Population changes in residential clusters in Japan

Takuya Sekiguchi[1,2], Kohei Tamura[3], Naoki Masuda[4] *

1 National Institute of Informatics, Chiyoda-ku, Tokyo, Japan, **2** JST, ERATO, Kawarabayashi Large Graph Project, c/o Global Research Center for Big Data Mathematics, NII, Chiyoda-ku, Tokyo, Japan, **3** Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Sendai, Miyagi, Japan, **4** Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

* naoki.masuda@bristol.ac.uk

Abstract

Population dynamics in urban and rural areas are different. Understanding factors that contribute to local population changes has various socioeconomic and political implications. In the present study, we use population census data in Japan to examine contributors to the population growth of residential clusters between years 2005 and 2010. The data set covers the entirety of Japan and has a high spatial resolution of 500 × 500 m^2, enabling us to examine population dynamics in various parts of the country (urban and rural) using statistical analysis. We found that, in addition to the area, population density, and age, the shape of the cluster and the spatial distribution of inhabitants within the cluster are significantly related to the population growth rate of a residential cluster. Specifically, the population tends to grow if the cluster is "round" shaped (given the area) and the population is concentrated near the center rather than periphery of the cluster. Combination of the present results and analysis framework with other factors that have been omitted in the present study, such as migration, terrain, and transportation infrastructure, will be fruitful.

Introduction

Population change is a central precondition to be considered in policy making and urban planning. In urban areas with high population concentrations, decentralization policies may be designed to mitigate congestion and environmental problems [1]. In developing countries, rapid growth of the number of urban dwellers is forecasted to exacerbate water shortage [2]. In rural areas facing population aging and shrinkage, how to ensure convenience of public transportation [3] and health care services [4] is a crucial issue.

The choice of the residential location is a main determinant of spatial patterns of population changes over time. People have been suggested to choose the residential location by considering residential environment attributes such as the accessibility to workplace measured by commute distance [5–7], school quality [8, 9], and the crime rate [8, 10]. Residential mobility is also affected by the individual's life course and household attributes such as age and income [7, 10], job change [5], marital status [11], the numbers of children and drivers [10], and home ownership [7, 11].

In addition to these factors, spatial characteristics of the city and inhabited areas, which shape socioeconomic and geographical environments, may also impact spatio-temporal patterns of population changes. For example, urban sprawl is considered to be a consequence of uncoordinated and unplanned urban development [12] and results in scattered spatial patterns of employment and residences in suburban areas [13–16]. These spatial patterns would cause a long commute time due to poor accessibility to workplaces [17]. In contrast, compact urban growth and the diversity of land uses within the region enhance the accessibility to both work and non-work activities [18, 19]. If the accessibility to workplaces and other activities influences residential decision-making, spatial patterns of inhabited regions are expected to affect dynamics of population changes.

There have been studies relating the population size or its change to spatial patterns of urban areas. For example, the population size of a region was shown to obey a power-law relationship with the area of the region in Norfolk in England [20] (also see [21] for an analysis of approximately 70000 cities in the world). In 78 regions in Israel, the population growth rate in sprawl regions was higher than in compact regions, where the sprawl and compact regions were defined in part by the shape of their boundaries [22]. Fractal dimensions are also useful tools for relating the population size/growth and spatial patterns of residential areas. For example, the fractal dimension of the central part of Tel Aviv metropolis and its population size concomitantly increased over time, and the observed fractal dimension was larger than that of the wider Tel Aviv [23]. In 20 urban areas in the US, the fractal dimension and the population size were positively correlated [24].

To the best of our knowledge, past studies on the relationship between spatial characteristics of regions and population changes examined a single or a small number of metropolitan areas of interest. Therefore, it seems to be unknown whether the relationship between spatial characteristics of regions and population changes can be generalized to a large number of metropolitan and non-metropolitan areas, even within a country. To address this question, one needs longitudinal data of population density with a high spatial resolution. Remote sensing technologies and the recent prevalence of mobile phones offer promising data on population dynamics at relatively low cost [25–27]. For example, the spatial distribution of the number of workers estimated from mobile phone data closely matched the counterpart calculated from the US census data [28]. The population density can also be estimated from the amount of night-time lights in satellite imagery [29, 30]. Such data enable estimation of short-term human mobility within a day or week [31, 32].

However, the accuracy of data obtained with these technologies is unclear. Furthermore, the population dynamics estimated by these methods may be susceptible to changes in the accuracy and coverage of the technology over time. In the present study, we use population census data of Japan with a high spatial resolution measured five years apart. To date, census data are probably advantageous to mobile phone or satellite data in tracking long-term population changes with a high accuracy. In fact, census data have been used for evaluating the accuracy of other techniques [28, 29].

We explore spatial factors that contribute to the population growth in local clusters of inhabited areas. We hypothesize that the shape of the cluster of inhabited patches significantly affects the population change in the cluster. To test the hypothesis, we carry out statistical analysis to relate population changes in a cluster over five years, from 2005 to 2010, to the cluster's shape and other demographic and socioeconomic variables. We resolve the aforementioned limitations of the previous studies by exhaustively analyzing clusters of inhabited areas across Japan and by using the census data with which the local populations are accurately estimated.

Materials and methods

Data set

We used data obtained from the population census of Japan in 2005 and 2010; the census is conducted every five years. The data consist of demographic information on a grid of cells of 500 m × 500 m covering the entire Japan [33–35]. There are 1,944,711 cells in total including completely water-surface cells (e.g., sea and lake), of which 4,82,181 cells were populated in 2005 and 477,172 cells in 2010. The population was 127,767,994 in 2005 (65,419,017 females and 62,348,977 males) and 128,057,352 in 2010 (65,729,615 females and 62,327,737 males). The numbers of female inhabitants, that of male inhabitants, and the latitude and longitude of the center of the cell are available for each cell. We denote the year (i.e., 2005 or 2010) by t.

City clustering algorithm

To determine the boundary of an inhabited area, we applied the city clustering algorithm [21, 36–38]. The algorithm calculates the connected components of populated cells, i.e., cells that contain at least one inhabitant, where we have defined the adjacency of cells by the von Neumann neighborhood (i.e., each cell has four neighbors in the north, south, east, and west). To find the connected components, we used the "decompose.graph" function provided by the R package 'igraph' [39]. This function takes a list of the pairs of connected cells and returns the list of the connected components. We refer to each connected component as cluster. We obtained 24165 and 24707 clusters in 2005 and 2010, respectively. In the following analysis, we focused on population changes over time in the clusters identified in 2005, which we denote by c. In other words, we compared the number of inhabitants in each cluster c between 2005 and 2010. It should be noted that we did not use the clusters identified in 2010.

Dependent variable

We denote by $n_i(t)$ the number of inhabitants in cell i at time t. We investigated dynamics of the number of inhabitants in each cluster c identified in 2005 (Fig 1). To this end, we adopted
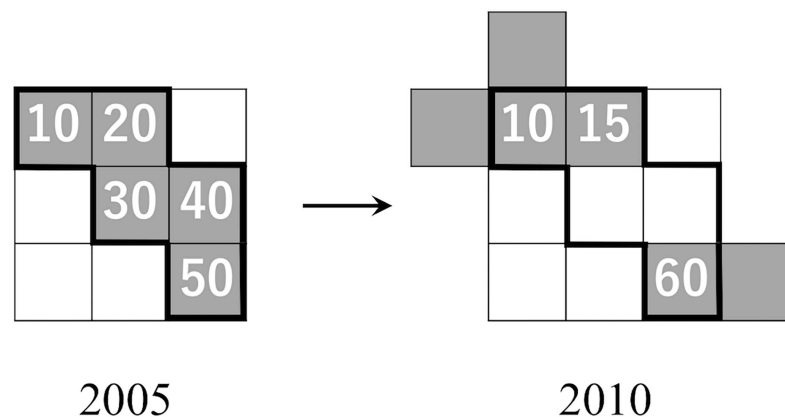
Fig 1. A hypothetical example of the population change in a cluster over five years. The number of inhabitants in a cell is indicated for inhabited cells shown in gray. The bold lines indicate the boundary of cluster c observed in year 2005. This cluster has $n_c^{\text{cluster}}(2005) = \sum_{i \in \text{cluster } c} n_i(2005) = 10 + 20 + 30 + 40 + 50 = 150$ and $\tilde{n}_c^{\text{cluster}}(2010) = \sum_{i \in \text{cluster } c} n_i(2010) = 10 + 15 + 60 = 85$ inhabitants in 2005 and 2010, respectively. Although cluster c is split into different clusters in 2010, each of which extends beyond the border of cluster c determined in 2005, we neglect the split to calculate the population change in cluster c. Therefore, cluster c has lost 150–85 = 65 inhabitants in the five years.

regression models whose dependent variable is defined by

$$\tilde{n}_c^{\text{cluster}}(2010) = \sum_{i \in \text{cluster } c} n_i(2010). \tag{1}$$

In other words, $\tilde{n}_c^{\text{cluster}}(2010)$ is the number of inhabitants in cluster c as of 2010. We used log $(n_c^{\text{cluster}}(2005))$, where

$$n_c^{\text{cluster}}(2005) = \sum_{i \in \text{cluster } c} n_i(2005), \tag{2}$$

i.e., the number of inhabitants in cluster c as of 2005, as the offset variable (see Eq (7)). In this manner, we aimed to compare $\tilde{n}_c^{\text{cluster}}(2010)$ and $n_c^{\text{cluster}}(2005)$, i.e., the number of inhabitants at two time points contained in each cluster c that existed in 2005.

Cells in a cluster c observed in 2005 may belong to different clusters recalculated in 2010. Furthermore, some inhabited cells in 2010 do not belong to any cluster observed in 2005 (Fig 1). Reflecting the latter fact, the total population of Japan in 2005 is equal to $\sum_c n_c^{\text{cluster}}(2005) = 127,767,994$, whereas the sum $\sum_c \tilde{n}_c^{\text{cluster}}(2010) = 127,901,037$ is smaller than the total population of Japan in 2010, where the summation is taken over the clusters identified in 2005. The present definition of cluster may discount the population growth of a cluster when it grew in terms of both the area and the number of inhabitants. This is because the inhabitants that emerged in the area that existed in 2010 but were absent in 2005 were not used in the calculation.

Independent variables

We used the following independent variables for each cluster observed in 2005 to explain the population change between 2005 and 2010.

First, the area of the cluster (denoted by S and referred to as *Area*) is defined by the number of cells constituting the cluster. Second, the population density (referred to as *Density*) is equal to the number of inhabitants in the cluster divided by S.

We quantified the shape of the cluster by the following two indices. We defined what we refer to as *Roundness*, originally proposed in Ref. [40], as S divided by the area of the circle whose diameter is equal to the longest Euclidean distance between two cells belonging to the cluster. We measured the position of a cell by the two-dimensional coordinate of the center of the cell. For example, the clusters shown in Fig 2 have four cells and have the longest Euclidean distance equal to $\sqrt{5}$ (in the unit of the linear length of a cell), yielding a *Roundness* value of 1.019. A cluster whose shape is close to a circle yields a large *Roundness* value. For a given S, the line-shaped cluster yields the smallest *Roundness* value. *Roundness* can be regarded as a simplified variant of the box-counting fractal dimension [41]. The second shape-related index,
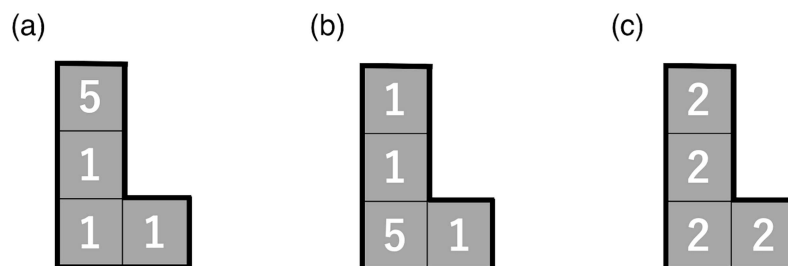

Fig 2. Three clusters, each composed of four cells (i.e., $S = 4$) and eight inhabitants. The number of inhabitants in each cell is indicated.

Irregularity, is defined by

$$\frac{2\log L}{\log S} \tag{3}$$

where *L* is the perimeter of the cluster. For a fixed *S*, *Irregularity* is small when the cluster is close to square-shaped. The perimeter was used for characterizing spatial patterns of urban regions [20]. Frenkel and Ashkenazi [22] applied Eq (3) to quantify the level of urban sprawl. We note that measures similar to *Irregularity* were proposed decades ago [42, 43]. Because of the scaling relation $S = L^{2/Irregularity}$, *Irregularity* can be interpreted as the fractal dimension of the cluster [44, 45].

We quantified the hypothesized efficiency of communication or transportation within a cluster by the following two indices. We defined the expected distance between uniformly randomly selected two inhabitants in the cluster by

$$\frac{\sum_{i \leq j: i,j \in \text{cluster } c} n_i(2005) n_j(2005) d_{ij}}{\binom{n_c^{\text{cluster}(2005)}}{2}} \tag{4}$$

where d_{ij} is the distance between cells *i* and *j*, and the denominator of Eq (4) is a binomial coefficient. It should be noted that $n_i(2005) n_j(2005) / \binom{n_c^{\text{cluster}(2005)}}{2}$ is the probability that randomly selected two inhabitants in cluster *c* belong to cells *i* and *j*. Because Eq (4) has the dimension of the length, it may give rise to multicollinearity with *S* in multivariate regression. To mitigate this potential problem, we divided Eq (4) by $\sqrt{S}$, which has a dimension of the length, to define

$$\frac{\sum_{i \leq j: i,j \in \text{cluster } c} n_i(2005) n_j(2005) d_{ij}}{\binom{n_c^{\text{cluster}(2005)}}{2}} / \sqrt{S} \tag{5}$$

We have assumed the normalization factor of $\sqrt{S}$ to make Eq (5) dimensionless if clusters are two-dimensional (with a large *Roundness* and/or small *Irregularity* value). In fact, clusters may be line-shaped or fractal-like, in which case Eq (5) would have a dimension of the length to some power. However, we expect that Eq (5) is less correlated with *S* than Eq (4) is. Therefore, we adopted Eq (5) as a dependent variable and referred to it as characteristic length (*CL*). We also adopted the coefficient of variation, which is defined by the standard deviation divided by the mean, of the number of inhabitants in a cell belonging to the focal cluster. This index quantifies spatial heterogeneity in the distribution of inhabitants within the cell and is referred to as *Heterogeneity*.

Fig 2 illustrates the difference among *Density*, *CL*, and *Heterogeneity*. The three clusters shown in the figure have the same *Area* (= 4) and *Density* (= 2.00). However, *Heterogeneity* for the clusters shown in Fig 2(A) and 2(B) (= 1.00) is larger than that for the cluster shown in Fig 2(C) (= 0.00). *CL* is smaller for the cluster shown in Fig 2(B) (= 0.623) than that shown in Fig 2(A) (= 0.747), because in Fig 2(B) the most populated cell is located in the center of the cluster. Note that the distribution of the number of inhabitants in a cluster is the same between Fig 2(A) and 2(B). *CL* is the largest for the cluster shown in Fig 2(C) (= 0.874). The distance between the uppermost and the bottom-right cells is equal to $\sqrt{5}$.

We used the following two demographic dependent variables. First, *Gender* refers to the fraction of female inhabitants in the cluster. Second, we estimated the average age of the inhabitants in a cluster, referred to as *Age*, as follows. Because the data set did not contain the average age for each cell, we approximated it by the average age of inhabitants in the prefecture to

which a cluster belongs. The average age of inhabitants in each prefecture is available from the prefecture-level population census data carried out in 2005 [46]. The prefecture of a cluster was defined as the prefecture to which the cell with the largest closeness centrality [47, 48] in the cluster belongs. In the calculation of the closeness centrality, we regarded the cluster as a network in which a cell was a node and two nodes were adjacent if they shared a side. Using the R packages 'rjson'[49] and 'RCurl'[50], we submitted the latitude and longitude of the cell with the largest closeness centrality to the reverse geocoding service provided by National Agriculture and Food Research Organization, Japan [51] and detected the prefecture in which the cell was located. When the reverse geocoding service returned no output because the submitted cell was located in the sea or for other reasons, we used a different data set with which one can determine the prefecture to which cells of 1 km × 1 km belong [52]. The 1 km × 1 km cells in this data set and the 500 m × 500 m cells in the census data were coaligned with each other in the sense that the division of a 1 km × 1 km cell into four cells yielded four 500 m × 500 m cells in the census data. If the 1 km × 1 km cell to which the 500 m × 500 m cell in question belonged to multiple prefectures, we plotted the latitude/longitude of the 500 m × 500 m cell on the map provided by Geospatial Information Authority of Japan [53] and visually determined the prefecture. If multiple cells had the same largest closeness centrality value, we used the average latitude and longitude of these cells to determine the cluster's prefecture.

Although the procedure for calculating *Age* is complicated, we decided to include it in addition to *Gender* for two reasons. First, *Gender* and *Age* are not strongly correlated (see Correlation coefficients section for the result). Second, these variables are likely to impact the birth and death rates in a cell in different ways. As for *Age*, the birth rate is relatively high among women in their twenties and thirties [54]. Therefore, a cluster having a large fraction of individuals in reproductive ages is expected to have a relatively large rate of population growth. However, if the value of *Age* is even larger, the population growth rate within the cluster is expected to be smaller because the death rate increases with age [55, 56]. The value of *Gender* may reflect the efficiency of matching between male and female depending on the sex-ratio balance. The extant results are mixed regarding whether a male-biased or female-biased sex ratio drives marriage squeeze [57, 58]. However, to the least, marriage squeeze may negatively impact the fertility rate [59] especially in countries such as Japan, where people tend to have children after marriage. In fact, that percentage of children born out of wedlock in Japan has been around 2% [60] and much lower than in other countries [61].

As a socioeconomic factor, we used the fraction of workers in the tertiary industry in the prefecture to which the cluster belongs [46] and referred to it as *Tertiary*. We determined the prefecture of a cluster in the same manner as in the case of *Age*.

Regression models

For analysis of count data, a Poisson regression model is often used (e.g., [62]). This model assumes that the dependent variable ($\tilde{n}_c^{\text{cluster}}(2010)$ in the present case) obeys a Poisson distribution given by

$$\Pr(\tilde{n}_c^{\text{cluster}}(2010) = k) = \frac{\exp(-\mu_c)\mu_c^k}{k!} \tag{6}$$

where the conditional mean μ_c is determined by

$$\log(\mu_c) = \log(n_c^{\text{cluster}}(2005)) + \beta_0 + \beta_1\log(Area_c) + \beta_2\log(Density_c) + \sum_{i=3}^{9}\beta_i X_{i,c} \tag{7}$$

In Eq (7), Eq (2) is used as the offset variable, the logarithmic link function is used, β_0 is the intercept, β_i ($i = 1,\ldots, 9$) is a regression coefficient, X_i ($i = 3,\ldots, 9$) is the ith independent

variable (i.e., *Roundness*, *Irregularity*, *CL*, *Heterogeneity*, *Gender*, *Age*, and *Tertiary*), and subscript c on the right-hand side indicates that the values of the independent variables are for cluster c.

In the Poisson regression model, the conditional mean of the dependent variable is assumed to be equal to its conditional variance. However, as we will show in Descriptive statistics section, the conditional variance of the dependent variable is considerably larger than its conditional mean for the present data. This situation is called the overdispersion, which we tested by running an overdispersion test [63, 64] (see also [65] for the usage of the R package 'AER'). The overdispersion test is carried out based on the statistic

$$\frac{\sum_{c=1}^{N}((\tilde{n}_c^{\text{cluster}}(2010) - \hat{\mu}_c)^2 - \tilde{n}_c^{\text{cluster}}(2010))}{\sqrt{2\sum_{c=1}^{N}\hat{\mu}_c^2}} \tag{8}$$

which asymptotically obeys the normal distribution with mean 0 and standard deviation 1 under the assumption of the Poisson model. In Eq (8), $\hat{\mu}_c$ is the maximum likelihood estimate of the dependent variable under the Poisson model (i.e., the null hypothesis).

Because the null hypothesis was rejected (Descriptive statistics section), we used the negative binomial regression model. A negative binomial regression model [62] assumes that the dependent variable obeys a negative binomial distribution given by

$$\Pr(\tilde{n}_c^{\text{cluster}}(2010) = k) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)\Gamma(k+1)}\left(\frac{\theta}{\mu_c+\theta}\right)^{\theta}\left(\frac{\mu_c}{\mu_c+\theta}\right)^{k} \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function, and θ is a parameter that is assumed to be the same for all clusters. In Eq (9), the conditional mean, μ_c, is given by Eq (7). The variance of the distribution given by Eq (9) is $\mu_c + \mu_c^2/\theta$. To fit the model, we maximized the likelihood with respect to β_i ($i = 0, \ldots, 9$) (Eq (7)) and θ using the glm.nb() function in the R package 'MASS' [66].

The *Area* and *Density* variables obeyed long-tailed distributions (Fig 3(A) and 3(C); also see Descriptive statistics section). Therefore, in Eq (7), we logarithmically transformed *Area* and *Density* to improve linearity between the dependent and independent variables. In fact, the logarithm of *Area* obeyed a much less long-tailed distribution (Fig 3(B)), and the logarithm of *Density* obeyed a distribution that roughly looks like a normal distribution (Fig 3(D)). For these two independent variables, a 1% increase in *Area* (or *Density*) corresponds to a β_1 (or β_2) % increase in the number of inhabitants in 2010 in a cluster observed in 2005. For X_i ($i = 3, \ldots, 9$), an increase in X_i by one unit increases the number of inhabitants $\exp(\beta_i)$ times. The distributions of these independent variables are shown in Fig 3(E)–3(K). We used the same offset term Eq (2) in the multivariate and univariate regressions.

We also searched the multivariate regression model that minimized the Akaike information criterion (AIC) among the models that had any of the independent variables as main effects and any of pairwise interaction terms between the independent variables. To avoid large variance inflation factor (VIF) values due to the pairwise interaction terms, we normalized all independent variables to have a zero mean [67]. We used the stepwise backward elimination method to find the best model, i.e., by sequentially excluding the least significant term in terms of the AIC [68].

Results

Descriptive statistics

Statistics of the dependent, offset, and independent variables are shown in Table 1. We find that the area of a cluster, S, the number of inhabitants in a cluster, and the population density
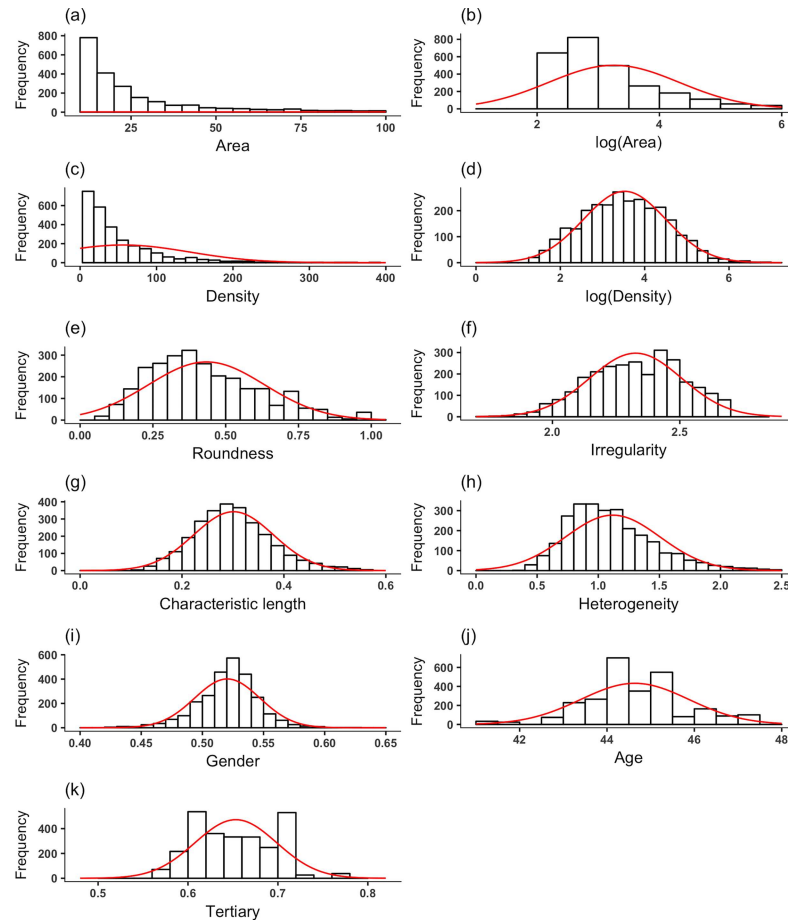
Fig 3. Distributions of the independent variables. (a) *Area*. **(b)** log(*Area*). **(c)** *Density*. **(d)** log(*Density*). **(e)** *Roundness*. **(f)** *Irregularity*. **(g)** *Characteristic length* (*CL*). **(h)** *Heterogeneity*. **(i)** *Gender*. **(j)** *Age*. **(k)** *Tertiary*. The clusters whose *Area* was less than 10 were omitted from the calculation of the distributions. Some of the distributions are truncated for a visibility reason. The curve shown in each panel represents the normal distribution with the sample mean and standard deviation.

https://doi.org/10.1371/journal.pone.0197144.g003

in a cluster are heterogeneously distributed, as suggested by large coefficient of variation (CV) values for these variables. Moreover, the skewness for these variables is large. This observation is confirmed by long-tailed distributions of these quantities shown in Fig 4.

In the following statistical analysis, we restricted ourselves to the clusters whose areas are at least ten cells because the geometry of smaller clusters would be strongly affected by the spatial discreteness.

We ran the overdispersion test to confirm that the assumption of the Poisson distribution of the dependent variable was violated ($p < 0.001$). Therefore, in the following we report the result of the negative binomial regression model.

Correlation coefficients

The Pearson, Spearman, and Kendall correlation coefficients between pairs of independent variables are shown in Tables 2–4, respectively. The signs of almost all of the correlation coefficients are consistent across the three types of correlation coefficient.

Table 2 indicates that log(*Area*) and *Irregularity* are strongly correlated (Pearson correlation coefficient = −0.794). This result is consistent with the positive correlation previously observed

Table 1. Descriptive statistics for the clusters composed of at least ten cells.

	Min	Max	Median	Mean	SD	CV	Skewness
Dependent variable							
$\tilde{n}_c^{\text{cluster}}(2010)$	22	42492718	623	47193.57	964925.07	20.446	35.962
Offset							
$n_c^{\text{cluster}}(2005)$	39	41403322	696	47086.07	946332.04	20.098	35.551
Independent variables							
log(*Area*)	2.303	11.187	2.944	3.256	1.070	0.329	2.375
log(*Density*)	1.291	7.684	3.509	3.526	0.977	0.277	0.172
Roundness	0.066	1.273	0.397	0.434	0.200	0.460	0.877
Irregularity	1.746	2.685	2.336	2.327	0.181	0.078	−0.181
CL	0.110	0.720	0.296	0.301	0.078	0.260	0.710
Heterogeneity	0.342	3.190	1.052	1.119	0.387	0.346	1.308
Gender	0.236	0.694	0.523	0.520	0.027	0.051	−1.179
Age	39.1	47.1	44.4	44.6	1.240	0.028	−0.803
Tertiary	0.570	0.774	0.653	0.653	0.045	0.070	0.185
Area	10	72194	19	160.50	1896.4	11.816	26.795
Density	3.636	2172.5	33.40	56.19	86.59	1.541	10.636

There are *N* = 2689 clusters. The clusters whose *Area* was less than 10 were omitted from the calculation of the statistics shown in the table. By definition, all independent variables are calculated for the data in 2005. SD represents the standard deviation. CV represents the coefficient of variation. The non-standardized *Area* and *Density* variables are not used in the regression models but shown for completeness.

between the city size and the spatial compactness of the city measured by a fractal dimension [24, 69]. However, we concluded that the multicollinearity problem was not present because the VIF values were sufficiently small (4.206 and 3.247 for log(*Area*) and *Irregularity*, respectively). In general, VIF values for independent variables should be less than 10, preferably less than 5, for multivariate regression analysis to be justified [70, 71].

Regression analysis

The results of the negative binomial regression are shown in Table 5. The contributions of log (*Area*) and log(*Density*) were significant at the 0.1% level, *Irregularity* and *Age* at the 1% level, and *CL* at the 10% level. The other variables (i.e., *Roundness*, *Heterogeneity*, *Gender*, and *Tertiary*) were not significant. Table 5 also indicates that a 1% increase in *Area* and *Density* is associated with an increase in the number of inhabitants in a cluster in 2010 (as compared to 2005) by 0.0113% and 0.0227%, respectively. An increase in *Irregularity*, *Age*, and *CL* by 1% is associated with a decrease in the number of inhabitants in a cluster by 3.27×10^{-4} (= 1−exp(−0.0327×0.01)) times, 2.40×10^{-5} (= 1−exp(−0.0024×0.01)) times, and 3.62×10^{-4} (= 1−exp(−0.0362×0.01)) times, respectively. Because the total population in Japan only changed by 0.23% between 2005 and 2010 (Data set section), the contribution of these factors to the population change is non-negligible.

The results for univariate regressions are also shown in Table 5. The signs of all the significant regression coefficients in the multivariate regression (i.e., negative binomial regression) were consistent with the results for the univariate regression, lending support to the results obtained from the multivariate analysis.

We carried out the model selection in terms of the Akaike Information Criterion (AIC) among the negative binomial regression models that were allowed to include any main effects and pairwise interaction terms. The regression coefficients of the selected model are shown in
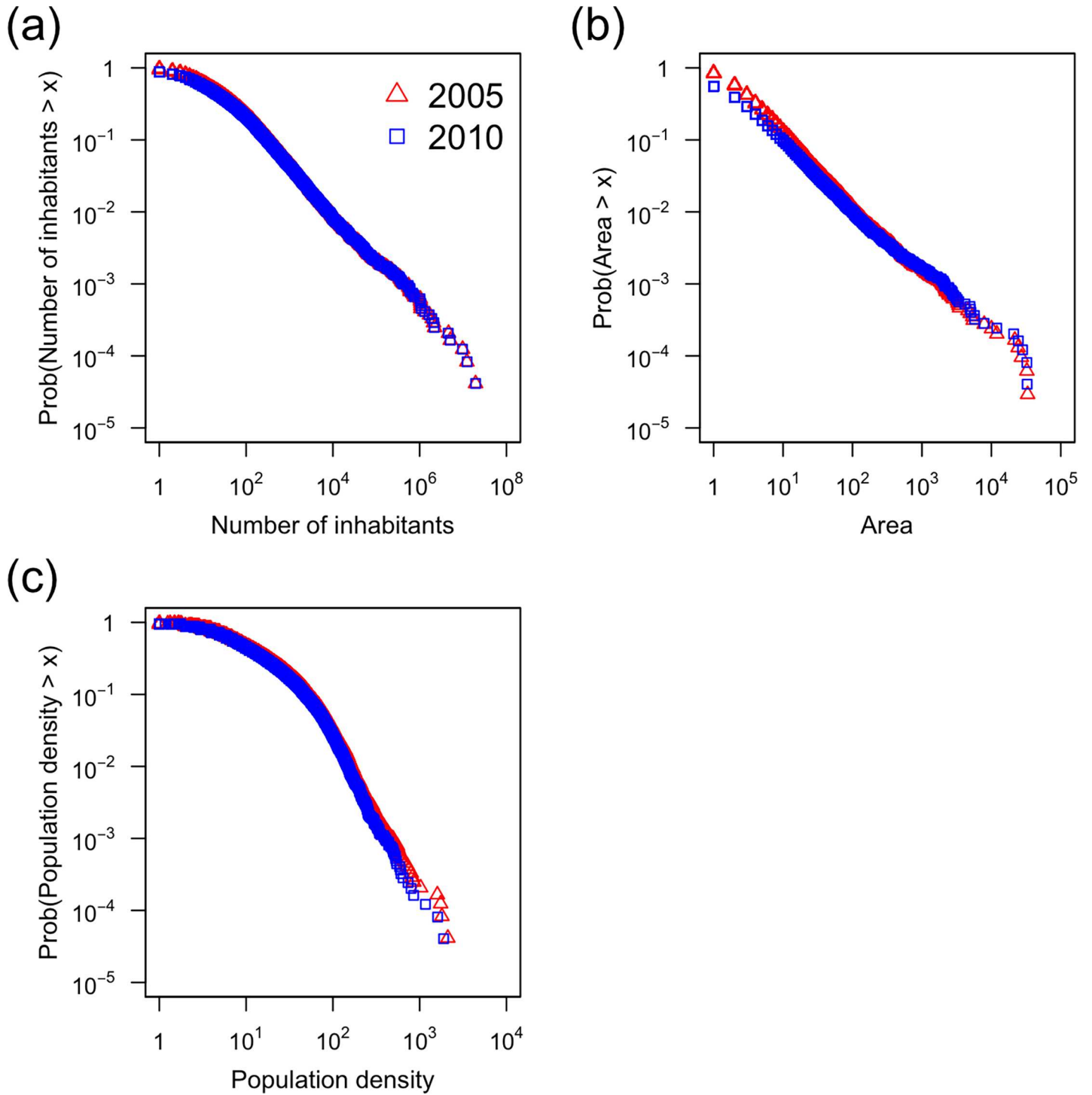
Fig 4. Complementary cumulative distributions for three properties of a cluster. (a) Number of inhabitants. **(b)** Area. **(c)** Population density. Differently from Fig 3, we used all the clusters to calculate the distributions.

Table 6. The selected model contained all independent variables. The result that the main effects of log(*Area*), log(*Density*), and *CL* are significant is consistent with that for the multivariate regression. However, the main effects of *Irregularity* and *Age*, which were significant in the

Table 2. Pearson correlation coefficient between the independent variables for the clusters with at least ten cells observed in 2005.

	log(*Area*)	log(*Density*)	*Roundness*	*Irregularity*	*CL*	*Heterogeneity*	*Gender*	*Age*	VIF
log(*Area*)	–								4.206
log(*Density*)	0.431***	–							1.536
Roundness	−0.479***	−0.167***	–						2.163
Irregularity	−0.794***	−0.494***	0.230***	–					3.247
CL	0.232**	−0.132***	−0.594***	0.009	–				2.353
Heterogeneity	0.516***	0.361***	−0.215***	−0.460***	−0.265***	–			1.943
Gender	0.062	0.172**	−0.051**	−0.067***	0.003	0.056**	–		1.065
Age	−0.002**	−0.068***	−0.014	0.010	0.008	0.007	0.160***	–	1.153
Tertiary	−0.004	−0.111	0.042**	−0.007	−0.135***	0.149***	−0.068***	−0.291***	1.190

$N = 2689$.

**$p < 0.01$

***$p < 0.001$.

https://doi.org/10.1371/journal.pone.0197144.t002

Table 3. Spearman rank correlation coefficient between the independent variables for the clusters with at least ten cells observed in 2005.

	log(*Area*)	log(*Density*)	*Roundness*	*Irregularity*	*CL*	*Heterogeneity*	*Gender*	*Age*
log(*Area*)	–							
log(*Density*)	0.373***	–						
Roundness	−0.560***	−0.204***	–					
Irregularity	−0.870***	−0.456***	0.273***	–				
CL	0.219***	−0.173***	−0.616***	0.049**	–			
Heterogeneity	0.473***	0.437***	−0.218***	−0.466***	−0.307***	–		
Gender	0.084***	0.192***	−0.047**	−0.089***	−0.022	0.101***	–	
Age	0.024	−0.022	−0.003	−0.023	−0.012	0.044**	0.162***	–
Tertiary	0.002	−0.113***	0.037†	−0.003	−0.115***	0.092***	−0.020	−0.224***

$N = 2689$.

†$p < 0.1$

**$p < 0.01$

***$p < 0.001$.

https://doi.org/10.1371/journal.pone.0197144.t003

Table 4. Kendall rank correlation coefficient between the independent variables for the clusters with at least ten cells observed in 2005.

	log(*Area*)	log(*Density*)	*Roundness*	*Irregularity*	*CL*	*Heterogeneity*	*Gender*	*Age*
log(*Area*)	–							
log(*Density*)	0.260***	–						
Roundness	−0.395***	−0.138***	–					
Irregularity	−0.700***	−0.318***	0.181***	–				
CL	0.151***	−0.115***	−0.446***	0.033**	–			
Heterogeneity	0.333***	0.297***	−0.146***	−0.323***	−0.210***	–		
Gender	0.057***	0.129***	−0.032**	−0.058***	−0.014	0.067***	–	
Age	0.017	−0.016	−0.003	−0.016	−0.007	0.028**	0.110***	–
Tertiary	0.002	−0.079***	0.026†	−0.002	−0.079***	0.063***	−0.010	−0.171***

$N = 2689$.

†$p < 0.1$

**$p < 0.01$

***$p < 0.001$.

https://doi.org/10.1371/journal.pone.0197144.t004

Table 5. Coefficients of multivariate and univariate negative binomial regressions.

	Multivariable		Univariate	
	Estimate	CI	Estimate	CI
(Intercept)	−0.0146	[−0.1419, 0.1126]	–	–
log(*Area*)	0.0113***	[0.0076, 0.0151]	0.0202***	[0.0183, 0.0222]
log(*Density*)	0.0227***	[0.0198, 0.0257]	0.0320***	[0.0294, 0.0347]
Roundness	0.0040	[−0.0128, 0.0208]	−0.0414***	[−0.0549, −0.0279]
Irregularity	−0.0327**	[−0.0552, −0.0101]	−0.1428***	[−0.1565, −0.1292]
CL	−0.0362†	[−0.0762, 0.0039]	−0.0302†	[−0.0619, 0.0016]
Heterogeneity	−0.0007	[−0.0084, 0.0070]	0.0375***	[0.0310, 0.0441]
Gender	−0.0433	[−0.1442, 0.0576]	0.0821	[−0.0309, 0.1951]
Age	−0.0024**	[−0.0043, −0.0005]	−0.0039***	[−0.0059, −0.0018]
Tertiary	−0.0098	[−0.0650, 0.0455]	0.0367	[−0.0221, 0.0956]

$N = 2689$. CI: 95% confidence interval.

†$p < 0.1$

**$p < 0.01$

***$p < 0.001$.

multivariate regression, were not significant in the selected model, while some interaction effects between other variables and *Irregularity* or *Age* were significant. This result implies that the effects of *Irregularity* and *Age* qualitatively depend on other variables. Lastly, the main effect of *Heterogeneity*, which was not significant in the multivariate regression, was significant in the selected model.

On the basis of the results for the multivariate regression, univariate regression, and model selection, we conclude that the main effects of *Area*, *Density*, and *CL* are significant according to the different criteria. In other words, the population growth of a cluster is associated with an

Table 6. The selected model.

	Estimate		Estimate
(Intercept)	−0.1202***	log(*Area*) × log(*Density*)	−0.0035**
log(*Area*)	0.0205***	log(*Area*) × *Gender*	−0.4752***
log(*Density*)	0.0230***	log(*Area*) × *Age*	0.0022*
Roundness	0.0067	log(*Area*) × *Tertiary*	0.0834**
Irregularity	−0.0127	*Density* × *Gender*	−0.1239*
CL	−0.0660**	*Density* × *Age*	−0.0019
Heterogeneity	−0.0100*	*Density* × *Tertiary*	−0.0537
Gender	−0.0592	*Roundness* × *Gender*	−1.0902***
Age	−0.0008	*Irregularity* × *Gender*	−1.2191*
Tertiary	−0.0464	*CL* × *Age*	−0.0316*
		CL × *Tertiary*	−1.1941***
		Age × *Tertiary*	−0.0638***

$N = 2689$. AIC = 29062.

*$p < 0.05$

**$p < 0.01$

***$p < 0.001$.

increase in *Area*, an increase in *Density*, and a decrease in *CL*. In addition, the main effects of *Irregularity* and *Age* were also significant in the multivariate and univariate regression (but not in the model selected by the AIC).

Discussion

Summary

We searched for potential drivers of population changes in terms of demographic, geometrical, and other properties of a cluster of inhabited cells. Unsurprisingly, we found that the area and the population density of the cluster were positively correlated with the population growth rate.

In addition, we found that a shape parameter for the cluster, *Irregularity*, and the mean distance between inhabitants within the cluster, *CL*, had negative impacts on the population growth. *Age* also had a negative impact on the population growth. In contrast, the fraction of female inhabitants, *Gender*, and that of tertiary-industry workers, *Tertiary*, had no significant contribution. The present results suggest that the population change is predictable to a certain degree from spatial characteristics intrinsic to the cluster, irrespectively of demographic factors.

Effects of variables characterizing the shape and heterogeneity of a cluster

Roundness was significantly correlated with *Area*. This result is inconsistent with the previous result showing no significant correlation between the city size and the anisometry, where the anisometry was defined by the ratio of the length of the major axis and that of the minor axis of the ellipse including the city cluster [69] and hence similar to *Roundness*. This inconsistency may originate from the different terrains in different cities and countries, the pattern of centralization of the population to urban areas of Japan such as Tokyo [72] in the present study, or other reasons; we do not have a clear explanation. Because urban sprawl is often negatively associated with the compact city [17, 22], it is intriguing to associate urban sprawl with *Roundness* or *Irregularity*. However, urban sprawl is not solely characterized by the shape of urban areas but also by a discontinuous development of suburban areas, which may reduce the intra- and inter-region accessibility [13]. To relate our approach to urban sprawl, we probably need to consider relationships between different clusters and the role of each cluster in wider geographical regions.

CL had a negative impact on the population growth rate. By definition, *CL* is small when highly populated cells are located near the geographical center of a cluster (Fig 2(B)) rather than when they are located in the periphery of the cluster (Fig 2(A)). Therefore, our results suggest that a cluster's population tends to grow if many inhabitants are located near the center of the cluster. A previous study showed that the values of indices characterizing urban regions (e.g., Moran, Geary, and Gini coefficients) were sensitive to the distribution of inhabitants in a confined region [73]. The present study suggests that the spatial distribution of inhabitants may affect the population growth rate as well as such urban indices. Investigating this issue warrants future work.

We did not pay attention to the change in the shape of the cluster over years. In fact, processes of urban growth, which are characterized by, for example, the population size, economic performance, and development of transportation systems, occur in tandem with changes in the shape of urban areas (e.g., [23, 74, 75]). Socioeconomic factors reflected in the shape of urban areas may influence inhabitants' residential decision making, which may in turn change the shape of urban areas.

Effects of the population size of a cluster

We used the population of the cluster in 2005 as an offset variable, not independent variable. We additionally analyzed the following linear regression model with the population of the cluster in 2005 as an independent variable: $\log(\tilde{n}_c^{\text{cluster}}(2010)/n_c^{\text{cluster}}(2005)) = \beta\log(n_c^{\text{cluster}}(2005))$. The population of the cluster in 2005 was positively correlated with the growth rate in the cluster over the five years ($\beta = 0.021$, $p < 0.001$). This result is inconsistent with the previous studies showing a smaller growth rate for clusters with a larger population [36] and the lack of correlation between the population of administratively defined cities and their growth rate [76]. The reason for this discrepancy is unclear. It may be because of the different definitions of the cluster change in the two studies or the aforementioned centralization of the inhabitants to urban areas of Japan.

Comparison with the gravity model

The gravity model and its variants explain spatio-temporal migration and population changes in various data [77–79]. The statistical explanation of population changes that we have found is different from the mechanisms implied by the gravity model and its variants.

First, let us assume that the unit of analysis is a cluster. Then, the gravity model assumes that migratory population flows are influenced by the attractiveness (often identified with the number of inhabitants) of the origin cluster and the destination cluster. In contrast, we ignored any interaction between clusters. Therefore, in our statistical approach, the rate of population change does not depend on the population of different clusters, differently from the prediction obtained by the gravity model. We neglected effects of other clusters because we did not have migration data. However, this decision does not imply that migration effects are unimportant (see the Limitations section below for more discussion).

Second, the proposed mechanism is also different from that provided by the gravity model even if one uses a single cell as the unit of analysis and applies the gravity model to population dynamics within a cluster. Given the shape of a cluster, we found that the *CL* negatively impacted on the population growth rate. This result implies that the spatial distribution of inhabitants within a cluster affects the population growth rate. In contrast, the gravity model applied to population dynamics within a cluster would describe migration dynamics within a cluster. Because intra-cluster migration implies that the number of inhabitants is preserved over time, the gravity model applied to a cluster would not predict whether the population of the cluster tends to increase or decrease. In sum, the present analysis is orthogonal to what the gravity model aims to explain.

Limitations

An important limitation of the present study is that we did not have an access to migration data. In general, the population change is decomposed into the natural increase (i.e., births minus deaths) and the migratory increase (i.e., immigration minus emigration). Because the census data used in the present study did not include the information about the population flow, we could not distinguish between the natural and migratory increases. Another limitation is that some dependent variables (i.e., *Age* and *Tertiary*) were estimated at the prefecture level due to the lack of data at the level of single cells.

We did not consider other information such as land use as independent variables, either. For example, steeper slopes and higher elevations negatively impact on urban expansion [80, 81]. Regarding transportation systems, the distance to highways and major roads negatively impact on urban expansion [80]. Network structures of transportation systems are also related

to the urbanization [74, 81]. For example, the treeness of street networks is negatively correlated with the metropolitan population [81]. Urban planning is also an important factor driving urban expansion. For example, Ref. [82] evaluated effects of urban master plans on urban expansion in Beijing between 1947 to 2008 and showed that the effects were positive in all periods. Further longitudinal analyses including any of these variables with an appropriate spatial resolution will be valuable.

Author Contributions

Conceptualization: Naoki Masuda.

Data curation: Takuya Sekiguchi, Kohei Tamura.

Formal analysis: Takuya Sekiguchi.

Investigation: Takuya Sekiguchi, Kohei Tamura, Naoki Masuda.

Writing – original draft: Takuya Sekiguchi, Naoki Masuda.

Writing – review & editing: Takuya Sekiguchi, Kohei Tamura, Naoki Masuda.

References

1. Schwanen T, Dijst M, Dieleman FM. Policies for urban form and their impact on travel: the Netherlands experience. Urban Stud. 2004; 41: 579–603.

2. McDonald RI, Green P, Balk D, Fekete BM, Revenga C, Todd M, et al. Urban growth, climate change, and freshwater availability. Proc Natl Acad Sci USA. 2011; 108: 6312–6317. https://doi.org/10.1073/pnas.1011615108 PMID: 21444797

3. Tanaka K, Imai M. A review of recent transportation geography in Japan. Geogr Rev Jpn Ser B. 2013; 86: 92–99.

4. Tanaka K, Iwasawa M. Aging in rural Japan: limitations in the current social care policy. J Aging Soc Policy. 2010; 22: 394–406. https://doi.org/10.1080/08959420.2010.507651 PMID: 20924894

5. Prillwitz J, Harms S, Lanzendorf M. Interactions between residential relocations, life course events, and daily commute distances. Transp Res Rec. 2007; 2021: 64–69.

6. Chen J, Chen C, Timmermans H. Accessibility trade-offs in household residential location decisions. Transp Res Rec. 2008; 2077: 71–79.

7. Lee BH, Waddell P. Residential mobility and location choice: a nested logit model with sampling of alternatives. Transportation (Amst). 2010; 37: 587–601.

8. Nechyba TJ, Strauss RP. Community choice and local public services: a discrete choice approach. Reg Sci Urban Econ. 1998; 28: 51–73.

9. Bayoh I, Irwin EG, Haab T. Determinants of residential location choice: how important are local public goods in attracting homeowners to central city locations? J Reg Sci. 2006; 46: 97–120.

10. Weisbrod G, Lerman SR, Ben-Akiva M. Tradeoffs in residential location decisions: transportation versus other factors. Transportation Policy and Decision-Making. 1980; 1: 13–26.

11. Clark WAV, Huang Y. The life course and residential mobility in British housing markets. Environ Plan A. 2003; 35: 323–339.

12. Harvey RO, Clark WAV. The nature and economics of urban sprawl. Land Econ. 1965; 41: 1–9.

13. Ewing R. Is Los Angeles-style sprawl desirable? J Am Plann Assoc. 1997; 63: 107–126.

14. Galster G, Hanson R, Ratcliffe MR, Wolman H, Coleman S, Freihage J. Wrestling sprawl to the ground: defining and measuring an elusive concept. Hous Policy Debate. 2001; 12: 681–717.

15. Johnson MP. Environmental impacts of urban sprawl: a survey of the literature and proposed research agenda. Environ Plan A. 2001; 33: 717–735.

16. Dieleman F, Wegener M. Compact city and urban sprawl. Buil Environ. 2004; 30: 308–323.

17. Bhatta B. Analysis of urban growth and sprawl from remote sensing data. Heidelberg: Springer-Verlag; 2010.

18. Cervero R. Jobs-housing balancing and regional mobility. J Am Plann Assoc. 1989; 55: 136–150.

19. Srinivasan S. Quantifying spatial characteristics of cities. Urban Stud. 2002; 39: 2005–2028.

20. Longley PA, Batty M, Shepherd J. The size, shape and dimension of urban settlements. Trans Inst Br Geogr. 1991; 16: 75–94.

21. Fluschnik T, Kriewald S, Ros AGC, Zhou B, Reusser DE, Kropp JP, et al. The size distribution, scaling properties and spatial organization of urban clusters: a global and regional percolation perspective. ISPRS Int J Geoinf. 2016; 5: 110.

22. Frenkel A, Ashkenazi M. Measuring urban sprawl: how can we deal with it? Environ Plann B Plann Des. 2008; 35:56–79.

23. Benguigui L, Czamanski D, Marinov M, Portugali Y. When and where is a city fractal? Environ Plann B Plann Des. 2000; 27: 507–519.

24. Shen G. Fractal dimension and fractal growth of urbanized areas. Int J Geogr Inf Sci. 2002; 16: 419–437.

25. Wu C, Murray AT. Population estimation using Landsat enhanced thematic mapper imagery. Geogr Anal. 2007; 39: 26–43.

26. Manfredini F, Pucci P, Tagliolato P. Mobile phone network data: new sources for urban studies. In: Borruso G, Bertazzon S, Favretto A, Murgante B, Torre CM, editors. Geographic information analysis for sustainable development and economic planning: new technologies. Hershey: IGI Global; 2012. pp 115–128.

27. Pucci P, Manfredini F, Tagliolato P. Mobile phone data to describe urban practices: an overview in the literature. In: Pucci P, Manfredini F, Tagliolato P, editors. Mapping urban practices through mobile phone data. Heidelberg: Springer; 2015. pp 13–25.

28. Becker RA, Caceres R, Hanson K, Loh JM, Urbanek S, Varshavsky A, et al. A tale of one city: using cellular network data for urban planning. IEEE Pervasive Comput. 2011; 10: 18–26.

29. Sutton P, Roberts D, Elvidge C, Meij H. A comparison of nighttime satellite imagery and population density for the continental United States. Photogramm Eng Remote Sensing. 1997; 63: 1303–1313.

30. Sutton P, Roberts D, Elvidge C, Baugh K. Census from Heaven: an estimate of the global human population using night-time satellite imagery. Int J Remote Sens. 2001; 22: 3061–3076.

31. Ahas R, Silm S, Saluveer E, Järv O. Modelling home and work locations of populations using passive mobile positioning data. In: Gartner G, Rehrl K, editors. Location based services and telecartography II: from sensor fusion to context models. Heidelberg: Springer; 2008. pp 301–315.

32. Ahas R, Aasa A, Silm S, Tiru M. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. Transp Res Part C Emerg Technol. 2010; 18: 45–54.

33. Statistics Bureau of Japan. Heisei 17 nen kokusei chosa (sekai sokuchikei 500m mesh). 2005. Available from: http://e-stat.go.jp/SG2/eStatGIS/page/download.html.

34. Statistics Bureau of Japan. Heisei 22 nen kokusei chosa (sekai sokuchikei 500m mesh). 2010. Available from: http://e-stat.go.jp/SG2/eStatGIS/page/download.html.

35. Tamura K, Masuda N. Effects of the distant population density on spatial patterns of demographic dynamics. Roy Soc Open Sci. 2017; 4: 170391.

36. Rozenfeld HD, Rybski D, Andrade JS, Batty M, Stanley HE, Makse HA. Laws of population growth. Proc Natl Acad Sci USA. 2008; 105: 18702–18707. https://doi.org/10.1073/pnas.0807435105 PMID: 19033186

37. Rozenfeld HD, Rybski D, Gabaix X, Makse HA. The area and population of cities: new insights from a different perspective on cities. Am Econ Rev. 2011; 101: 2205–2225.

38. Rybski D, Ros AGC, Kropp JP. Distance-weighted city growth. Phys Rev E. 2013; 87: 042114.

39. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006; 1695.

40. Reock EC. A note: measuring compactness as a requirement of legislative apportionment. Midw J Pol Sci. 1961; 5: 70–74.

41. Batty M, Longley PA. Fractal cities: a geometry of form and function. London: Academic Press; 1994.

42. Cox EP. A method of assigning numerical and percentage values to the degree of roundness of sand grains. J Paleontol. 1927; 1: 179–183.

43. Schwartzberg JE. Reapportionment, gerrymanders, and the notion of compactness. Minn Law Rev. 1966; 50:443–452.

44. Lovejoy S. Area-perimeter relation for rain and cloud areas. Science. 1982; 216: 185–187. https://doi.org/10.1126/science.216.4542.185 PMID: 17736252

45. Mandelbrot BB. The fractal geometry of nature. New York: WH freeman; 1983.

46. Statistics Bureau of Japan. Population of Japan (Final Report of the 2005 Population Census). 2010. Available from: http://www.e-stat.go.jp/SG1/estat/ListE.do?bid=000001025191&cycode=0.

47. Beauchamp MA. An improved index of centrality. Behav Sci. 1965; 10: 161–163. PMID: 14284290

48. Newman MEJ. Networks: an introduction. Oxford: Oxford University Press; 2010.

49. Couture-Beil A. rjson: JSON for R. 2014.

50. Lang DT, the CRAN team. RCurl: General Network (HTTP/FTP/ . . .) Client Interface for R. 2016.

51. National Agriculture and Food Research Organization. Simple reverse geocoding service. Available from: https://www.finds.jp/rgeocode/index.html.en

52. Statistics Bureau of Japan. Shi ku cho son betsu mesh code ichiran. 2015. Available from: http://www.stat.go.jp/data/mesh/m_itiran.htm

53. Geospatial Information Authority of Japan. Chiriinn chizu. Available from: https://maps.gsi.go.jp/

54. Human Fertility Database (HFD). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available from: https://www.humanfertility.org

55. Wang H, Dwyer-Lindgren L, Lofgren KT, Rajaratnam JK, Marcus JR, Levin-Rector A, et al. Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012; 380: 2071–2094. https://doi.org/10.1016/S0140-6736(12)61719-X PMID: 23245603

56. Statistics Bureau of Japan. Nenrei betsu shibousuu oyobi shibouritsu. 2015. Available from: https://www.stat.go.jp/data/nenkan/back64/zuhyou/y0226000.xls

57. Schacht R, Kramer KL. Patterns of family formation in response to sex ratio variation. PLoS One. 2016; 11: e0160320. https://doi.org/10.1371/journal.pone.0160320 PMID: 27556401

58. Schacht R, Smith KR. Causes and consequences of adult sex ratio imbalance in a historical US population. Philos T Roy Soc B. 2017; 372: 20160314.

59. Heer DM, Grossbard-Shechtman A. The impact of the female marriage squeeze and the contraceptive revolution on sex roles and the women's liberation movement in the United States, 1960 to 1975. J Marriage Fam. 1981; 43: 49–65.

60. Statistics Bureau of Japan. Seinenjibetsu ni mita chakushutsu denai ko no shusseisuu oyobi wariai. 2018. Available from: https://www.e-stat.go.jp/dbview?sid=0003214687

61. OECD. Share of births outside of marriage. 2016. Available from: http://www.oecd.org/els/family/SF_2_4_Share_births_outside_marriage.pdf

62. Dobson AJ, Barnett A. An introduction to generalized linear models. Boca Raton: CRC Press; 2008.

63. Cameron AC, Trivedi PK. Econometric models based on count data: comparisons and applications of some estimators and tests. J Appl Econ. 1986; 1: 29–53.

64. Yang Z, Hardin JW, Addy CL, Vuong QH. Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. Biom J. 2007; 49: 565–584. https://doi.org/10.1002/bimj.200610340 PMID: 17638291

65. Kleiber C, Zeileis A. Applied econometrics with R. New York: Springer-Verlag; 2008.

66. Venables WN, Ripley BD. Modern applied statistics with S. New York: Springer; 2002.

67. Cronbach LJ. Statistical tests for moderator variables: flaws in analyses recently proposed. Psychol Bull. 1987; 102: 414–417.

68. Han J, Kamber M. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann; 2000.

69. Zhou B, Rybski D, Kropp JP. The role of city size and urban form in the surface urban heat island. Sci Rep. 2017; 7: 4791. https://doi.org/10.1038/s41598-017-04242-2 PMID: 28684850

70. Stine RA. Graphical interpretation of variance inflation factors. Am Stat. 1995; 49: 53–56.

71. Tufféry S. Data mining and statistics for decision making. Chichester: Wiley; 2011.

72. Statistics Bureau of Japan. Population of Japan. Heisei 27 nen kokusei chosa (jinkou sokuhou shukei kekka) Available from: www.stat.go.jp/data/kokusei/2015/kekka/pdf/gaiyou.pdf

73. Tsai YH. Quantifying urban form: compactness versus 'sprawl'. Urban Stud. 2005; 42: 141–161.

74. Lu Y, Tang J. Fractal dimension of a transportation network and its relationship with urban growth: a study of the Dallas-Fort Worth area. Environ Plann B Plann Des. 2004; 31: 895–911.

75. Batty M. Building a science of cities. Cities. 2012; 29:S9–S16.

76. Devadoss S, Luckstead J. Growth process of US small cities. Econ Lett. 2015; 135: 12–14.

77. Simini F, González MC, Maritan A, Barabási AL. A universal model for mobility and migration patterns. Nature. 2012; 484: 96–100. https://doi.org/10.1038/nature10856 PMID: 22367540

78. Masucci AP, Serras J, Johansson A, Batty M. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. Phys Rev E. 2013; 88: 022812.

79. Batty M. The new science of cities. Cambridge, MA: MIT Press; 2013.

80. Li X, Zhou W, Ouyang Z. Forty years of urban expansion in Beijing: what is the relative importance of physical, socioeconomic, and neighborhood factors? Appl Geogr. 2013; 38: 1–10.

81. Levinson D. Network structure and city size. PLoS One. 2012; 7: e29721. https://doi.org/10.1371/journal.pone.0029721 PMID: 22253764

82. Long Y, Gu Y, Han H. Spatiotemporal heterogeneity of urban planning implementation effectiveness: Evidence from five urban master plans of Beijing. Landscape Urban Plan. 2012; 108: 103–111.