A Block-level Estimation of Residential Characteristics Using Survey and Spatial Microdata

Masaya UESUGI*, Yasushi ASAMI**

Abstract

A wide variety of microdata offer great potential for the advancement of small area demographic estimation. This study proposes a method to estimate residents' attributes at the street block level using survey microdata and spatial microdata (GIS building data), demonstrating the method in a Tokyo ward area. The proposed method is shown to be superior to alternatives using limited microdata. This method includes (1) modeling housing composition for each block based on dwelling unit information from spatial microdata, (2) establishing the statistical relationship between housing and demographic attributes through a small area cross tabulation technique using survey microdata, and (3) allocating demographic attributes to street blocks based on this small area-specific relationship. Housing composition is defined by building type and dwelling size and is constrained by small area totals of the census. The Tokyo ward area contains high-density built-up districts and is characterized by vertical population extension and resident diversity. In spite of this complex residential structure, the accuracy assessment indicates that the proposed method performs well and presents a more accurate estimation than any other methods using limited microdata. This finding stresses the effectiveness of combining different types of microdata, which not only allows spatial refinement but also reflects neighborhood heterogeneity of the relationship between housing and demography.

Keywords: survey microdata; spatial microdata; street block; small area demographic estimation

1. Introduction

Population characteristics of small geographic units such as blocks or sub-blocks are important in both practical and theoretical terms. For example, estimating the precise target population for a facility or transportation service zone is imperative for urban planning and business. However, the lack of fine-resolution information about the population may lead to estimation errors because these zones do not necessarily match the administrative unit used by the population census. Academics also recognize the modifiable areal unit problem (MAUP) as a serious issue in spatial analysis; the manner in which spatial phenomena are aggregated influences the analytical result.¹⁾ In most cases, small area census data released to the public do not appear at a block or sub-block level and are often restricted to one-dimensional variables owing to the need to preserve privacy and confidentiality. Therefore, there is an urgent need for a reliable and applicable estimation approach for multi-dimensional demographic attributes on a finer scale.

* Ritsumeikan University / JSPS, E-mail: m-uesugi@gst.ritsumei.ac.jp

** The University of Tokyo

Thus, many recent studies have attempted to conduct small area population estimations using various types of microdata. We divide these microdata into two major types: survey and spatial microdata. We refer to a set of individual records of persons or households containing their attributes collected through some kind of public survey, census, or bespoke questionnaire as survey microdata. These data typically contain rich attributes of the population, but the spatial resolution is relatively low for the purposes of protecting privacy. On the other hand, spatial microdata are defined in this study as the individual data of buildings or housing units with their location information at a fine scale. These data are typically derived from GIS data, including remote sensing data or digital topographical maps, but provide little attribute information.

This article proposes a method to estimate residential characteristics at the street block level by incorporating survey and spatial microdata, and illustrates the superiority of the proposed method over alternatives using limited microdata. Although this study demonstrates the estimation for household size as residents' attributes at the street block level in Tokyo ward area because of verifiability issues, the potential of its application to different contexts (including other countries) and other residential characteristics (such as household income and householder age) strongly associated with housing attributes is also discussed. Using such microdata is notable in a metropolitan area, which traditionally falls under the radar of most researchers; it is challenging to conduct a finer estimation for a vertical population distribution due to the concentration of high-rise buildings and population diversity, which is a combination of complex demographic attributes.

2. Literature Review

2.1. Estimation using survey microdata

The approach using the first type of microdata, survey microdata, contains cross tabulation at the small area level and the generation of synthetic spatial microdata.²⁾³⁾ For the cross tabulation approach, we use the iterative proportional fitting (IPF) technique to impute a joint distribution consistent with the known marginal distribution at the small area level, based on the cross tabulation of survey microdata samples for a larger geographical area representing the relationship between two variables. The IPF method, originally developed by Deming and Stephan⁴⁾, has proved to be an influential approach for small area estimation in the field of geography.⁵⁾⁶⁾⁷⁾

Conversely, synthetic spatial microdata are generated by weighting individual or household records from survey microdata. This approach, including generalized regression weighting and the combination optimization procedure, is known as the reweighting method and is often conducted as part of spatial micro-simulation.⁸⁾⁹⁾ The approach adds a weight to each individual or household record based on microdata imputing and adjusts the weight until the aggregation is consistent with the known marginal distribution at the small area level. It is also possible to generate synthetic spatial microdata by adding the attributes from multi-dimensional cross tables estimated by the IPF procedure into individual or household records through the Monte Carlo method,¹⁰⁾ which is referred to as synthetic reconstruction.

These two approaches have both advantages and limitations. The advantage of cross tabulation is that it reproduces accurately known marginal distributions. This is in contrast to the generation of synthetic spatial microdata, which involves a probabilistic simulation and random allocation process. The cross tabulation approach provides a deterministic estimation using an explicit algorithm. Conversely, the generation of synthetic spatial microdata using the reweighting method can retain all attributes that the microdata originally contained, while the dimensional extension

is limited in the cross tabulation approach. However, both approaches use only survey microdata and thus have limitations with regard to spatial disaggregation because the spatial resolution is confined to the zone level at which the constraint tables are available.

2.2. Estimation using spatial microdata

In order to estimate the population at a finer scale, such as at the street block level, spatial microdata, including building polygon and address point, are needed. Notably, information related to housing characteristics is important for population estimation at fine resolutions. Wu et al.'s¹¹⁾ review of population estimation methods using GIS and remote sensing indicate two approaches of possible application with regard to building and housing location data: areal interpolation methods with ancillary information and statistical modeling based on the relationships between populations and dwelling units. Especially, the most common method in the first category is dasymetric mapping.¹²⁾ The process disaggregates the census population into a spatially finer area with homogeneous density partitioned by additional data sources, the so-called ancillary data including postal address point data¹³⁾ and open-access building data.¹⁴⁾

Detailed building information makes an important contribution to these methods. Compared with building area data derived from building footprints, three-dimensional building information (i.e., building volume) is significantly useful and allows the source population to be assigned to each building, some of which are modeled as a framework¹⁵⁾ and can actually be demonstrated.¹⁶⁾¹⁷⁾¹⁸⁾¹⁹⁾²⁰⁾ Moreover, introducing building use (e.g., residential or non-residential) and type (e.g., multi-family or single-family) improves accuracy of estimation.²¹⁾²²⁾²³⁾ At a finer level, Jia et al.²²⁾ discussed the importance of focusing on dwelling units, since this unit is the finest and the most meaningful basic spatial unit corresponding to collection of household information.¹⁵⁾¹⁹⁾

Nonetheless, there are possibilities of further development. First, with some exceptions, most studies using these approaches estimated only the population count and ignored socio-demographic attributes such as race and age.²⁴⁾ Leyk et al.²⁵⁾ allocated a microdata-based census population with multi-attributes to a sub-tract level using maximum entropy dasymetric modeling. Although they used no building or dwelling unit data, their argument about combining the idea of spatial allocation and area interpolation (which had been studied separately in the previous literature) is suggestive. Moreover, the combination of census housing variables having a strong association with building type allows certain advancements.²⁰⁾²⁶⁾ As another example, Akiyama et al.²⁷⁾ allocated census population to individual housing units using detailed building information, and stochastically added the population attributes in a sequential way according to the housing unit floor space, but the assignment of demographic attributes was conducted based on municipality data, which do not reflect neighborhood variation.

Thus, the second task to be tackled is the spatial heterogeneity of the relationship between target zone variables and demographic characteristics within an area. The spatial variation of this relationship is treated by a limited approach including geographically weighted regression modeling,²¹⁾²⁸⁾ iterative adjustment process for each building¹⁹⁾ and spatial allocation with census tract-specific weighting.²⁵⁾ Additionally, the regression approach does not necessarily preserve the pycnophylactic property, wherein the total population does not change after allocation.²⁹⁾

2.3. Summary

A comprehensive literature review shows that survey microdata offer a strong benefit in small

Figure 1. Tokyo ward area, central district of the greater Tokyo metropolitan area, Japan: (a) land use and (b) 23 wards constituting the district.

Source: Land use mesh data for urban areas (2007), National Land Numerical Information, and Ministry of Land, Infrastructure, Transport and Tourism.

area estimations with multiple demographic attributes and in spatial disaggregation in areas for which constraint tables are available. Spatial microdata, especially GIS building data, display great ability to refine the resolution. Hence, incorporating the small area estimation approach using these two types of microdata overcomes the shortcomings of each. Nevertheless, few studies have used both types of microdata in combination for small area estimation and evaluating their effectiveness, especially for high-density urban areas. This study contributes to the limited literature on small area demographic estimation by integrating these different kinds of microdata.

Regarding survey microdata, we adopt the small area cross tabulation approach using the IPF technique so as to completely reproduce the known marginal distribution. The IPF method provides reliable small area estimation even in the Japanese context³⁰⁾³¹⁾³²⁾ and its applicability to the housing variables listed in the population census of Japan has already been demonstrated.³³⁾ Establishing the statistical relationship between housing and demography for small area data reflects the differences between neighborhoods as well as population heterogeneity. Regarding spatial microdata, we explicitly consider residential dwelling units, which allow a finer and more appropriate estimation at the household level. Individual building and housing unit information including the height, number of floors, and use is indispensable for advanced high-resolution estimation.

3. Data

The Tokyo ward area, located at the center of the greater Tokyo metropolitan area in Japan, was selected as the study area. This area comprises 23 wards and is a high-density built-up area with a high rate of apartment residences and considerable mixed-use buildings in business, commercial, and residential districts (Figure 1). According to the 2010 population census, the number of households and residents in this area is 4,540,746 and 8,945,695 within 622 km², respectively.

Census data (small area statistics)

The latest population census of Japan (2010) provides basic statistics and small area statistics, and is available from the website of the Statistics Bureau, Ministry of Internal Affairs and Communications. The basic statistics summary tables provide some household and population attributes including two-dimensional cross tables for the whole country, prefectures, and municipalities, whereas the small area statistics summary tables provide limited variables, usually univariate, such as the number of households by type of dwelling, dwelling unit floor space, household size, and others for *cho-aza* districts (blocks and other sub-divisions of municipalities; hereafter referred to as *cho* districts).

In the Tokyo ward area, the average number of blocks per *cho* district is about 30, and each block has about 80 residents on average, although there are, of course, population variations. For example, a block including a huge apartment building may have an extremely large population within it. A *cho* district, a kind of administrative unit, is the minimum unit used to represent population with attributes in the population census (i.e., small area statistics), the boundary being defined by geographical conditions or historical background. Considering that the average number of residents per *cho* district in the study area is less than 3,000, a *cho* district roughly corresponds to a census block group in the U.S., which is located between a census tract and a census block in terms of population size, despite *cho* districts being much smaller than the census tract in terms of geographical area.

Spatial microdata

Disaggregated data relevant to individual buildings are obtained from two data sources. The first source is a digital residential map called Zmap-Town II (2008/09) provided by a private map-making company, Zenrin Co., Ltd. This nationwide large-scale map gives the name as well as the address of each building and residence investigated by field surveyors house-by-house and is updated every one to five years. Therefore, there is no concern about maintaining privacy, and it is commercially available. The biggest advantage of this detailed map is that it has unique appended tables which include building information on the number and use of rooms in apartment buildings and shops in multi-tenanted buildings. In addition, the digital data format facilitates data processing and provides additional information such as the building footprint from the building polygons. The second data source is the building use inventory data of the City Planning Basic Survey (2006) from the Administrative GIS provided by the Bureau of Urban Development, Tokyo Governmental Office. These data are updated by field surveys and aerial photography every five years. The building attributes in the administrative GIS include building type, building structure, and number of floors in every building. Although these data have several attributes in common with the residential map, they represent a more detailed classification of building use. Therefore, we can identify mixed-use buildings, which are usually difficult to find using only the residential map. Table A1 in the Appendix shows the building attributes extracted from each spatial microdata used in this article.

We generate an original building map by combining these digital residential map and the administration GIS data. All building polygons on the digital residential map are converted to point data based on its centroid and the building information of the administration GIS is added by overlaying the building polygon on it, using the spatial join function of ArcGIS. The attached attributes include the building use, number of floors, and coefficient of gross floor space.

Figure 2. Flowchart of the proposed approach for estimating residential characteristics at the block level using microdata and census data.

Survey microdata

We use the individual data of the Housing and Land Survey (2008), which is one of the official statistical surveys to investigate not only the current state of housing and buildings but also household living conditions. It is conducted by the Ministry of Internal Affairs and Communications every five years. This is a sample survey of all households in Japan, including a total of 175,000 sample households in the study area (the Tokyo ward area). Although these individual data are unavailable to the public, the recent availability of microdata collected via the census and official governmental surveys is increasing. The merit of these data is that they include housing-related variables such as building type and dwelling size, which are common with the population census housing variables, along with basic household information. Given that the survey period of each data differs slightly, we use the most recent data sources described above.

4. Methods

We explain the estimation procedure of residential characteristic distribution at the block scale in this section. Given the lack of data to validate the results, household size, that is, number of people per household, is chosen as the household attribute to be estimated in this study. The proposed methodology includes major three steps: (1) modeling the housing composition for each block using spatial microdata, (2) establishing the statistical relationship between housing composition and demographic attributes at the small area level using survey microdata, and (3) allocating demographic attributes to street blocks based on this small area-specific relationship (Figure 2). Table 1 summarizes these residential characteristics including demographic and housing (dwelling unit) attributes. Housing composition is determined by the combination of building type and dwelling unit size. Census data are also employed in order to constrain small area totals. Here,

Table 1. Residential characteristics used in this study.

Household attributes	Variables
<i>Demographic attributes</i>	
Household size	Number of persons in households
<i>Housing (dwelling unit) attributes</i>	
Building type	Detached housing / Apartment housing
Dwelling unit size	Dwelling unit floor space

“small area” refers to the small unit represented by the census summary tables such as the census tract (as in the U.S.), output area (as in the U.K.), or *cho* district (as in Japan). Street blocks nested in the small area unit basically represent a minimum zone composed of a space surrounded by road networks. Finally, we illustrate the validation strategy for evaluating the estimation accuracy and its superiority over alternative approaches using limited microdata.

The proposed method uses a variety of data such as building data, survey microdata, and census aggregated data. This integration is enabled by using common housing variables contained in the census data and both types of microdata. In order to obtain joint building environment and household attributes, we select the variables of building type and dwelling size because they are assumed to be strongly related with residents’ demographic attributes. Here, building type refers to the building style of residential units, like one-family detached or attached, or buildings with 2–4 units (as in the U.S. census) and detached, semi-detached, or a flat (as in the U.K. census). Dwelling size indicates gross floor space or the number of rooms in the residential dwelling unit. If the census includes only the number of rooms, the link between the number of rooms and gross floor space needs to be established, because the GIS building map provides only gross floor space including that of the bathroom and kitchen. These variables can be derived from GIS maps and obtained as typical survey items from the population census or housing-related surveys in most countries. In this study, the building type of a residential dwelling is classified as detached and apartment housing, and dwelling size is categorized into six levels of dwelling floor space, according to the categories used in Japan’s population census.⁽¹⁾

In general, dwelling type and size are strongly associated with demographic attributes. Studies on housing demand found that the aged, householders who are professionals, and households with school-age children are not inclined to choose rented apartment housing³⁴⁾ and that an increase in household size, age, or income is associated with demand for larger dwellings.³⁵⁾ However, this trend is not uniform across the city. Welsh and Wadley³⁶⁾ found, for example, that there was significant variability in household size by dwelling size across and within metropolises. Therefore, it is reasonable to assume there is a relationship between housing and demographic variables at the small area level.

4.1. Modeling housing composition at the block level using spatial microdata

First, “tentative” dwelling unit attributes (i.e., building type and dwelling size) are modeled using building and address data. Then, individual dwelling units are weighted so that the small area total count can match the population census. Housing composition based on dwelling units at the block level is estimated by aggregating the data according to blocks.

Table 2. Classification of building type by residential map and administrative GIS data.

	<i>Administrative GIS data</i>					
	No linkage	Independent house	Multifamily building	Residential & commercial	Residential & industrial	Others
<i>Residential map</i>						
Personal housing	D	D	D	D	D	D
Empty or not reported	N	D	N	D	D	N
Business establishment	N	D	N	D	D	N
Building and apartment	A	A	A	A	A	A

Notes: D, A, and N represent detached, apartment housing, and non-residential use buildings, respectively.

4.1.1. Modeling dwelling unit attributes

1) Building use and type

First, it is essential to differentiate residential and non-residential use for individual buildings before conducting the building-based population estimation. Then, each residential building must be classified by building type, such as detached or apartment housing. The intent of this step is to identify the residential buildings by building type according to the census category. Some studies classify buildings into single-family and multi-family houses and non-residential units using satellite imagery, land use maps, and address point data²³⁾ or even Light Detection and Ranging (LiDAR) data, as done by Lu et al.³⁷⁾ in the U.S.' context. For modeling building type, open access high-resolution topographic maps have also been used in the U.K.³⁸⁾ and in Germany.³⁹⁾

In the Japanese context, as described in Section 3, we can directly obtain building use and type from administrative GIS data as well as a nationwide digital residential map, which has proved to be useful for small area population estimation.²⁷⁾⁴⁰⁾ As shown in Table 2, all buildings included in the residential map are classified as detached, apartment, or non-residential based on the information integrating two GIS building maps. In particular, we need to include a part of mixed-use buildings when detached housing is identified because a housing unit with a store or workshop is regarded as detached housing in the census definition although it is difficult to identify it using only the digital residential map.⁽²⁾ The effectiveness of employing spatial microdata and combining different spatial microdata sources is also assessed in Section 4.4 by comparing the alternative methods BI (using no spatial microdata) and BII (using only one digital residential map as spatial microdata) against the proposed method A.

2) Number of dwelling units

For detached housing, the count for the dwelling unit is considered to be equal to that for detached housing. On the other hand, in order to estimate the number of residential dwelling units for apartment housing, information on the internal structure of the building is required. The total count of residential units within a building can be obtained directly by aggregating the number of address point data, which are available for several countries,¹³⁾ or if available, tax roll records may be used.¹⁵⁾ The number of housing units at the block level may also be calculated by regression analysis using land use data.⁴¹⁾ For this study, the number of residential dwelling units is obtained by aggregating the number of residential units in the buildings contained in the appended tables of the digital residential map.⁽³⁾

3) Dwelling unit size

Depending on the identified building type, dwelling size (i.e., unit floor space) is derived based on the building footprint and number of floors in the building, and allocated to a category of housing floor space from the population census. We calculate the unit floor space s for each detached housing unit as follows:

$$s = BF \quad (1)$$

where B is the building footprint, and F is the number of floors.

Conversely, for apartment housing, the following three steps are carried out. First, for all the floors of each building, the number of floors housing only residential units (i.e., dwelling or vacant) are identified and used in the next step as the number of floors for residence. Orford⁴²⁾ pointed out issues of living space estimation of flat and proposed that common space and commercial space should be subtracted from the gross floor space of the building. We use this concept to calculate the number of floors for residence and introduce what we refer to as the rate of exclusively owned space in the next step. Although some studies identified non-residential usage floor using OpenStreetMap tag data⁴³⁾ or simply regarded only one floor for residential use,⁴²⁾ we derive the residential information of each building from the appended tables of the digital residential map.

Second, the average unit floor space $\bar{s}$ is calculated for each building:

$$\bar{s} = BF_r r / S \quad (2)$$

where B is the building footprint, F_r is the number of floors for residence, r is the rate of exclusively owned space, and S is the total number of units within the building. Data on building footprint are obtained from open data topographic maps¹⁴⁾⁴⁴⁾ as well as satellite images or aerial photography.¹⁵⁾¹⁹⁾²⁰⁾²³⁾ To subtract the common space, the rate of exclusively owned space is kept constant at 0.8.²⁷⁾

Third, the distribution of the dwelling unit floor space inside the building is estimated because some apartment buildings contain various unit sizes. Based on the unit floor space distribution by the average unit floor space and number of floors, the dwelling units for each building are proportionally allocated to any one of the dwelling size categories. While this information is included in additional data sources such as property data in the estate agents' surveys,⁴²⁾ we use the survey microdata sample of the Housing and Land Survey.

4.1.2. Constraining housing variables at the small area level

Before determining the dwelling unit composition at the block level based on the tentative dwelling unit attributes modeled in previous section, the total count of dwelling units is controlled by building type and dwelling size to match the census summary tables at the small area level. This is because tentative housing attributes contain unavoidable errors. The small area statistics of the census generally provide one-dimensional variables such as household counts by building type and dwelling size category. First, using these common census housing variables as constraints (marginal distributions), the joint distribution of households by building type and dwelling size at the small area level is generated through the IPF procedure,⁴⁾ and these are used as control tables.⁵⁾ Here, the initial distribution is created by aggregating tentative housing attributes at the small area level.

Table 3. The relationship between housing variables and household size based on survey microdata (whole of the Tokyo ward area).

Housing variables		Household size (person)					
		1	2	3	4	5	≥ 6
Detached housing	0–29m ²	343	114	53	20	5	1
	30–49m ²	916	953	448	215	70	23
	50–69m ²	1,607	2,192	1,335	895	239	86
	70–99m ²	1,934	3,402	2,599	2,283	692	257
	100–149m ²	1,664	3,308	2,802	2,488	967	429
	≥150m ²	943	1,879	1,536	1,319	672	479
Apartment housing	0–29m ²	28,706	2,784	767	268	63	18
	30–49m ²	14,453	8,391	3,362	1,477	357	96
	50–69m ²	7,252	7,971	5,018	3,413	664	182
	70–99m ²	2,789	4,110	3,177	2,743	635	135
	100–149m ²	498	754	575	421	161	60
	≥150m ²	117	161	134	104	39	31

Second, individual dwelling unit attributes are weighted under the given total count of the number of households by building type and dwelling size category for each small area. For detached housing, the weight for an individual dwelling unit is calculated to match the total counts of the number of households by building type when aggregated at the small area level. Next, the dwelling size category is assigned to correspond with the number of households by the dwelling unit floor space in ascending order. For apartment housing, on the other hand, after weighting in the same manner as for detached housing, the dwelling size category is allocated through the IPF method at the small area level. In detail, we impute the joint distribution by the weighted number of dwelling units in each building and the number of households living in apartments by dwelling size for each small area. In this study, the initial distribution is sourced from the same sample data used in the last step in Section 4.1.1. Finally, the dwelling unit composition at the block level was obtained by aggregating the units by building type and the dwelling size by street block.

4.2. Establishing the small area-specific relationship between housing and demographic variables using survey microdata

The relationship between dwelling units and demographic attributes at the small area level was established using the cross tabulation method through the IPF procedure. It is usually assumed that there is a strong relationship between housing variables and demographic attributes. As to the building type, for example, single-person households tend to live in an apartment while a large household tends to live in detached housing. Additionally, as to the unit floor space, larger families tend to live in a larger house. Table 3 confirms these trends, illustrating the relationship between housing and household size based on survey microdata. The chi-square test of independence between housing variables (building type and dwelling size) and household size using our survey microdata rejects the null hypotheses of dependence for all wards ($p < 0.001$).⁽⁶⁾ This implies that household size distribution significantly varies according to housing composition.

The IPF procedure, an algorithm estimating unknown joint distributions under known marginal distributions, is briefly described here. Assuming a two-dimensional case, let P_{ij} denote an initial distribution (or seed variables), and let A_i and B_j denote marginal distributions. The joint distribution in a two-dimension cross table is imputed as follows:

Step 1: Set initial distribution $P_{ij}^{(0)} = \pi_{ij}$ and marginal distribution $Q_i = A_i, Q_j = B_j$

Step 2: $P_{ij}^{(n+1)} = P_{ij}^{(n)} \left(Q_i / P_{i+}^{(n)} \right), P_{ij}^{(n+2)} = P_{ij}^{(n+1)} \left(Q_j / P_{+j}^{(n+1)} \right)$

Step 3: Convergence condition $\left| Q_{i+} / P_{i+}^{(n+2)} - 1 \right| < \varepsilon$ and $\left| Q_{+j} / P_{+j}^{(n+2)} - 1 \right| < \varepsilon$

Step 4: Repeat Steps 2 and 3 until the condition above is satisfied.

Here, n is the repetition number, and ε is the fixed convergence criterion.

The relationship between housing and demography can be captured by small area cross tabulation including the number of households by household size, building type, and dwelling size category. This procedure is implemented by employing this IPF method for each small area, and it imputes the joint distribution constrained by a) the number of households by building type and dwelling size category obtained from Section 4.1 and b) the number of households by household size⁽⁷⁾ from the population census small area statistics. In this study, the survey microdata sample of the Housing and Land Survey is used to create the initial distribution by aggregating households by household size, building type, and dwelling size category at the municipality level. In order to underline the importance of introducing these relationships and survey microdata, the performance of alternative method CI (assuming independence between the housing and demographic variables by setting each cell of the seed variables to 1 in this process, and therefore using no survey microdata) is compared with that of the proposed and other alternative methods in Section 4.4.

Since these relationships between housing and demography are derived for each small area, the spatial heterogeneity between neighborhoods can be reflected by this estimation approach. In Section 4.4, the effectiveness of small area constraints and the role of survey microdata are also assessed by comparing the alternative methods CII (using census data at the municipality level as seed variables instead of survey microdata) and CIII (with the same condition as CII but controlled by small area census variables).

4.3. Allocating demographic attributes to blocks

After integrating the dwelling unit composition at the block modeled in Section 4.1 and the small area-specific relationship between the dwelling unit and demographic attributes obtained from Section 4.2, we assign household attributes to blocks assuming that one household lives in one dwelling unit. Sharing common housing variables (i.e., building type and dwelling size) linking blocks and households enables this allocation. Let $H_{tsi}(A)$ denote the number of households with size i living in the dwelling unit of building type t and dwelling size s in small area A . Then, in small area A , the number of households with size i living in dwelling unit of building type t and dwelling size s in block b ($b \in A$) is calculated as $\hat{H}_{tsi}(b)$ as follows:

$$\hat{H}_{tsi}(b) = D_{ts}(b) * \frac{H_{tsi}(A)}{\sum_i H_{tsi}(A)} \quad (3)$$

where $D_{ts}(b)$ is the total number of dwelling units of building type t and dwelling size s in block b included in small area A .

4.4. Accuracy assessment

Although no reference data are publicly available to directly compare and validate outcomes, we are able to evaluate them in an indirect manner using the population census statistic summary for basic unit blocks (BUBs). These data provide only the number of households and population at BUBs, which are geographically roughly equivalent to street blocks. We associated the BUBs' location points with street blocks and aggregated the number of households and population by street blocks. Our strategy for examining estimation performance is to compare the number of households and people calculated from the estimated distribution of household size with that of census-based data by blocks using some measures of accuracy. The estimated population P in a block is calculated from the household size distribution as follows:

$$P = \sum_k k h_k \quad (4)$$

where h_k is the number of households with k people as family members, which was obtained as the output in Section 4.3.

Then, we use three measures of the accuracy of each estimation method: the R^2 value of the correlation between the estimated and census-based population (or households) counts for each block, root mean square error (RMSE), and mean absolute percent error (MAPE). RMSE and MAPE are often used for evaluating the estimation performance and comparing it with alternative approaches, and each index is obtained as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p'_i - p_i)^2} \quad (5)$$

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^n \left| \frac{p'_i - p_i}{p_i} \right| \quad (6)$$

where p'_i is the estimated population (or number of households) in block i , p_i is the census-based population (or number of households) in block i , and n is the number of blocks.

In order to demonstrate the effectiveness of the current method using survey and spatial microdata, we compare some alternative estimation approaches using limited microdata. These are summarized as follows. Relative to the current method A integrating spatial and survey microdata, BI and BII employ limited spatial microdata. BI employs the area weighting method⁴⁵⁾ where the population (or households) is (are) allocated spatially according to areal units without any ancillary data, and BII uses only the residential map as spatial microdata. In contrast, CI, CII, and CIII do not employ any survey microdata. In establishing the relationship between housing and demography in Section 4.2, CI assumes no relationship between these variables, while CII and CIII refer to aggregated census households in a larger (municipal) area.⁽⁸⁾ CII is directly applied to the relationship derived from the cross table at the municipality level, and CIII adds constraints using small area cross tabulation.

Figure 3. Estimated percentage of single-person households at the block level: (a) around a station and (b) along a main road.

5. Results

The method described in the previous section is implemented for street blocks located within the entire Tokyo ward area. The outcome provides the number of households with multiple attributes, including household size, building type, and dwelling size of housing at the block level, which can reproduce census cross tables when aggregated at the small area (*cho* district) level.

Before evaluating the accuracy of estimation performance, we briefly illustrate some characteristics which we are unable to identify from the *cho* district mapping, describing the spatial distribution of single-person households obtained from the estimated results. The most distinctive feature is extremely high rates of single-person households around rail stations (Figure 3a). The *cho* district, which contains rail stations, is subject to uneven distribution. This might reflect residential preference patterns in urban areas where single people tend to place a high value on accessibility and many apartments are supplied for them. Another feature is the relatively high rates of single people along the periphery of a main road (Figure 3b). In this case, apartment buildings stand on one side of blocks adjacent to a main road, and in contrast, detached housing exclusively for families spreads out on the other. This location pattern might be subject to zoning control. Such micro areal characteristics inside or beyond the administrative boundary are successfully described in block scale mapping.

Using the results, we compare the estimated number of households and population with census-based (BUB) data and evaluate their performance relative to alternative approaches following the method in Section 4.4. In implementing this approach, we randomly choose 520 blocks out of all the blocks across the study area. Notably, we face a limitation for full-scale evaluation of household attributes because this strategy summarizes the distribution of household attributes in terms of the number of people. Using the R^2 values, RMSE, and MAPE as measures of accuracy, Table 4 shows the comparative statistics of the estimated results obtained from the proposed method and some alternatives using limited microdata for sample blocks. As expected, the current method A outperforms the others, as indicated through the highest value of R^2 and the lowest value of RMSE and MAPE in terms of both household and population. Figure 4 also shows a high correlation between estimation and census data in terms of both households and population, even in blocks with a small population size. Although a comparative analysis by Brinegar and Popick⁴⁶⁾

Table 4. Comparative statistics for sample block estimates.

	Spatial microdata	Survey microdata	Household			Population		
			R^2	RMSE	MAPE	R^2	RMSE	MAPE
A*	✓	✓	0.987	35.19	16.56	0.984	80.63	20.38
BI	—	✓	0.871	132.93	73.43	0.889	293.16	61.30
BII	△ ^{a)}	✓	0.874	99.62	20.29	0.876	227.96	22.15
CI	✓	—	0.987	35.19	16.56	0.979	88.93	24.61
CII	✓	— ^{b)}	0.987	35.19	16.56	0.976	97.51	22.27
CIII	✓	— ^{c)}	0.987	35.19	16.56	0.981	86.37	21.11

Notes: * denotes the proposed model, a) means we use only one residential map (without administrative GIS map), b) means we use census data for a larger area, and c) means we use census data for a larger area with small area constraints.

revealed that existing interpolation methods tend to perform poorly in areas having small populations or low population density, our method using microdata has the potential to overcome such a weakness.

BI and BII, both methods using limited spatial microdata, significantly reduce the accuracy of estimation relative to the proposed method A (Table 4). In particular, BI resulted in the worst estimation, because the area weighting method assumes a uniform population density and employs no spatial microdata, and therefore, it lacks informative variables such as housing composition. BI generally suffers from an underestimation of both household and population as shown in Figure 4, perhaps because the household is assigned regardless of housing unit location, although the population density varies between areas. Figure 4 also shows that BI performs much worse than BII in blocks with a small population despite their R^2 values being quite similar. Although using spatial microdata, namely a residential map, substantially improves the errors of block-level estimation relative to BI, both methods face a common challenge in the estimation of population attributes in particular, judging from the errors in population estimation which are larger than that of households, except for the MAPE of BI.

On the other hand, CI–CIII, which use spatial but no survey microdata, perform better than BI and BII and produce outcomes similar to those of the present model A in terms of accuracy (Table 4). This highlights the significant role of spatial microdata for population estimates at finer scales. Since the number of households for each block has already been identified using spatial microdata, and therefore the number of estimated households is the same for CI–CIII, we focus on evaluating the population estimation. Among them, CIII provides the most accurate estimation, as the small area estimation is based on the relationship between housing and demography in a larger area, using the census data at the municipality level. Figure 4 indicates that the proposed method and CIII have quite similar scatterplots although the latter has more errors in blocks with less than 1,000 residents. In contrast, the worst estimation is presented by CII, where the relationship between housing and households in a larger area is applied without small area estimation.

Figure 4. Scatterplots of the estimated counts versus census-based counts at the block level: household (left) and population (right).

6. Discussion

The results show the advantage of using survey and spatial microdata to study the block-level geography of demography not captured by conventional small area statistics. We map the estimated distribution of residential characteristics and compare the proposed method with alternative methods. Some possible sources of error remain if we use spatial microdata with only a residential map. First, the definition of building type differs between the residential map and the census. While the census defines housing to include dwellings with shops and workshops as well as those used exclusively for residential purposes, it is difficult to extract mixed-use building data from the residential map. Indeed, compared with the population census at the municipality level, we can identify detached housing in less than 60% of the central business districts of the Tokyo area (Chiyoda, Chuo, and Minato) and Taito, which is highly commercial, using only the residential map. As Kunze & Hecht⁴³⁾ claimed, combining more than two data sources improves the estimation accuracy. The proposed method incorporates administrative GIS data and also allowed the identification of dwellings used for business or other purposes, and consequently, the differences between the number of identified detached housing units derived from our building map and the number of households derived from the census is within 10% for all wards except Chiyoda. Chiyoda ward differs from the others in that it is located in a business district, and the number of residential units is limited.

Second, a considerable number of appended tables in the residential map have imperfections. As only 70% of apartment buildings in the Tokyo ward area have appended tables, the rest needed to be estimated. Considering the high rate of households living in apartment buildings in Tokyo, it is necessary to correct these defects for a better estimation. Another defect of the residential map is that some large-scale apartment complexes with multi-tenants are labeled as business establishments.⁽⁹⁾ Despite three-dimensional building information being highly useful in estimating vertical population distribution, this lack of data and mislabeling of spatial microdata could cause significant errors because all households living in apartment buildings are allocated to the incorrect housing unit category.¹⁹⁾ This issue can be noticed in Figure 4, where we can observe the outliers in some blocks from the BII outcome. The remaining errors presumably arise on account of the data belonging to different survey periods and the presence of unreported households, shared households, and vacant dwellings.⁽¹⁰⁾ Although it is difficult to completely remove the errors from the spatial microdata, constraining the census housing variables as described in Section 4.1, will play a role in improving the estimation accuracy.

The findings concerning survey microdata emphasize the need for small area tabulation. The population census usually provides two- or three-dimensional cross tables including housing and household variables at the municipality level or higher. However, the relationship between them varies even within the same municipality. For example, Figure 5 shows the variations in the rates of single-person households whose housing type is an apartment and dwelling size is 50–69 m², and we can indeed see that while there is great variability in the central business district area (Chiyoda, Chuo, and Minato), the rates are relatively homogeneous in Arakawa. This means that the relationship at the municipality level alone is not enough for some areas. Therefore, an estimation constraining variables to the small area level can provide a satisfactory result if survey microdata are completely unavailable. In addition, as shown by the outcome of CI, even if the estimation is implemented at the small area (*cho* district) level, a lack of a relationship between variables might prevent improved performance. An assumption of independence produces large errors in cross tabulation with a strong relationship between variables.⁽⁶⁾

Figure 5. Boxplot of *cho* district percentage of single-person households whose housing type is apartment and dwelling size is 50–69 m².

Source: Population census of Japan (2010).

Although this article focuses on household size as a demographic attribute for considering verifiability, using survey microdata is essential in estimating other residential characteristics such as household generation and household income because these are also assumed to have a strong association with housing variables. Indeed, for example, the relationship between housing composition and household generation⁽¹¹⁾ is not independent ($p < 0.001$) according to a chi-square test of independence using our sample data of Housing and Land Survey for all wards in Tokyo. The same holds for household income.⁽¹²⁾ Though strict accuracy evaluation is difficult owing to lack of reference data at the street block level, a degree of reliability is assured in that aggregation at the small area level at least matches the small area totals. It is expected that such estimated socio-economic and demographic data of a fine-scaled population may serve as input data for studies in this field.⁽²⁴⁾

Concerning the spatial microdata used in the proposed method, though the nationwide residential map is unique to Japan, there is great potential to apply the current framework to different contexts by using other data sources. Particularly, there may be increased benefit in using address point data⁽¹³⁾⁽⁴⁷⁾⁽⁴⁸⁾ and points-of-interest information⁽⁴³⁾⁽⁴⁴⁾ because the key feature of information needed for more sophisticated estimation in high-density urban areas is the internal structure of mixed-use buildings and apartment housing. For the country and the area where fine-level census data with rich attribution is not available, including developing countries, the current method is meaningful in that it can be applied using data for larger area or with limited variables. Especially, this study stresses that spatial microdata can play a certain role to disaggregate coarse population data without small area census statistics. In such situations, satellite imagery or aerial photography would supplement spatial microdata as some previous studies proposed.

The biggest obstacle of the current method might be the absence of dwelling unit floor space information in the census of some countries, including the U.S. and the U.K. Instead, additional steps are needed to convert the number of rooms into floor space in order to be representative of the dwelling size recorded by the census statistics. This issue poses an important area for future research.

7. Conclusion

This article proposed a method to estimate residents' characteristics at the street block level using survey and spatial microdata. The proposed model not only reflected neighborhood characteristics of the relationship between housing and household variables from the survey microdata but also improved spatial refinement by using residential dwelling unit information obtained from a combination of two types of GIS building data. We demonstrated our method using an entire Tokyo ward area, characterized by vertical population extension and residential diversity. In spite of this complex residential structure, the accuracy assessment indicated that the proposed model performs well and presents a more accurate estimation than any other approaches using limited microdata. In particular, we noted the role of spatial refinement played by spatial microdata and housing unit level information, which is significant in a high-density area like Tokyo. Furthermore, a combination of more than one type of spatial microdata is very effective because using a single data source sometimes leads to unavoidable errors depending on regional conditions and data accuracy. On the other hand, survey microdata are important with respect to estimating the multi-dimensional attributes of residents. Without microdata, however, the second best way is to use census-aggregated population data from a larger area.

This study suffers from some limitations, which may be addressed in future research. First, it is possible to estimate other residential attributes, which are assumed to have a strong correlation with housing variables, using our method as discussed in the previous section, because the dimensional extension is allowed in generating cross tabulation, at least to some degree. However, currently, no data are available at the block level for evaluating the estimated demography. Second, an examination of data accuracy and more efficient correction will be needed because cost and time will increase as the diversity of microdata grows and the geographical range of application is expanded. It could hardly be said that the impact from the data defects and mislabeling inherent in our spatial microdata are completely removed in spite of supplementing and manually correcting them. Thus, future studies may consider improving the technique by combining more informative data, for example, automated data collection via the Internet.

In spite of these limitations, by successfully highlighting the effectiveness of various microdata for small area demographic estimation, this research discusses the application potential of the proposed approach. In addition to increasing the development of spatial microdata, such as volunteered geographic information, availability of public survey or census microdata has expanded in several countries even after data have been processed for anonymity or fabrication.³⁾⁴⁹⁾ This increasing availability of various microdata will ensure further progress in small area estimation, potentially broadening the field of application to demographic analyses related to public health, crime and safety, and environmental justice.

Acknowledgements

This research used the residential maps provided by Zenrin Co., Ltd as the joint research with Center for Spatial Information Science, the University of Tokyo (No. 411) and was supported by JSPS KAKENHI, Grant-in-Aid for JSPS Fellows.

Notes

- (1) Six categories of dwelling unit floor space from the population census: 0–29 m², 30–49 m², 50–69 m², 70–99m², 100–149 m² and more than 150 m².
- (2) Detached housing defined here includes ‘tenement houses’ and ‘others’ as well as ‘detached houses’ of the census category for building type while apartment housing is from ‘apartment houses or flats’. Although it is difficult to identify tenement houses and others from our building map, the number of them is limited in the Tokyo ward area (1.2% and 0.3%, respectively). For the residential map, vacant and not reported are also involved because the number of housing units including this type of housing is more consistent than that which excludes them.
- (3) Although the appended tables provide some information of occupants inside the building, not all apartment buildings and multi-tenanted buildings have them. In order to supplement this information, the total number of units are estimated by means of regression analysis using known building information from the appended tables as training data.
- (4) This procedure is explained in detail in Section 4.2.
- (5) One problem regarding this process is that the estimated joint distributions are not necessarily guaranteed to match census cross tables for a larger geographical area (i.e., municipality) because constraints are imposed only at the small area level.³²⁾ In order to address this issue, the initial distribution was adjusted making use of census cross tables at the municipality level, and adding a correction factor to the initial distribution. Let i denote building type and j denote dwelling size category, and the correction factor m_{ij}^k for cell (i, j) of initial distribution in *cho* district k is as follows:

$$m_{ij}^k = \frac{P_{ij} - \sum_K p_{ij}^k}{\sum_K p_{ij}^k}$$

where P_{ij} is the value of the cell (i, j) of census cross table for the municipality to which *cho* district k belongs and p_{ij}^k is the value of the cell (i, j) of the initial distribution in *cho* district k .

- (6) Chi-square values are ranging between 657.0 (Chiyoda) and 4965.9 (Setagaya).
- (7) Six categories of household size: one person, two people, three people, four people, five people and more than six people.
- (8) In fact, the municipality cross table is estimated sequentially by IPF using census data of a larger area (i.e., municipality) because the three-dimensional cross table in question is unavailable as it is even at municipality level without disaggregated data.
- (9) In order to address this problem when applied to Tokyo, the present method involved manually correcting the total number of residential units of apartment buildings searching property information on the Internet by the building name, discovering that *cho* districts were underestimated by over 20% compared to the population census at the *cho* district level. Consequently, the R^2 value representing the correlation between the number of identified

residential units and that of households based on the population census exceeds 0.95 for all wards except Chiyoda and Chuo (0.85 in Chiyoda and 0.94 in Chuo).

- (10) According to the House and Land Survey (2008), the rate of vacant dwellings in the whole of the Tokyo ward area is 11.3%, which ranged from 7.6% in Setagaya to 25.8% in Chiyoda.
- (11) Three categories of household generation: household including under the age of 18 children, household composed of over the age of 65, and others.
- (12) Three categories of household income: household in the top 20%, household in the bottom 20%, and others. The household income distribution at *cho* district level is estimated by Uesugi & Asami⁵⁰⁾ because Japan's population census do not provide income data.

References

- 1) Openshaw, S. (1983). The modifiable areal unit problem. *Concepts and techniques in modern geography*, 38, Norwich: Geobooks.
- 2) Kurban, H., Gallagher, R., Kurban, G., & Persky, J. (2011). A beginner's guide to creating small-area cross-tabulations. *Cityscape: A Journal of Policy Development and Research*, 13(3), 225–235.
- 3) Leyk, S., Battenfield, B. P., & Nagle, N. N. (2013). Modeling ambiguity in census microdata allocations to improve demographic small area estimates. *Transaction in GIS*, 17(3), 406–425.
- 4) Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4), 427–444.
- 5) Johnston, R., & Pattie, C. (1993). Entropy maximizing and the IPF procedure. *The Professional Geographer*, 45, 317–322.
- 6) Simpson, L., & Tranmer, M. (2005). Combining sample and census data in small area estimates. *The Professional Geographer*, 57(2), 222–234.
- 7) Wong, D. W. S. (1992). The reliability of using the iterative proportional fitting procedure. *The Professional Geographer*, 44(3), 340–348.
- 8) Hermes, K., & Poulsen, M. (2012). A review of current methods to generate synthetic spatial microrodata. *Computers, Environment and Urban Systems*, 36, 281–290.
- 9) Rahman, A., Harding, A., Tanton, R., & Liu, S. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation*, 3(2), 3–22.
- 10) Birkin, M., & Clarke, M. (1988). SYNTHESIS – a synthetic spatial information system for urban and regional analysis: Methods and examples. *Environment and Planning A*, 20(12), 1645–1671.
- 11) Wu, S., Qiu, X., & Wang, L. (2005). Population estimation methods in GIS and remote sensing: A review. *GIScience and Remote Sensing*, 42(1), 80–96.
- 12) Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass*, 3, 727–745.
- 13) Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS*, 15, 5–27.
- 14) Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis*, 45(3), 324–344.

- 15) Xie, Z. (2006). A framework for interpolating the population surface at the residential housing-unit level. *GIScience and Remote Sensing*, 43, 233–251.
- 16) Alahmadi, M., Atkinson, P., & Martin, D. (2013). Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41, pp. 167–176.
- 17) Lwin, K., & Murayama, Y. (2009). A GIS estimation of building population for microspatial analysis. *Transactions in GIS*, 13(4), 401–414.
- 18) Qiu, F., Sridharan, H., & Chun, Y. (2010). Spatial autoregressive model for population estimation at the census block level using LIDAR-derived building volume information. *Cartography and Geographic Information Science*, 37, 239–257.
- 19) Sridharan, H., & Qiu, F. (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis*, 45, 238–258.
- 20) Wu, S., Wang, L., & Qiu, X. (2008). Incorporating GIS building data and census housing statistics for sub-block-level population estimation. *The Professional Geographer*, 60, 121–135.
- 21) Cockx, K., & Canters, F. (2015). Incorporating spatial non-stationarity to improve dasymetric mapping of population. *Applied Geography*, 63, 220–230.
- 22) Jia, P., Qiu, Y., & Gaughan, A.E. (2014). A fine-scale spatial population distribution on the high-resolution gridded population surface and application in Alachua County, Florida. *Applied Geography*, 50, pp. 99–107.
- 23) Ural, S., Hussain, E., & Shan, J. (2011). Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation*, 13, 841–852.
- 24) Dmowska, A., & Stepinski, T. F. (2014). High resolution dasymetric model of US demographics with application to spatial distribution of racial diversity. *Applied Geography*, 53, 417–426.
- 25) Leyk, S., Nagle, N. N., & Battenfield, B. P. (2013). Maximum entropy dasymetric modeling for demographic small area estimation. *Geographical Analysis*, 45(3), 285–306.
- 26) Leyk, S., Battenfield, B. P., Nagle, N. N., & Stum, A. K. (2013). Establishing relationships between parcel data and land cover for demographic small area estimation. *Cartography and Geographic Information Science*, 40(4), 305–315.
- 27) Akiyama, Y., Takada, T. & Shibasaki, R. (2013). Development of micropopulation census through disaggregation of national population census. In *Paper presented at Computers in Urban Planning and Urban Management 2013 conference*.
- 28) Lin, J., Cromley, R., & Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS*, 17(1), 1–14.
- 29) Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74, 519–530.
- 30) Igarashi, G., Tokunaga, Y., Miyamoto, K., & Sugiki, N. (2005). A forecasting model of household distribution by zone and detailed attributes for a metropolitan area. *Journal of the City Planning Institute of Japan*, 40(3), 943–948 (in Japanese).
- 31) Ishizaka, K., Utsumi, K., & Ohashi, Y. (2011). Estimation method of multi-dimensional census data in small district for analyzing housing conditions. *Proceedings of the Housing Studies Symposium 2011*, 169–176 (in Japanese).

- 32) Yamada, M., Itoikawa, E., & Sato, T. (2012). Habitation structure after the reconstruction of Hanshin-Awaji great earthquake disaster area: The method of estimation for habitation micro data, and analysis of habitation structure of Kobe. *Journal of Architecture and Planning*, 677, 1663–1672 (in Japanese).
- 33) Asami, Y., & Kido, K. (1998). Statistical accuracy of contingency tables estimated with the iterative proportional fitting method applied to the housing related data of population census: Case study of Tokyo metropolitan area. *Journal of Architecture and Planning*, 514, 185–189 (in Japanese).
- 34) Cho, C.J. (1997). Joint choice of tenure and dwelling type: A multinomial logit analysis for the city of Chongju. *Urban Studies*, 34(9), 1459–1473.
- 35) Skaburskis, A. (1999). Modelling the choice of tenure and building type. *Urban Studies*, 36 (13), 2199–2215.
- 36) Welsh, A., & Wadley, D. (2014). The Relationship of dwelling size and population potential: Practical implications for strategic urban planning. *Urban Policy and Research*, 32(1), 85–105.
- 37) Lu, Z., Im, J., Rhee, J., & Hodgson, K. (2014). Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landscape and Urban Planning*, 130, 134–148.
- 38) Orford, S., & Radcliffe, J. (2007). Modelling UK residential dwelling types using OS Mastermap data: A comparison to the 2001 census. *Computers, Environment and Urban Systems*, 32(2), 206–227.
- 39) Meinel, G., Hecht, R., & Herold, H. (2009). Analyzing building stock using topographic maps and GIS. *Building Research & Information*, 37(5–6), 468–482.
- 40) Greger, K. (2014). Spatio-temporal building population estimation for highly urbanized areas using GIS, *Transactions in GIS*, 19(1), 129–150.
- 41) Deng, C., & Wu, C. (2013). Improving small-area population estimation: An integrated geographic and demographic approach. *Annals of the Association of American Geographers*, 103(5), 1123–1141.
- 42) Orford, S. (2010). Towards a data-rich infrastructure for housing-market research: Deriving floor-area estimates for individual properties from secondary data sources. *Environment and Planning B: Planning and Design*, 37(2), 248–264.
- 43) Kunze, C., & Hecht, R. (2015). Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems*, 53, 4–18.
- 44) Bakillah, M., Liang, S.H.L., Mobasheri, A., Arsanjani, J., & Zipf, A. (2014). Fine resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, 28(9), 1940–1963.
- 45) Goodchild, M. F., & Lam, L. S. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing*, 1, 297–312.
- 46) Brinegar, S. J., & Popick, S. J. (2010). A comparative analysis of small area population estimation methods. *Cartography and Geographic Information Science*, 37(4), 273–284.
- 47) Mesev, V. (2005). Identification and characterization of urban building patterns using IKONOS imagery and point-based postal data. *Computers, Environment and Urban Systems*, 29(5), 541–557.

- 48) Swanson, D. A., & McKibben, J. N. (2010). New directions in the development of population estimates in the United States? *Population Research and Policy Review*, 29(6), 797–818.
- 49) Hanaoka, K. (2012). The possibility of using anonymized data from Japanese official statistics for implementing geospatial analysis at a small area level: Creating a geographically disaggregated synthetic microdata set by using a spatial microsimulation. *Japanese Journal of Human Geography*, 64(3), 195–211 (in Japanese).
- 50) Uesugi, M., & Asami, Y. (2011). Neighborhood segregation by household income since the late 1990s: Case study of Ota ward, Tokyo. *Geographical review of Japan*, 84(4), 345–357 (in Japanese).

Appendix

Table A1. Building information available from spatial microdata used in this study.

	Residential map (Zenrin Co., Ltd.)	Administrative GIS data (Tokyo Governmental Office)
Number of floors	Above ground floor	Underground and above ground floor with coefficient of gross floor space
Use	Building and apartment Personal housing Business establishment Landmark Vacant or not reported	Public office Educational and cultural Health and welfare Supply and disposal services Business office Exclusively commercial Residential with commercial Hotel and leisure Sport and entertainment Independent house Multifamily building Exclusively industrial Residential with industrial Warehouse and transport Agriculture, Forestry and Fisheries
Name	Nameplate / building name	-
Address	Minimum level: sub-block	Minimum level: <i>cho</i> district (numeric code)
Internal structure	Appended tables for some apartment buildings provide: room number, resident and tenant name, use (landmark, dwelling, business, vacant or not reported and others)	-
Others	Positional coordinates and footprint of building are obtained from building geometry	Positional coordinates and footprint of building are obtained from building geometry