Lee et al. *EPJ Data Science* (2021) 10:41
https://doi.org/10.1140/epjds/s13688-021-00298-3

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Spatiotemporal distributions of population in Seoul: joint influence of ridership and accessibility of the subway system

Ji-Hye Lee[1], Segun Goh[2], Keumsook Lee[3] and M.Y. Choi[1*]

*Correspondence:
mychoi@snu.ac.kr
[1]Department of Physics and Astronomy and Center for Theoretical Physics, Seoul National University, 08826 Seoul, Korea
Full list of author information is available at the end of the article

Abstract

Moving along with daily life, urban residents and commuters create characteristic spatiotemporal patterns which vary extensively with the time of day. These patterns are formed via traffic flows: accordingly, understanding the impact of transportation system is essential for urban planners to evaluate expected urban activities. To explore them, we examine specifically population distributions in Seoul City by analyzing hourly population data based on mobile phone location records in combination with a couple of indicators of the Seoul Subway system. Through clustering and principal component analyses, we first demonstrate that the spatial distribution of the population is categorized according to the time of day, i.e., night, daytime, and evening, variations across which reflect the morphology of land use. We then examine the influence of the subway system on the population, employing ridership and accessibility as indicators. Our linear regression analysis shows that both are associated with the daytime and the evening populations, which implies that only commercial activities are substantially coupled to the subway system. Further, we find that the distinctive difference of night population is encoded in the probability distributions; this is elucidated by means of a multiplicative growth model for the morphological evolution of Seoul, revealing decentralization of residential areas and centralization of commercial areas. This study sheds light on the interplay of a public transportation system and land use, which is of relevance to planners and policymakers wishing to develop neighborhoods in support of sustainable modes.

Keywords: Spatiotemporal distribution of population; Mobile phone location data; Urban transportation network; Ridership; Accessibility; Urban growth; Urban morphology

1 Introduction

The spatiotemporal distribution of the urban population, which reflects the vivid rhythm of life and space utilization behaviors of urban residents, is very important for building plans of urban land use and transportation systems as well as for understanding the urban structure. In daily life, people move to places where related facilities are located according to the time of day, which leads the population distribution to change over time [1]. Specifically, urban residents create traffic flows according to various purposes, such as shopping,

public affairs and social activities, and leisure and cultural activities, as well as commuting between residences and workplaces. Meanwhile, the travel time may vary depending on the purpose of travel, and the locations of each passenger's residence and various facilities are different. Accordingly, the distribution of urban population can vary extensively over time and in particular differ greatly from the distribution of the number of residents, depending on the time of day. It is also expected that the resulting spatiotemporal distribution of population should be intimately related to that of traffic flows.

However, due to the absence of time and space information on people's movements, existing studies were conducted mostly to analyze the activity patterns at the individual level by utilizing rather limited survey data [2–6] while some spatiotemporal distributions of the passengers using subway stations were also examined [7, 8]. In recent years, there have been several attempts to develop the methodology to capture the real-time distribution of population better by utilizing big data from smart media such as mobile phones [9–15].

This study seeks to understand how the spatiotemporal distribution of the population is formed in Seoul, with emphasis on the influence of accessibility of the subway system as well as the difference between residential and commercial activities. We first examine the hourly population of each subway station area, estimated from the mobile phone signal data. Performing clustering analysis and principal component analysis (PCA) [16], we find that the spatial distribution of population varies depending on the time of day: night, daytime, and evening. In accord, the hourly population is categorized into three groups corresponding to the three time slots.

To understand the relationship between the population distribution and subway transportation, we consider two different measures of the Seoul subway station. One of the employed measures is ridership which characterize rather direct influences of the subway system on the population of the station area. The other quantity we take in this study is accessibility, which generally refers to an indicator measuring the ease of reaching other locations or activities distributed in space. As various theoretical approaches and their operational implications have been discussed, accessibility as a key indicator of transport planning is gaining consensus [17–20]. Our linear regression between the population and either of theses two measures as an explanatory variable shows that ridership or accessibility alone is not enough to account for the population. If both quantities are used as explanatory variables, our model predictions are significantly improved, confirming the impact of the subway system on the population. However, such an impact turns out relevant only to the daytime and evening populations, leaving the night population distribution largely unexplored.

To fill this gap, we further investigate the population distributions in detail and find the distinctive difference of night population. The population distribution functions are in general skewed and fit into Weibull distributions for night population and log-normal distributions for daytime as well as evening populations. The subsequent application of a multiplicative growth model [21, 22] to the morphological evolution of Seoul elucidates that decentralization of residential areas and centralization of business areas play a crucial role in the emergence of such skewed night and daytime/evening population distributions, respectively. Therefore, we suggest that such decentralization imprinted in the night population distribution could be responsible for the decoupling of the population and the subway network.

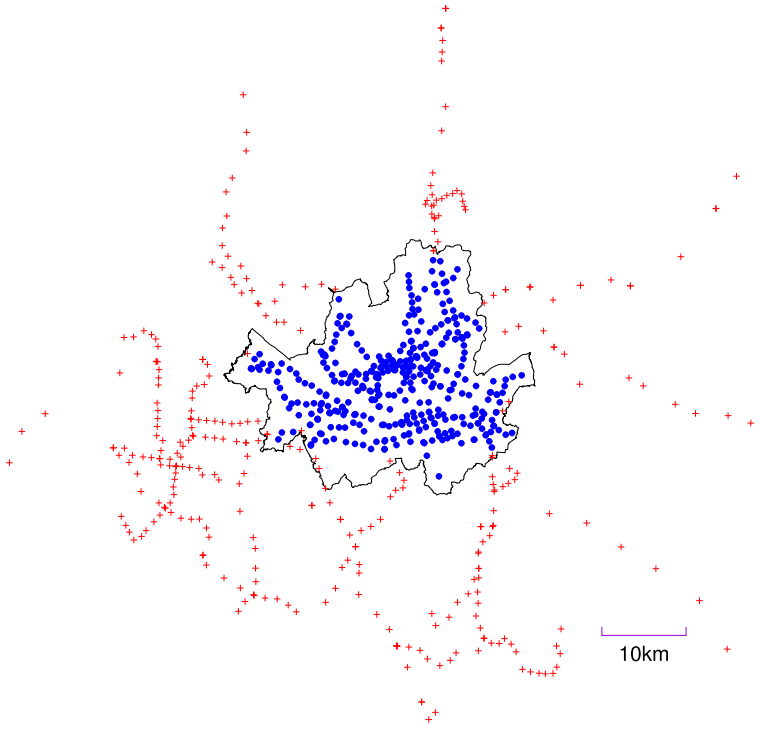

Figure 1 *Seoul Metropolitan Subway system*. Each station is represented by a (blue or red) dot whereas the (black) solid line depicts the boundary of Seoul City. Accessibility, as well as the number of passengers at each station, is obtained from the data for the whole subway system. In the case of population, 289 stations (blue dots) inside Seoul out of 615 stations are used since the population data are available only for Seoul

2 Data analysis

The study area is presented in Fig. 1, which displays the stations in the Seoul Metropolitan Subway system, along with the boundary of Seoul City. The subway system consists of 22 lines including 615 stations as of November 2019 and covers most of Seoul City and its suburbs, providing a major means of transportation in the Seoul metropolitan area as well as the municipal area. Its modal share occupies more than 40% of all passenger trips in 2018 [23].

The key data sources used in this study are the rail network geographic information system (GIS) database [24] and the ridership data [25] as well as the population data [26]. The rail network GIS database contains the inter-station distances as well as the locations of stations, provided by the *Ministry of Land, Infrastructure and Transport*. The ridership data, provided by the *Seoul Metropolitan Government*, contain the daily numbers of boarding and alighting passengers for 510 out of 615 stations in November 2019. The data were collected by the *Seoul Metropolitan Government* through the smart card called the *T-money* card. To pass through a turnstile of the subway system, every passenger should touch it to the card reader of the turnstile. A passenger who does not have one can purchase a one-time usable card from a vending machine. In this manner, accurate ridership data can be extracted from smart-card data since every riding and alighting at each station is recorded. The population data, provided by the *Seoul Metropolitan Government*, contain the hourly numbers of people on Thursday, 21 November, 2019 in each unit district called *Jipkyegu*, the minimum area unit established by the *National Statistical Office*

to provide statistical information. The population data of 19,153 unit districts were estimated from the long-term evolution (LTE) signal records, the communication histories between mobile phones and LTE base stations. The amount of data generated by a person is 8640 records on monthly average, 70 times larger than call or text-based data. There is little difference among individuals, and data are generated evenly at any time. Specific information on the population data is provided by *Seoul Metropolitan Government* [27].

2.1 Population

To investigate the influence of the subway transportation system on the population distribution, we consider the population of each station, defined to be the number of people in the unit districts within the distance of 500 m from the station. Since the population data are available only for Seoul, 289 out of 615 stations are selected for analysis, represented by blue dots in Fig. 1.

Examining hourly populations in a day consisting of 24 time windows, we observe high positive correlations between populations in some time windows, which allow categorizing the time slots. Such correlations are probed by means of clustering analysis, specifically, the hierarchical agglomerative clustering method [28], in which the 24 time windows start in their own clusters and are then merged in the greedy manner. The dissimilarity in the population distribution between two time windows t_1 and t_2 is measured by

$$D_{t_1 t_2} = 1 - \rho_{t_1 t_2} \tag{1}$$

with the Pearson correlation coefficient

$$\rho_{t_1 t_2} \equiv \frac{\sum_i [m_{t_1}(i) - \mu_{t_1}][m_{t_2}(i) - \mu_{t_2}]}{\sqrt{\sum_i [m_{t_1}(i) - \mu_{t_1}]^2}\sqrt{\sum_i [m_{t_2}(i) - \mu_{t_2}]^2}}, \tag{2}$$

where $m_t(i)$ is the population around station i $(= 1, 2, \ldots, N = 289)$ in time window t $(= 0, 1, \ldots, 23)$, and $\mu_t \equiv N^{-1} \sum_i^N m_t(i)$ is the mean of $m_t(i)$ over all stations. In determining the dissimilarity between clusters, a linkage criterion is additionally required. We use three linkage criteria: single, unweighted-average, and complete criteria. In the single-linkage criterion, the dissimilarity between clusters is defined to be the minimum of the dissimilarity between time windows belonging to each cluster. In the unweighted-average-linkage criterion, the dissimilarity between clusters is defined as the average of the dissimilarity between time windows belonging to each cluster. In the complete-linkage criterion, the dissimilarity between clusters are defined as the maximum of the dissimilarity between time windows belonging to each cluster.

The Pearson correlations among hourly populations are shown in Fig. 2, where each cell color represents the correlations between two populations. The hourly populations in the nighttime are highly correlated with each other; so are those in the daytime. To categorize time windows, we carry out clustering analysis. The dendrogram of hierarchical agglomerative clustering is shown in Fig. 3, illustrating how each cluster is composed via each linkage criterion. At the start of the dendrogram, each time window is labelled with t $(= 0, 1, \ldots, 23$ in 24-hour notation). From bottom to top, time windows with high correlations are preferentially merged, which are depicted by inverted U-shaped links whose heights represent the dissimilarity obtained via each criterion from Eq. (1). Following the dendrograms, one can find the clusters commonly observed in all three criteria. Those

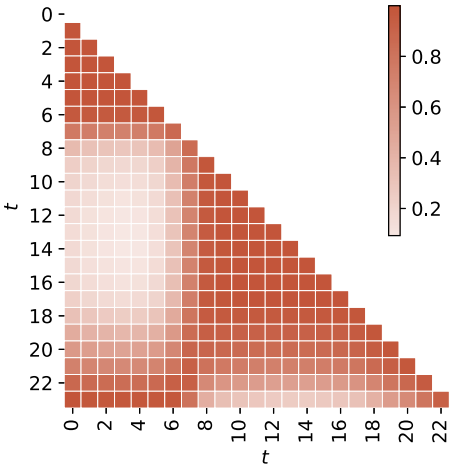

Figure 2 *Pearson correlation coefficient*. Each cell color represents the Pearson correlation coefficient between populations at different time windows t_1 and t_2, given by Eq. (2). Both time windows are given in 24-hour notation. The values that colors represent are illustrated in the color bar: darker colors indicate higher correlations between two populations. The hourly populations in night time are highly correlated with each other, while those in the daytime are also highly correlated with each other

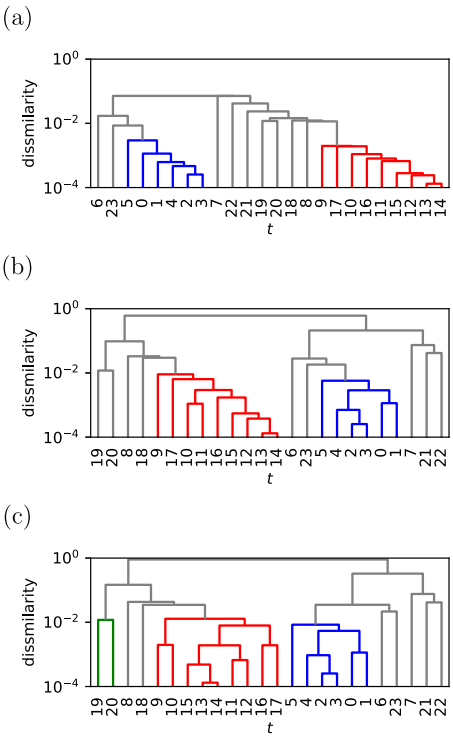

Figure 3 *Dendrograms of hierarchical agglomerative clustering*. The dendrograms illustrate how clusters are composed according to the (**a**) single, (**b**) average, and (**c**) complete linkage criteria. At the start of the dendrogram, each time window is labelled with t (= 0, 1, . . . , 23 in 24-hour notation). From bottom to top, time windows with high correlations are preferentially merged, as depicted by the inverted U-shaped link with the height representing the dissimilarity under each criterion with Eq. (1). Following the dendrograms, one can find the clusters commonly observed in all three criteria. Those with the least dissimilarity are the time windows from $t = 0$ to 5 (blue links) and from $t = 9$ to 17 (red links). At this level of dissimilarity in the complete criterion, a cluster including $t = 19$ and 20 (green links) is additionally observed

with the least dissimilarity are the time windows from $t = 0$ to 5 (blue links) and from $t = 9$ to 17 (red links). At this level of dissimilarity in the complete criterion, a cluster including $t = 19$ and 20 (green links) is additionally observed.

To clarify this ambiguity, we also carry out PCA, taking the eigenvectors with the three largest eigenvalues of the correlation matrix obtained from Eq. (2). These eigenvectors constitute the principal bases, where the projections have the largest variances. We then examine the components in each eigenvector and identify how hourly population distributions in the 24 time windows attribute to the variances. Note that the correlated spatial distributions of hourly populations in different time windows give rise to similar distributions of components. Specifically, we first find that the ratios of the three largest eigen-

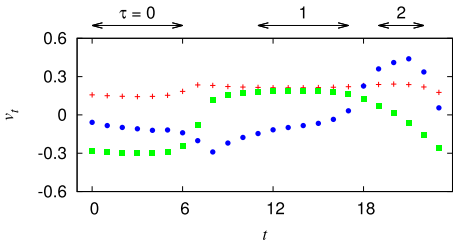

Figure 4 *Components v_t of the first three normalized principal bases*. Unlike the first principal basis showing more or less the same contributions of time windows (red crosses), the second principal basis differentiates night time from the daytime (green squares). Furthermore, removing the first and second variances, one observes high contributions of the evening time (blue circles). Here the time window t (= 0, 1, . . . , 23) is expressed in 24-hour notation; The double-headed arrows designate the night, day, and evening time slots labeled by $\tau = 0, 1, 2$, respectively

values to the sum of all eigenvalues are given by 0.68, 0.30, and 0.011, respectively. Then the components v_t of the corresponding normalized eigenvectors $\mathbf{v} = (v_0, v_1, \ldots, v_{23})$ are presented in Fig. 4. The first principal basis shows that the contributions of various time windows are more or less the same (red crosses). In contrast, the second principal basis manifests the conspicuous difference between the daytime and nighttime (green squares). Accordingly, it is reasonable to define the daytime population and the night population as the mean values in the time windows 11 to 17 and 0 to 6, respectively. Furthermore, when the first and the second variances are removed, the time windows from $t = 19$ to 22 make high contributions (blue circles), suggesting the evening time slot. In this fashion, we obtain three categories for the population distribution: night population ($0 \leq t < 6$), daytime population ($11 \leq t < 17$), and evening population ($19 \leq t < 22$). Henceforth the night, day, and evening time slots are labelled by $\tau = 0, 1$, and 2, respectively (double-headed arrows in Fig. 4).

With this categorization, we now analyze the spatial distributions of mean hourly populations in the three time slots, which exhibit distinct differences in Fig. 5. The color in each circle, located at a station, represents the population in the *Jipgyegu* districts within the distance of 500 m from the station, as depicted by the color bar with the range from the minimum to the maximum of each time-slot population. While night population exhibits a dispersed pattern, daytime and evening populations exhibit concentrated patterns in space. The spatiotemporally floating population on a day is determined by people's daily activities such as sleep at night and work in the daytime, which implies that the observed spatial patterns are intrinsically related to the land-use structure in Seoul. In this context, it is straightforward to understand the night population: Comparison with the population of registered residents confirms that areas populated in the night time overlap mostly with residential areas. Specifically, we have examined all 424 administration districts, each called *dong*, to observe that the night population is highly correlated with the residential population, with the Pearson correlation coefficient given by 0.85. On the other hand, areas populated in the daytime tend to be localized geographically within the central business districts (CBDs), particularly in the traditional CBD. Business services are integrated in these areas serving as employment hubs [29–31]. Meanwhile, in the evening, the population still concentrates on the CBDs as the daytime distribution, but exhibits a more dispersed pattern characterized by decreases in the traditional CBD and increases in another CBD as well as regional centers relevant to entertainment activities. These results demonstrate that the dispersed pattern in night population is related to the residential activities while the concentration of population during awakening periods is related to

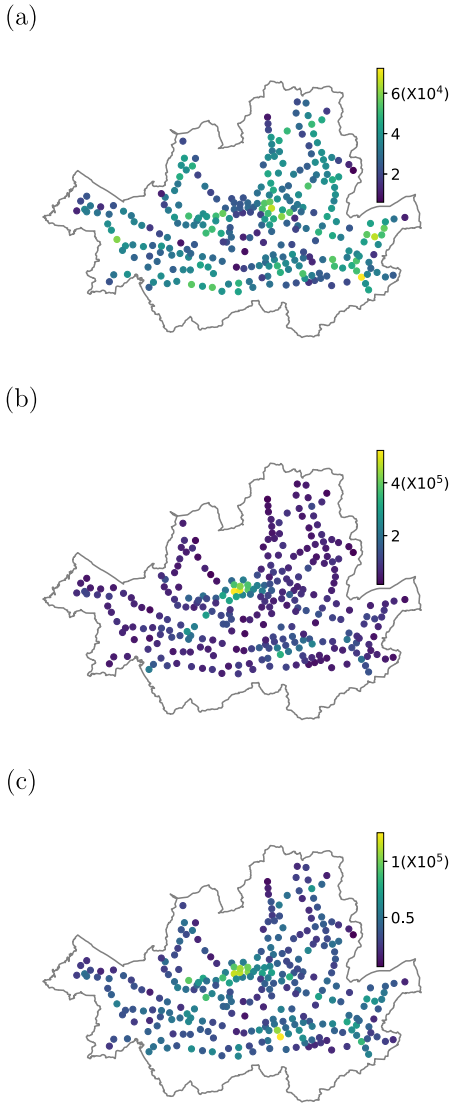

Figure 5 *Spatial distributions of* (**a**) *night,* (**b**) *daytime, and* (**c**) *evening populations*. Each circle color represents the population in the *Jipgyegu* districts within the distance of 500 m from the station, and its value is illustrated in the color bar ranging from the minimum to the maximum of each time-slot population. While the night population exhibits a dispersed pattern, the daytime and evening populations exhibit concentrated patterns. The spatial distribution of night population results largely from the residential activities, displaying the Pearson correlation coefficient 0.85 with the population of registered residents. In contrast, spatial distributions of the daytime and evening populations are largely due to the commercial activities, as manifested by the concentrated areas in the centers of Seoul. The difference between daytime and evening reflects the difference in the purpose of activities, work and entertainment

the commercial activities. The seemingly different purposes of the commercial activities during the daytime and evening induce similar but different patterns.

2.2 Ridership and accessibility

In the ridership data, collected through *T-money* cards and providing the daily numbers of passengers [25], the mean daily numbers of boarding and alighting passengers are essentially the same except for a slight discrepancy arising from a few passengers' failures to tag smart cards when alighting. We therefore use the number of boarding passengers in the analysis. Plotting the data, we obtain the spatial distribution of the numbers of passengers shown in Fig. 6.

We next consider accessibility. To extract information distinctive of the ridership, we define accessibility exclusively based on the structure of the Seoul subway network. Public transportation networks are developed in accord with the morphological evolution of land use, and therefore, their structure can be used as a unique indicator of the morpho-

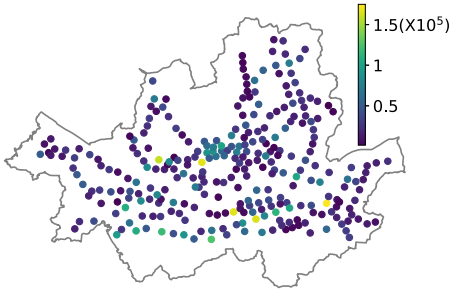

Figure 6 *Spatial distribution of ridership*. The color of each circle represents the ridership, defined to be the mean daily number of boarding passengers, of the corresponding station, and is explained in the color bar ranging from the minimum to the maximum of the ridership

logical evolution of land use. We also note that other accessibility indices such as those concerned with specific resources (e.g. jobs or facilities) [32] are not determined solely by the network structure and do not provide independent measures for the structure of the network. Ultimately, we aim to examine how relevant the ease of reaching locations is to the spatiotemporally varying population distribution.

We thus define the accessibility of station i specifically as

$$A_i \equiv \frac{1}{A_{\max}} \sum_j \left[h(d_{ij}) + h(d_{ji}) \right], \tag{3}$$

where d_{ij} is the shortest path distance from station j to station i and $A_{\max}$, the maximum of the summation on the right-hand side, takes care of the normalization. To compute d_{ij}, we use the rail network GIS which contains the information of distance between stations directly linked by a railway. The distance along this railway is used to define the element of the adjacency matrix. If two stations are not directly linked, the element is defined to be zero. The shortest path length on the graph with this adjacency matrix gives d_{ij}. Note that the distance corresponds approximately to the time on the subway, given that the speeds on railways are more or less similar and that transfers are not considered. Specifically, the speed ranges from 29.3 km/h to 35.9 km/h on the eight major lines [33]. If d_{ij} and d_{ji} are equal for all pairs, the second term in the summation on the right-hand side of Eq. (3) would be redundant. In the Seoul Metropolitan Subway system, however, the two can be different due to the six stations of *Line 6* connected in one direction. The function $h(d_{ij})$ describes how easy it is to get from j to i in the subway system. It should thus be a decreasing function of the distance d_{ij}, with two simple instances given by exponential and algebraic functions:

$$h(d_{ij}) = e^{-d_{ij}/d_0} \quad \text{and} \quad h(d_{ij}) = d_{ij}^{-\alpha}. \tag{4}$$

We first examine the difference between the exponential accessibility and algebraic accessibility, corresponding to the two functions in Eq. (4), respectively. Varying the characteristic distance d_0 and the decay exponent α, we compute the Pearson correlation coefficient between the two accessibility measures with parameters d_0 and α, which is shown in Fig. 7. It is observed that the two types of accessibility are mapped linearly quite well to each other in the typical range of the parameters. Note that this multicollinearity indicates that one accessibility is substantial for examining the relation to population. Since the variance inflation factor (VIF) [34] exceeding ten is commonly considered to have a

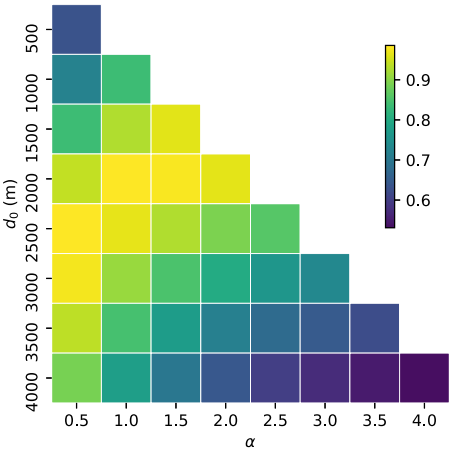

Figure 7 *Pearson correlations between accessibility measures*. Each cell color represents the Pearson correlation coefficient between exponential accessibility and algebraic accessibility given by Eq. (3) together with Eq. (4), with parameters d_0 and α, respectively. The values that colors represent are given in the color bar: brighter colors indicate higher correlations between the two accessibilities

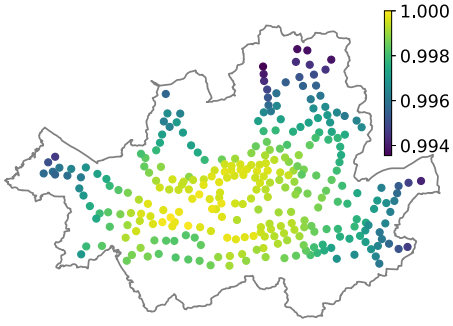

Figure 8 *Spatial distribution of exponential accessibility*. The whole subway system is used when computing the accessibility measure. The characteristic distance, obtained via fitting to the daytime population, takes the value $d_0 = 1796$ m

multicollinearity problem [35], the coefficient over $\sqrt{0.9} \approx 0.95$ can be set to define the typical range.

Therefore in the remaining part, we consider mainly the exponential accessibility unless stated otherwise. The spatial distribution of the values of the (exponential) accessibility is presented in Fig. 8, where the characteristic distance has been chosen as $d_0 = 1796$ m by fitting to the daytime population (see Sect. 3 for details). Comparison of the accessibility distribution with the population distributions in Fig. 5 may provide a useful guide to improving the public transportation system.

3 Influence of ridership and accessibility

In this section, we examine the relation of the population to the ridership and accessibility of the subway system. As a simple attempt, we first perform linear regression analysis of population $m_\tau(i)$ versus the (daily) passenger number $S(i)$ of station i:

$$m_\tau(i) = kS(i) + c, \tag{5}$$

where k and c are regression coefficients. Note that m_τ refers to actual population. Subtracting the mean value and dividing by the standard deviation reduce Eq. (5) to

$$\tilde{m}_\tau(i) = \kappa_S \tilde{S}(i) \tag{6}$$

Table 1 Fitting results in the ridership model and the accessibility model, given by Eqs. (6) and (7), respectively. κ_S and R^2_{adj} values of the ridership model are presented in the first and second columns, respectively. The third and fourth columns present the obtained values of κ_A and d_0 of exponential accessibility, respectively, while the corresponding values of R^2_{adj} are shown in the fifth column. Lastly, the sixth, seventh, and eighth columns are devoted to algebraic accessibility

Fitting results of the ridership model and the accessibility model								
	κ_S	R^2_{adj}	d_0	κ_A	R^2_{adj}	α	κ_A	R^2_{adj}
night	0.21 ± 0.06	0.04	551 ± 215	0.31 ± 0.06	0.1	2.50 ± 0.59	0.27 ± 0.06	0.08
daytime	0.54 ± 0.05	0.29	1794 ± 294	0.56 ± 0.05	0.32	1.40 ± 0.17	0.56 ± 0.05	0.31
evening	0.62 ± 0.05	0.38	1379 ± 251	0.52 ± 0.05	0.27	1.61 ± 0.18	0.52 ± 0.05	0.27

with standardized variables $\tilde{m}_\tau(i)$ and $\tilde{S}(i)$, where κ_S is estimated via the ordinary least square (OLS) method [36]. This defines the "ridership model". We also preform the same analysis with the accessibility. With the standardized accessibility $\tilde{A}_i$ of station i, we consider the "accessibility model" in the form

$$\tilde{m}_\tau(i) = \kappa_A \tilde{A}(i), \tag{7}$$

where we again use the OLS method to estimate κ_A and the accessibility parameter d_0 or α. (Recall that the two types of accessibility, exponential and algebraic, lead to similar results for the parameters of accessibility chosen suitably.) The results of the ridership and the accessibility models are summarized in Table 1. In all cases, p-values are lower than 0.001, indicating strong evidence against the null hypothesis that the coefficient vanishes. To examine how well the ridership and accessibility explain the population distribution, we look into the adjusted coefficient of determination R^2_{adj}. These models with a single explanation parameter show that the night population varies essentially regardless of the ridership ($R^2_{\text{adj}} = 0.04$) as well as the accessibility ($R^2_{\text{adj}} = 0.1$ for the exponential accessibility and $R^2_{\text{adj}} = 0.08$ for the power-law accessibility). On the other hand, variations in the daytime population, as well as the evening population, can be attributed partly to the ridership and accessibility. For the daytime and the evening populations, scatter plots of actual data versus the predictions via the ridership and accessibility models are presented in Fig. 9.

Next, as a better attempt, we consider multiple linear regression in the form:

$$\tilde{m}_\tau(i) = \kappa_1 \tilde{A}(i) + \kappa_2 \tilde{S}(i) \tag{8}$$

with two regression coefficients κ_1 and κ_2. In this "accessibility-ridership model", we use the OLS method to estimate three parameters κ_1, κ_2, and the accessibility parameter d_0 or α. The results are summarized in Tables 2 and 3. Table 2 presents the obtained values of d_0 (for the exponential accessibility) and α (for the algebraic accessibility), together with the corresponding values of the adjusted coefficient of determination R^2_{adj}. The first and the second columns display the characteristic distance d_0 and R^2_{adj} of the exponential accessibility, respectively; the third and the fourth columns display the decay exponent α and R^2_{adj} of the algebraic accessibility. The small values of R^2_{adj} in the first row again indicate that variations in the night population are attributable neither to ridership nor to accessibility. In contrast, the daytime population and the evening population are affected significantly by both ridership and accessibility. Specifically, in the case of daytime population, the parameter of accessibility is estimated as $d_0 = 1796 \pm 323$ m for exponential

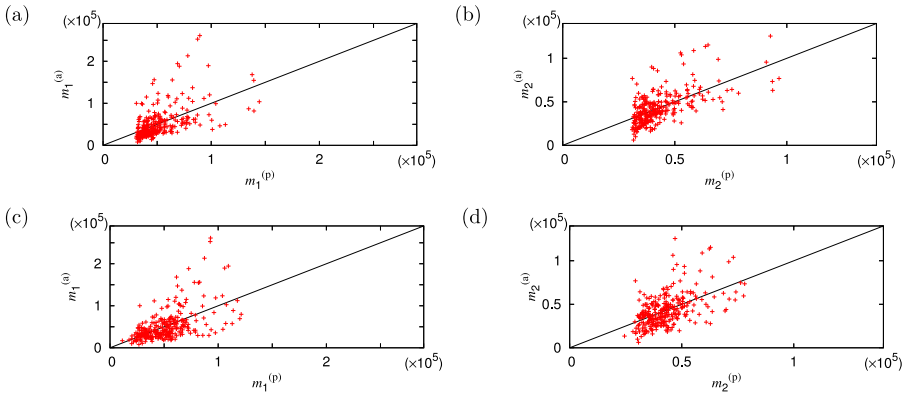

Figure 9 *Linear regression analysis of the ridership and accessibility models*. Actual population $m_\tau^{(a)}$ versus predicted population $m_\tau^{(p)}$ from the ridership model is plotted for (**a**) the daytime ($\tau = 1$) and (**b**) the evening ($\tau = 2$). In (**c**) and (**d**), actual population versus predicted population from the exponential accessibility is presented for the daytime and the evening, respectively. The coefficient of determination takes the values $R^2 = 0.29, 0.38, 0.32,$ and 0.27, respectively

Table 2 Fitting results in the accessibility-ridership model given by Eq. (8). The first and second columns present the parameter of exponential accessibility and corresponding values of R^2_{adj} while the third and fourth columns are devoted to algebraic accessibility

Fitting results of the accessibility-ridership model				
	d_0	R^2_{adj}	α	R^2_{adj}
night	473 ± 232	0.1	2.75 ± 0.87	0.09
daytime	1796 ± 323	0.48	1.54 ± 0.18	0.48
evening	1189 ± 248	0.51	1.88 ± 0.21	0.51

Table 3 Regression coefficients κ_1 and κ_2 in the accessibility-ridership model given by Eq. (8) for exponential accessibility and algebraic accessibility

Regression coefficients of the accessibility-ridership model		
Exponential accessibility	κ_1	κ_2
daytime	0.44 ± 0.045	0.42 ± 0.045
evening	0.38 ± 0.043	0.53 ± 0.043
Algebraic accessibility	κ_1	κ_2
daytime	0.45 ± 0.045	0.42 ± 0.045
evening	0.38 ± 0.043	0.52 ± 0.043

accessibility and $\alpha = 1.54 \pm 0.18$ for algebraic accessibility, with $R^2_{\text{adj}} = 0.48$. Note that R^2_{adj} takes similar values since the parameters d_0 and α are the ones with which the two types of accessibility are almost identical. In the case of evening population, the parameter of accessibility is estimated as $d_0 = 1189 \pm 248$ m and $\alpha = 1.88 \pm 0.21$, with $R^2_{\text{adj}} = 0.51$ again similar to each other. To disclose the multicollinearity between accessibility and ridership, we compute VIF [34] with the estimated values of d_0 and α. In all cases, VIF turns out to be approximately unity, indicating that accessibility and ridership are not in a linear relationship. Figure 10 presents the scatter plots of actual population versus predicted population (with exponential accessibility) in the daytime and the evening.

Employing both ridership and accessibility as explanatory variables, we improve the model and obtain better predictions of the population, as manifested by R^2_{adj} values. To de-

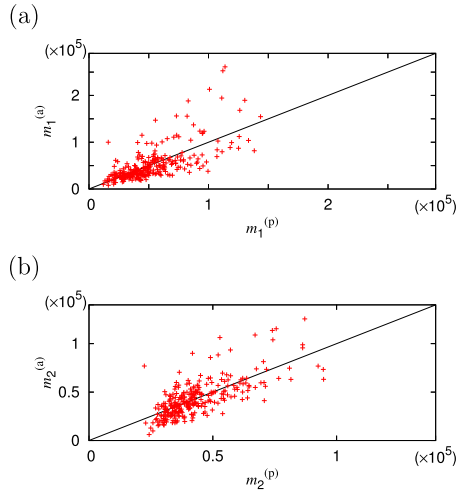

Figure 10 *Linear regression analysis of the accessibility-ridership model*. Actual population $m_\tau^{(a)}$ versus predicted population $m_\tau^{(p)}$ from the accessibility-ridership model is plotted for (**a**) the daytime ($\tau = 1$) and (**b**) the evening ($\tau = 2$). The coefficient of determination takes the values $R^2 = 0.48$ and 0.51, respectively

cide whether the improvement is just due to adding one more parameter, we compute the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) scores [36] and find that both scores drop substantially, by more than 60 for the accessibility-ridership model, e.g., compared with the ridership model; this confirms that it is indeed a better model. Here the contributions of the two variables, accessibility and ridership, to the population can be assessed by comparing the values of κ_1 and κ_2, summarized in Table 3. In the case of daytime population, both contributions are similar while ridership contributes more to the population in the evening. We also note that night population is observed to be rather decoupled with the subway network.

4 Growth of distribution functions

Examining the population distributions in the three time-slots, we observe that they are generally described by unimodal and skewed functions. Figure 11 presents the distributions of population m_τ at night ($\tau = 0$), in the daytime ($\tau = 1$), and in the evening ($\tau = 2$), with the data points plotted with black squares on the logarithmic scale. The probability density function (PDF) $f(m_\tau)$ has been obtained via data binning: The number of values in a given bin is counted and divided by the total number as well as the bin size.

It is shown in Figs. 11(b) and (c) that the distributions in the daytime and in the evening exhibit apparently symmetrical shapes, which turn out to fit into log-normal distributions

$$f_L(m) = \frac{1}{m\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln m - \mu)^2}{2\sigma^2}\right], \tag{9}$$

with two parameters μ and σ obtained via maximum likelihood estimation (MLE) (see the red lines). On the other hand, as shown in Fig. 11(a), the distribution at night is still skewed on the logarithmic scale and fit for the Weibull distribution

$$f_W(m) = \frac{\gamma}{\eta}\left(\frac{m}{\eta}\right)^{\gamma-1} \exp\left[\left(\frac{m}{\eta}\right)^{\gamma}\right] \tag{10}$$

with two parameters γ and η obtained again via MLE (see the blue line). For comparison, both log-normal and Weibull distributions are plotted with red and blue lines. The cor-

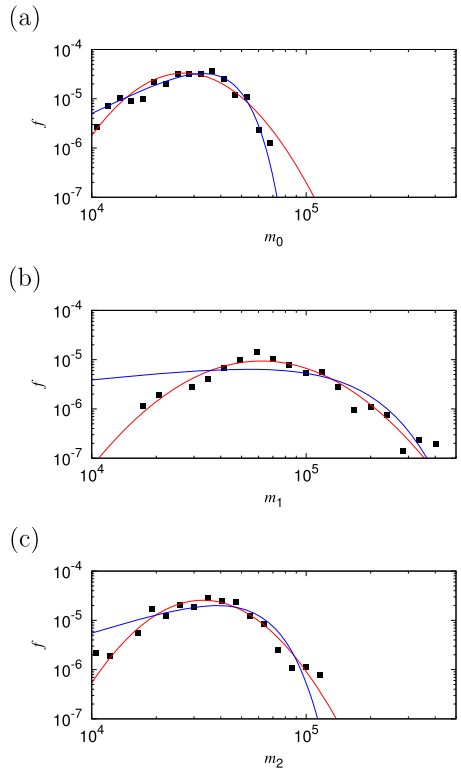

Figure 11 *Population distribution function for each time slot*. The distribution function f is plotted as a function of (**a**) night population m_0, (**b**) daytime population m_1, and (**c**) evening population m_2. Data points, plotted with squares, are fitted to the log-normal and Weibull functions depicted by red and blue lines, respectively, with parameters obtained via maximum likelihood estimation. In (**a**), the Weibull distribution gives a better description of the night population whereas (**b**) and (**c**) show that the log-normal distributions perform better for daytime and evening populations; this is confirmed by $|\Delta\mathcal{A}| > 20$

Table 4 Parameters of Eqs. (9) and (10) obtained via maximum likelihood estimation from night population ($\tau = 0$), daytime population ($\tau = 1$), and evening population ($\tau = 2$)

Distribution function parameters				
τ	μ	σ	γ	η
0	10.37 ± 0.03	0.41 ± 0.02	3.1 ± 0.1	$37{,}362 \pm 755$
1	11.37 ± 0.03	0.58 ± 0.03	1.5 ± 0.1	$116{,}499 \pm 5524$
2	10.59 ± 0.03	0.43 ± 0.02	2.4 ± 0.1	$48{,}657 \pm 1317$

responding four parameters, μ, σ, γ, and η for $\tau = 0, 1, 2$ are summarized in Table 4. We further compute the AIC score $\mathcal{A}$ and find that the absolute values of the difference are larger than 20 for all three populations: $|\Delta\mathcal{A}| > 20$. It turns out that Weibull distributions exhibit smaller scores in the case of the night population while log-normal distributions exhibit smaller scores for the daytime and evening populations. These analyses reveal that the patterns observed in the population distribution involve not only the dispersion across the space but also the asymmetrical change in the shape of the distribution, which we now illuminate with a modeling approach.

Log-normal and Weibull distributions emerge naturally from the multiplicative processes with growth and production, which can be described conveniently by master equations [21, 22, 37–39]. This approach provides lucid elucidation of how the observed distribution emerges as the population grows according to the land-use planning. Cities generally evolve from monocentric structures, where all activities are concentrated in rather restricted geographical areas, to more sprawled and polycentric structures, where commercial activities are constantly centralized while residential activities decentralized [40]. Likewise, in the early days of Seoul, both business-related and residential populations con-

centrated in the center [31]. Such populated areas attracted most activities, resulting in the city highly centralized. In accord, it is plausible that the population growth at this early stage follows the multiplicative process [41, 42], where the amount of an increment in size is proportional to the current size, i.e., $\Delta m = bm$. More precisely, our multiplicative growth is expressed in terms of the transition rate $\omega(m \to m') = \lambda\delta(m' - m - bm)$ with the growth factor b and rate λ. With the transition rate specified, one can establish the master equation governing the process and derive a time evolution equation for the PDF $f(m,t)$ of population m at time t. Specifically, since the growth of an area from size m to $(1+b)m$ can be interpreted as a combination of annihilation of the area of size m and creation of size $(1+b)m$, the time evolution equation for $f(m,t)$ consists of two terms as follows [21, 22]:

$$\frac{\partial}{\partial t}f(m,t) = -\lambda f(m,t) + \frac{\lambda}{1+b}f\left(\frac{m}{1+b}\right), \tag{11}$$

where the first and the second terms on the right-hand side describe the decrease and the increase of $f(m,t)$, respectively, as the population grows from size m to $(1+b)m$ and from $m/(1+b)$ to m. Note that the factor $1+b$ in the denominator of the prefactor in the second term is required for normalization of the probability density. The solution is given by the log-normal distribution in Eq. (9), where the parameters evolve in time according to [22, 38]:

$$\begin{aligned}\mu &= \lambda t \ln(1+b) + \mu_0,\\ \sigma^2 &= \lambda t\left[\ln(1+b)\right]^2 + \sigma_0^2\end{aligned} \tag{12}$$

with the initial values μ_0 and σ_0.

Heretofore, populations have been considered to grow without decentralization. In reality, as a city develops, residential areas are constructed mostly in the peripheral areas due to raised land values. Further, the population growth induces new centers in the city, which thus goes through centralization as well as decentralization. This morphologic evolution describes well the observed spatial distributions of population in Sect. 2.1. In this course, public transportation systems such as the subway system are constructed to enable access to the segregated residential and commercial areas. Namely, new subway lines and/or stations are added to connect the centers or enable the commute. Accordingly, it is expected that the business-related population around newly constructed stations is similar to that around existing stations, for which the master equation leads to Eq. (11) again, as the PDF of populations near newly added stations would have the same form as that of existing stations. Still, population is considered to grow proportionally, which naturally leads to the formation of urban centers and log-normal distributions. This explains the concentrated daytime and evening populations and their probability distributions.

Here the growth characteristics of daytime and evening populations can be further assessed by comparing the growth factors and rates. Fitting the data to Eq. (12), we obtain $b = 0.03$ and $\lambda = 384$ for daytime population ($\tau = 1$) together with $b = 0.018$ and $\lambda = 606$ for evening population ($\tau = 2$). Compared with the daytime population, the evening population exhibits a smaller value of b and a larger value of λ, which suggests a smaller scale and a larger rate of land-use change: the areas populated in the evening have been developed more frequently and gradually than the areas populated in the daytime.

Now, to take into account the decentralization in residential activities, we introduce additional terms, which correspond to production of newly accessible areas in smaller size, characterized by the reduction factor β as well as rate r. The master equation then leads to the time evolution equation for the PDF:

$$\frac{\partial}{\partial t}f(m,t) = -(\lambda + r)f(m,t) + \frac{\lambda}{1+b}f\left(\frac{m}{1+b}\right) + \frac{r}{1-\beta}f\left(\frac{m}{1-\beta}\right), \tag{13}$$

where the first term on the right-hand side describes the annihilation of existing area of size m with rate $\lambda + r$ due to growth as well as reduction while the second and third terms describe the creation from area of size $m/(1+b)$ by growth and of size $m/(1-\beta)$ by reduction, respectively. As already discussed for newly added accessible areas via the centralization mechanism, the term involving the factor β captures not only reduction of populations but also production of new areas with the reduced size $(1-\beta)m$, and ultimately, redistribution, i.e., decentralization of populations [8, 38], consisting of annihilation of an area of size m and creation of several areas of the reduced size $(1-\beta)m$.

Equation (13) in the rare-reduction limit bears the solution given by the Weibull distribution in Eq. (10) [38], where the parameters evolve in time according to

$$\gamma = \frac{1}{\sqrt{\gamma_0^{-2} + C\lambda t}}, \qquad \eta = \eta_0 e^{Dt} \tag{14}$$

with $C \equiv \lambda[\ln(1+b)]^2 + r[\ln(1-\beta)]^2$ and $D \equiv \lambda\ln(1+b) + r\ln(1-\beta)$ as well as the initial values γ_0 and η_0. Therefore, our observation that the night population fits better to the Weibull distribution demonstrates another distinctive feature inherent in the night population distribution: in addition to the spatially dispersed pattern, emergence of the Weibull distribution manifests that there involves decentralization in the development of residential areas in Seoul. In particular, we suggest that the decentralization of residential activities during suburbanization could have derived the dispersed spatial pattern observed in the night population distribution and resulted simultaneously in the decoupling with the subway system. Whereas the spatial dispersion is confirmed visually, the decentralization is elucidated via our multiplicative growth model.

Our growth model can also be applied to the ridership distributions. Specifically, we have examined the PDF of the numbers of passengers in the Seoul municipal area as well as the metropolitan area. The PDF for the municipal area turns out to follow a log-normal distribution with $\mu = 9.58 \pm 0.06$ and $\sigma = 0.96 \pm 0.05$ whereas the PDF for the metropolitan area follows a Weibull distribution with $\gamma = 1.03 \pm 0.04$ and $\sigma = 15{,}475 \pm 716$. The details of the analysis are largely similar to those in the previous study [37] and not presented here.

5 Discussion

Analyzing the hourly population data based on LTE signal records in Seoul via clustering analysis and PCA, we have demonstrated that the spatial distribution of population varies with the time of day: night, daytime, and evening. The spatial distribution visualizes the populated areas during each time slot, manifesting dispersion in residential activities of night population and concentration in commercial activities of daytime and evening populations. It is likely that the purposes of commercial activities give rise to the difference

between the daytime and the evening populations, which may be supported by the land-use pattern in Seoul [29–31]. Since daytime and evening activities are expected to focus on work and entertainment, data about the points of interests and the numbers of jobs as well as businesses would be useful for probing this question [43, 44]. Meanwhile, night population has been explained well by the residential population data.

The separation of residential activities from commercial activities in the urban core is the result of suburbanization, during which the residential population moves from the urban core to the suburbs. This process often occurs in economically developed cities [45] and is intimately associated with the introduction of a new mode of transportation [46], which is the subway system in the case of Seoul [31]. The subway system is a major transportation infrastructure in Seoul, with the modal share over 40% in 2019 [23]. To assess the impact of the transportation system on the population distribution, we have investigated the correlations between time-slot population and daily ridership of each station as well as local accessibility on the subway network via linear regression. Ridership as well as accessibility have turned out to correlate positively with the populations in daytime and in the evening. Namely, it has been shown that commercial activities concentrate on the areas which are easy to reach and have high demands for movement. The regression coefficients have further disclosed that accessibility is more important to daytime population than to night population, which implies that accessibility can serve as an indicator for the urban center. Note that ridership and accessibility measurement used in this study is one of basic indicators in a transportation system. Other indicators would give different results and interpretations. For instance, ridership data during the peak time in the morning are expected to indicate the demand for commuting [37]. Also, accessibility measurement of how accessible neighboring stations indicates the ease of reaching through multiple trips and can thus make a useful indicator for the connectivity of the subway network [47–50].

Finally, we have found that the distinctive difference of night population is also encoded in the probability distribution, fit into the Weibull distribution. Since people tend to visit populated areas, both populations related to residential and commercial activities could be considered to grow proportionally to their sizes. The multiplicative growth process described by the master equation has elucidated that the probability density function of such populations would follow log-normal distributions. In addition, due to the suburbanization, the population appearing in peripheral areas has been considered in the process, which leads to the emergence of the Weibull distribution for night population. In this interpretation, the probability distribution of ridership transforms from the log-normal distribution to the Weibull distribution as newly constructed stations are added in peripheral areas. Indeed, the probability distributions of ridership in municipal and metropolitan areas tend to follow log-normal distributions and Weibull distributions, respectively. Note that this transition can be implemented in defining the boundary of urban areas, which is an important issue in studying universal features in cities such as scaling laws [51, 52].

Acknowledgements
We are grateful to J. Jo and J.W. Kim for helpful discussions.

Funding
This work was supported by National Research Foundation of Korea through the Basic Science Research Program (Grant No. 2019R1F1A1046285).

Abbreviations
PCA, Principal component analysis; OLS, Ordinary least square; GIS, Geographic information system; LTE, Long-term evolution; CBD, Central business district; PDF, Probability density function; MLE, Maximum likelihood estimation; VIF, Variance inflation factor; AIC, Akaike information criterion; BIC, Bayesian information criterion.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the Korea Transport Database repository, https://www.ktdb.go.kr/www/selectPbldataChargerWebList.do?key=12 (rail network GIS dataset) and the Seoul Metropolitan Government repository, https://data.seoul.go.kr/dataList/OA-12914/S/1/datasetView.do (ridership dataset), and https://data.seoul.go.kr/dataList/OA-14979/F/1/datasetView.do (population dataset).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JHL and KL collected the data. JHL and SG analyzed the data. JHL, SG and MYC devised the model. KL and MYC conceived and supervised the study. All contributed to writing the manuscript and approved the final manuscript.

Author details
[1]Department of Physics and Astronomy and Center for Theoretical Physics, Seoul National University, 08826 Seoul, Korea. [2]Theoretical Physics of Living Matter, Institute of Biological Information Processing, Forschungszentrum Jülich, 52425 Jülich, Germany. [3]Department of Geography, Sungshin Women's University, 02844 Seoul, Korea.

Publisher's Note

Received: 23 March 2021 Accepted: 1 August 2021 Published online: 10 August 2021

References
1. Hägerstrand T (1970) What about people in regional science? Pap Reg Sci 24:7–24
2. Janelle D, Goodchild MF (1983) Diurnal patterns of social group distributions in a Canadian city. Econ Geogr 59:403–425
3. Goodchild MF, Klinkenberg B, Janelle DG (1993) A factorial model of aggregate spatiotemporal behavior: application to the diurnal cycle. Geogr Anal 25:277–294
4. Weber J, Kwan MP (2002) Bringing time back in: a study on the influence of travel timevariations and facility opening hours on individual accessibility. Prof Geogr 54:226–240
5. Zandvliet R, Dijst M (2006) Short-term dynamics in the use of places: a space-time typology of visitor populations in the Netherlands. Urban Stud 43:1159–1176
6. Yoon SY, Ravulaparthy SK, Goulias KG (2014) Dynamic diurnal social taxonomy of urban environments using data from a geocoded time use activity-travel diary and point-based businessestablishment inventory. Transp Res, Part A, Policy Pract 68:3–17
7. Chen C, Chen J, Barry J (2009) Diurnal pattern of transit ridership: a case study of the New York City subway system. J Transp Geogr 17:176–186
8. Lee K, Park JS, Choi H, Choi MY, Jung WS (2010) Sleepless in Seoul: 'The ant and the metrohopper'. J Korean Phys Soc 57:823–825
9. Kang C, Liu Y, Ma X, Wu L (2012) Towards estimating urban population distributions from mobile call data. J Urban Technol 19(4):3–21
10. Csáji BC et al (2013) Exploring the mobility of mobile phone users. Physica A 392(6):1459–1473
11. Deville P et al (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888–15893
12. Lee K, Kim HS, Lee SY (2015) The spatial characteristics of real-time population distribution in Seoul based on the media users' time-space information for the activity spaces. J Econ Geogr Soc Korea 18:87–102
13. Chen BY, Wang Y, Wang D, Li Q, Lam WH, Shaw SL (2009) Understanding the impacts of human mobility on accessibility using massive mobile phone tracking data. Ann Assoc Am Geogr 108:1115–1133
14. Lee K, Kim HS (2018) Spatio-temporal analysis of population distribution in Seoul via integrating transportation and land use information; based on four-dimensional visualization methods. J Econ Geogr Soc Korea 21:20–33
15. Liu Z, Ma T, Du Y, Pei T, Yi J, Peng H (2018) Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. Trans GIS 22(2):494–513
16. Joliffe IT, Morgan BJT (1992) Principal component analysis and exploratory factor analysis. Stat Methods Med Res 1(1):69–95
17. Mladenović MN et al (2017) Transport justice: designing fair transportation systems. Transp Rev 37(2):245–246
18. Miller EJ (2018) Accessibility: measurement and application in transportation planning. Transp Rev 38(5):551–555
19. Van Wee B, Geurs K (2011) Discussing equity and social exclusion in accessibility evaluations. Eur J Transp Infrastruct Res 11(4):350–367
20. Vecchio G, Tiznado-Aitken I, Hurtubia R (2020) Transport and equity in Latin America: a critical review of socially oriented accessibility assessments. Transp Rev 40(3):354–381
21. Choi MY, Choi H, Fortin JY, Choi J (2009) How skew distributions emerge in evolving systems. Europhys Lett 85(3):30006
22. Goh S, Kwon HW, Choi MY, Fortin JY (2010) Emergence of skew distributions in controlled growth processes. Phys Rev E 82:061115
23. Modal share 2019. https://news.seoul.go.kr/traffic/archives/289. Accessed 17 Sep 2020

24. GIS database for rail network. https://www.ktdb.go.kr/www/selectPbldataChargerWebList.do?key=12. Accessed 17 Sep 2020
25. Ridership. https://data.seoul.go.kr/dataList/OA-12914/S/1/datasetView.do. Accessed 17 Sep 2020
26. Population. https://data.seoul.go.kr/dataList/OA-14979/F/1/datasetView.do. Accessed 17 Sep 2020
27. Manual for population data. https://data.seoul.go.kr/dataVisual/seoul/seoulLivingPopulation.do. Accessed 17 Sep 2020
28. Müllner D (2011) Modern hierarchical, agglomerative clustering algorithms. arXiv:1109.2378
29. Land Use 2010. https://data.si.re.kr/node/108. Accessed 22 Jun 2021
30. Land Use 2015. https://data.si.re.kr/node/45. Accessed 22 Jun 2021
31. Yi C, Kim I, Song M, Jy Y (2019) Plans and actions for sustainable urban spatial structure of Seoul. The Seoul Institute, Seoul
32. Páez A, Scott DM, Morency C (2012) Measuring accessibility: positive and normative implementations of various accessibility indicators. J Transp Geogr 25:141–153
33. Speed. https://www.kric.go.kr/jsp/handbook/sta/statisticsTechList.jsp. Accessed 22 Jun 2021
34. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Elsevier, New York
35. O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41(5):673–690
36. Choi BS (1992) ARMA model identification. Springer, New York
37. Lee K, Goh S, Park JS, Jung WS, Choi MY (2011) Master equation approach to the intra-urban passenger flow and application to the Metropolitan Seoul Subway system. J Phys A 44(11):115007
38. Goh S, Kwon HW, Choi MY (2014) Discriminating between Weibull distributions and log-normal distributions emerging in branching processes. J Phys A 47(22):225101
39. Goh S, Choi J, Choi MY, Yoon BG (2017) Time evolution of entropy in a growth model: dependence on the description. J Korean Phys Soc 70(1):12–21
40. Mayer HM, Kohn CF (eds) (1959) Readings in urban geography University of Chicago Press, Chicago
41. Gibrat RL (1931) Inégalités Économiques. Librairie du Recueil Sirey, Paris
42. Arshad S, Hu S, Ashraf BN (2018) Zipf's law and city size distribution: a survey of the literature and future research agenda. Physica A 492:75–92
43. Baniukiewicz M, Dick ZL, Giabbanelli PJ (2018) Capturing the fast-food landscape in England using large-scale network analysis. EPJ Data Sci 7(1):39
44. Chong SK, Bahrami M, Chen H, Balcisoy S, Bozkaya B et al (2020) Economic outcomes predicted by diversity in cities. EPJ Data Sci 9(1):17
45. Carra G, Barthelemy M (2019) A fundamental diagram of urbanization. Environ Plan B, Plan Des 46(4):690–706
46. Antrop M (2004) Landscape change and the urbanization process in Europe. Landsc Urban Plan 67(1–4):9–26
47. Landau E (1895) Zur relativen Wertbemessung der Turnierresultate. Dtsch Wochenschach 11(366–369):3
48. Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43
49. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117
50. Lee JH, Kim JW, Lee K, Choi MY (2020) Generalized maximal entropy argument for the gravity law in human mobility. Europhys Lett 132(4):48001
51. Cottineau C, Hatna E, Arcaute E, Batty M (2017) Diverse cities or the systematic paradox of urban scaling laws. Comput Environ Urban Syst 63:80–94
52. Alvioli M (2020) Administrative boundaries and urban areas in Italy: a perspective from scaling laws. Landsc Urban Plan 204:103906