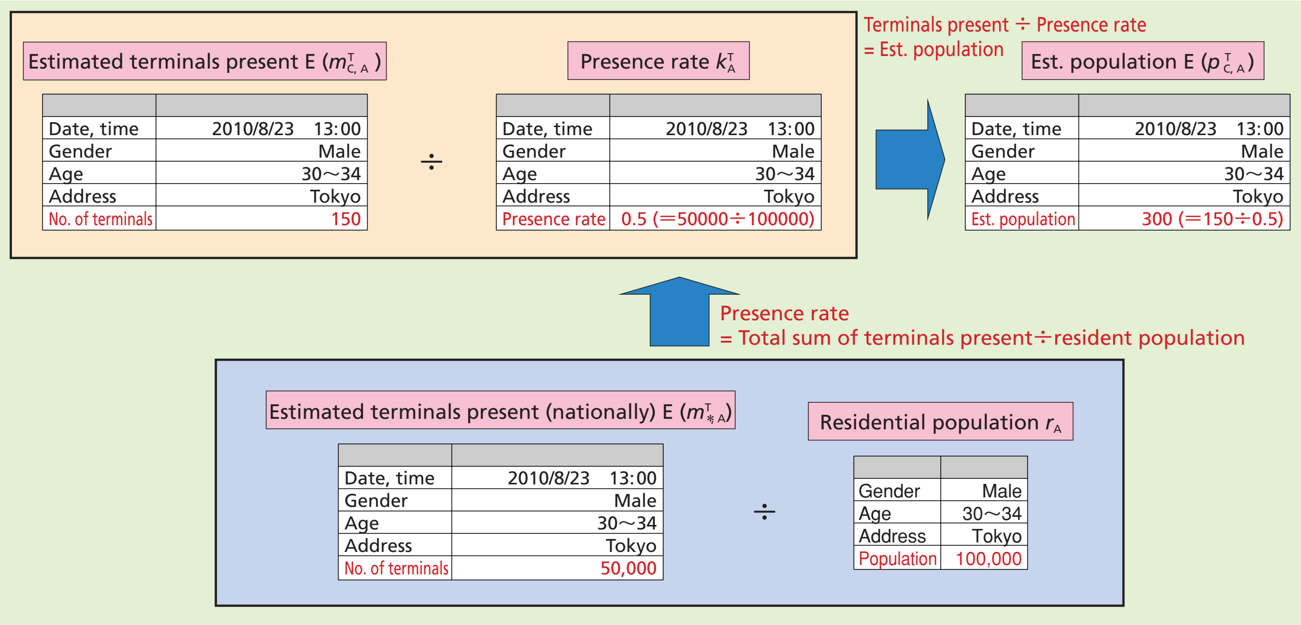
$$E(p^T_{C,A}) = E(m^T_{C,A}) / k^T_A \qquad (8)$$

The total population in cell C during T regardless of attribute values, p^T_C, is simply calculated as the total sum of the populations with attributes.

In this way, the population is extrapolated from the estimated number of mobile terminals without biases in adoption rates of NTT DOCOMO mobile terminals by gender, age-group and region, as well as fluctuations in mobile terminals present due to terminal power status.

3.3 Area Conversion

Area conversion converts the populations estimated for each cell in the extrapolation process into populations



Estimated terminals present $E(m^T_{C,A})$

Date, time	2010/8/23 13:00
Gender	Male
Age	30～34
Address	Tokyo
No. of terminals	150

÷

Presence rate k^T_A

Date, time	2010/8/23 13:00
Gender	Male
Age	30～34
Address	Tokyo
Presence rate	0.5 (＝50000÷100000)

Terminals present ÷ Presence rate = Est. population

Est. population $E(p^T_{C,A})$

Date, time	2010/8/23 13:00
Gender	Male
Age	30～34
Address	Tokyo
Est. population	300 (＝150÷0.5)

Presence rate = Total sum of terminals present÷resident population

Estimated terminals present (nationally) $E(m^T_{*,A})$

Date, time	2010/8/23 13:00
Gender	Male
Age	30～34
Address	Tokyo
No. of terminals	50,000

÷

Residential population r_A

Gender	Male
Age	30～34
Address	Tokyo
Population	100,000

Figure 2 Extrapolated estimation

in geographic units more commonly used in various types of analysis application, such as grids of 500 m, 1 km, or other sizes, or administrative boundaries such as cities and towns, as shown in **Figure 3**.

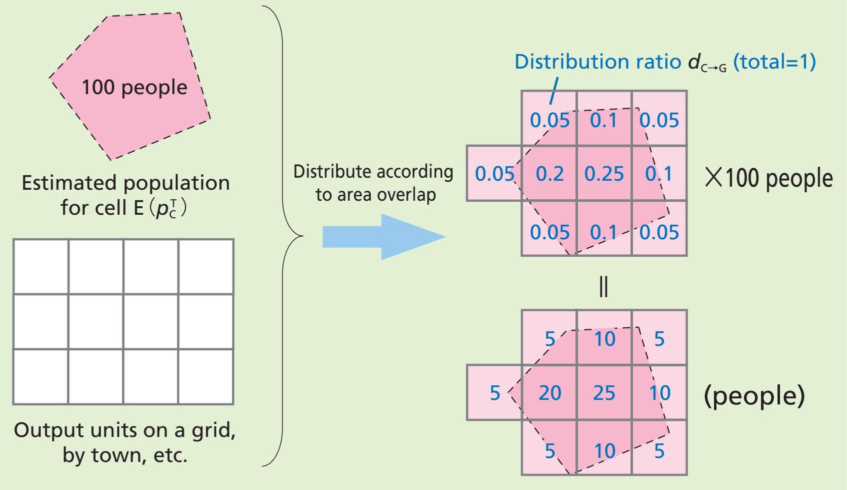

100 people
Estimated population for cell E (p_C^T)
Output units on a grid, by town, etc.
Distribute according to area overlap
Distribution ratio $d_{C\to G}$ (total=1)
0.05 0.1 0.05
0.05 0.2 0.25 0.1
0.05 0.1 0.05
×100 people
||
5 10 5
5 20 25 10
5 10 5
(people)

Figure 3 Computing area conversions

In the area conversion process, the populations estimated for each cell are distributed over a grid or other geographic unit, according to the size and shape of the actual cell area. The desired population value, p_G^T, for geographic unit G during observation period T, is estimated by the following.

$$E(p_G^T) = \Sigma_C d_{C\to G} E(p_C^T) \quad (9)$$

Here, $d_{C\to G}$ is the population distribution ratio from cell C to the geographic unit G, and the sum of these values for each cell is 1 (one). In other words, $\Sigma_C d_{C\to G} = 1$.

The distribution ratios, $d_{C\to G}$, are calculated based on the size of the geographical overlap between the cell area and each of the geographic units (proportional division by area). Let the area of base station area C be |C|, and the area of geographical overlap between cell C and geographic unit G be |C ∩ G|. Then the distribution ratio, $d_{C\to G}$, can be calculated by the following equation.

$$d_{C\to G} = \frac{|C\cap G|}{|C|} \quad (10)$$

In this way, populations estimated according to cell can be converted to populations in commonly used geographic units, such as grids, which are easier to use in various analysis applications.

4. Conclusion

In this article, we have described some issues with the estimation process, which is a core part of estimating populations with MSS. We have also described estimation methods to resolve these issues.

In fact, the actual estimation process of MSS also includes various types of corrections to increase the reliability of the resulting estimates such as correction of estimation errors caused by network faults or traffic restrictions, and sanitization of operational data to prevent the estimates from being biased by outliers.

As of the time of writing this article, these corrective measures continue to be evaluated and improved, and we continue research and development, carefully examining fields of application to provide MSS that are more reliable and sufficient to support even more important decision making in these fields.