CSIS Discussion Paper No. 150

The Size Distribution of 'Cities' Delineated with a Network Theory-based Method and Smartphone GPS Data*

Shota Fujishima[†] Naoya Fujiwara[‡]
Yuki Akiyama[§] Ryosuke Shibasaki[§] Hodaka Kaneda[¶]

October 10, 2017

Abstract

We apply a network theory-based method to delineate 'cities' independent of administrative boundaries in Japan. We divide the country into cells of 1×1 km^2 over which a random walker moves. Its switching probabilities are specified by GPS-based people flow data. We find the groups of the cells that describe the long-run behavior of the random walker in the most effective manner. The combination of two lognormal distributions yields a better fit of the city size distribution than the distribution with a Pareto upper tail. A jump diffusion process is a stochastic process of the city population underlying such a distribution.

Keywords: City size distribution; Power law; Mixture of distributions; Community detection; GPS data; Jump diffusion process

JEL Codes: R12; C46; C55

[*] We are grateful to Hiroyasu Inoue, Takaaki Aoki, the participants at the European Meeting of the 7th Urban Economics Association, and the seminar audience at the University of Tokyo and Kagawa University for their helpful comments and discussions. Ristu Sakuramachi helped us with processing the data and running the computer programs. Shota Fujishima acknowledges financial support from the Kikawada Foundation. Any remaining errors are our own.

[†] School of Management, Tokyo University of Science, 1-11-2 Fujimi, Chiyoda, Tokyo 102-0071, Japan. Phone/Fax: +81 3 3288 2563, Email: sfujishima@rs.tus.ac.jp

[‡] Institute of Industrial Science, the University of Tokyo.

[§] Center for Spatial Information Science, the University of Tokyo.

[¶] ZENRIN DataCom Co., Ltd.

1 Introduction

Economic activities are often geographically localized, where the spatial extent of economic activities is, for instance, measured by people's commuting flows. However, a spatial extent that is self-contained in terms of people's commuting flows is generally not guaranteed to coincide with preexisting administrative units such as city and county. This has motivated researchers and practitioners in regional science to delineate "metropolitan areas."

The U.S. Office of Management and Budget defines metropolitan statistical areas (MSAs), which we can obtain from the U.S. Census Bureau.[1] MSAs are constructed by merging counties that have strong economic and social ties. While MSAs are collections of legally bounded entities, Rozenfeld *et al.* (2011) consider the micro scale by considering cells of a given size. By using gridded population data, they delineate metropolitan areas with their *city clustering algorithm* in which they recursively grow the cluster by adding populated cells within a prescribed distance of the cluster.

In general, algorithmic methods such as that mentioned above iteratively merge geographical entities according to some criteria. These criteria depend on free parameters that researchers need to decide.[2] For example, Rozenfeld *et al.* (2011) must specify the prescribed distance, which decides the candidate neighboring cells to merge, to start their algorithm. Moreover, algorithmic methods do not explicitly construct an objective function to be optimized, which makes the underlying model unclear. This is particularly relevant when we want to relate the delineation of metropolitan areas to the decision problems of economic agents.[3]

In this study, we invoke insights from *community detection* in network theory.[4] A network is generally a collection of nodes and links that connect nodes. A community is then defined as a collection of nodes that have mutually strong relationships. Whether nodes have strong ties is decided by the link structure. For example, we can readily think of an urban economy as a network in which each geographical unit serves as a node and two nodes are linked if there is a flow of people between them. Thus, the communities in such a network would naturally correspond to metropolitan areas.

While various methods have been proposed in the literature on community detection, we use the *map equation* method developed by Rosvall and Bergstrom (2008). In our model, a random walker moves over geographical units. In particular, we consider the simplest possible model in which the switching probability between two units is proportional to the

[1] See `https://www.census.gov/programs-surveys/metro-micro.html`.

[2] In addition, the final result may be affected by the initial condition of the algorithm.

[3] We return to this point in the Conclusion.

[4] See Fortunato (2010) and Fortunato and Hric (2016) for overviews of this literature.

volume of the flow of people between them. Motivated by information theory, the optimal community structure minimizes the average code length necessary to describe the long-run behavior of the random walker. How clustering works to reduce the average code length is in the same spirit as the real address. Specifically, grouping geographical units into clusters allows us to assign the same name to multiple geographical units as long as they belong to different clusters. In fact, we see many Washington counties in the United States. This reduces the average code length. At the same time, however, clustering incurs a cost because we must assign names to the clusters. This tradeoff gives rise to the optimal clustering, where minimizing the "average" code length makes us assign a shorter code to those geographical units frequently visited by the random walker. Our task is formulated as a standard optimization problem because it follows that minimizing the average code length boils down to minimizing the entropy of the distribution of the random walker's long-run visit frequencies over nodes. As long as we accept the model, this method does not involve a free parameter that affects the resulting community structure. Moreover, we can readily capture the hierarchical structure of communities by making the entropy function hierarchical (Rosvall and Bergstrom, 2011). This approach thus enables us to look at metropolitan areas at various scales.[5]

We use GPS data on the flow of people in Japan taken from smartphone apps. Since we have high-resolution location information for each trip, we are free to choose the scale of geographical unit. Extracting about half-a-million commuting trips across Japan, we divide the country into cells of 1 km by 1 km and then identify the origin and destination cells for each trip. This yields the volumes of commuting flow for each pair of cells, which are used to compute the switching rates of the random walker.

Contrary to Rozenfeld *et al.* (2011) who use population data, we use only people flow data, which describe the *relationships* among geographical units.[6] If we rely on population data that describe the characteristics of each geographical unit, we typically need additional information. For example, it might become necessary to specify how the economic ties between geographical units decay over distance. In fact, Farmer and Fotheringham (2011), who use another popular method of community detection to delineate metropolitan areas in Ireland, make such an assumption.[7] On the contrary, people flow data already incorporate

[5] To our knowledge, this is the first work that delineates metropolitan areas while capturing their hierarchical structure.

[6] There is a large literature on detecting industrial agglomeration (see, e.g., Mori and Smith (2015) and references therein). In the context of industrial agglomeration, addressing the economic connections among geographical units requires micro data on, for example, transaction volumes among establishments, which is difficult to obtain. In fact, agglomeration is usually detected with only data on the spatial distribution of establishments.

[7] See the Appendix for more details.

the geographical structure because the volume of the flow of people tends to be small as the distance becomes large.

By adopting detected communities, or metropolitan areas, we study their size distribution, which has received considerable research attention in the urban economics literature (Gabaix and Ioannides, 2004). To simplify the expositions, we hereafter call metropolitan areas cities. Studying the city size distribution is particularly interesting here because our cities are not constrained to follow legal boundaries. Moreover, we can examine the size distribution of cities of all sizes. We understand that the main point at issue in the literature is whether the city size distribution is a single lognormal distribution (Eeckhout, 2004) or whether its upper tail is Pareto (Giesen and Suedekum, 2014; Ioannides and Skouras, 2013; Giesen *et al.*, 2010). Our conclusion is aligned with neither of these distributions. In fact, we find that a combination of two lognormal distributions is a better fit with the data than is a single lognormal distribution or a combination of lognormal and Pareto distributions. We further show that the stochastic process of the city population behind such a distribution is approximated by a *jump diffusion process*, which has a long history of application to finance since Merton (1976).

The rest of the paper proceeds as follows. In Section 2, we formally present our method. In Section 3, we explain the data we use. In Section 4, we visualize the detected communities and, in Section 5, we study their size distribution. The last section concludes.

2 The Model

We divide the whole of Japan into cells of approximately 1 km by 1 km and consider a model in which a random walker moves over the cells. Let n_{ij} be the number of workers commuting from cells i to j and m_{ij} be the number of workers returning home from cells i to j. The total number of commuting trips from cells i to j is then $N_{ij} = n_{ij} + m_{ij}$. The probability of a random walker switching from cells i to j, P_{ij}, is given by

$$P_{ij} = \frac{N_{ij}}{N_i}, \tag{1}$$

where $N_i = \sum_{j=1}^{S} N_{ij}$ and S is the total number of cells. Because $m_{ij} = n_{ji}$, $N_{ij} = n_{ij} + n_{ji}$ and thus N_i is the sum of the daytime and nighttime working populations of cell i.

We focus on the largest recurrent class of the Markov chain defined by (1). As we see in Section 4, this makes up around 95% of all populated cells in our data. Then, we are interested in the probabilities of a random walker staying in each cell in the long run, which we call the *long-run visit frequencies*. Because the vector of long-run visit frequencies

$p = (p_1, ..., p_S)$ is an invariant distribution of the Markov chain defined by (1), it satisfies

$$pP = p, \tag{2}$$

where $P = [P_{ij}]_{i,j}$ and S is the total number of cells in the largest recurrent class. Because we focus on the largest recurrent class, p uniquely exists.

We describe the long-run behavior of the random walker in our network by assigning binary codes, which are enumerations of numbers taking the values of either 0 or 1 such as '01' and '0010', to each state. Our objective is to code these as effectively as possible. To make this more precise, let ℓ_j be the length of the binary code assigned to state j. For example, the length of the binary code '0010' is 4. Then, we assign codes to states to minimize the following *average code length*:

$$\sum_j \ell_j p_j, \tag{3}$$

where p_j is the long-run probability of state j. The key idea here is to reduce the average code length by bringing cells together into several *communities* because this allows us to assign the same code to different cells as long as they belong to different communities. We also observe this type of information saving in real addresses: we see the same street name everywhere. However, grouping cells into communities makes it necessary for us to assign codes to each community. This tradeoff yields the optimal community structure.

We code by dividing the long-run behavior of the random walker into those inside each community and those across communities. Let $\{C^k\}$ be a community partition such that $C^k \cap C^\ell = \emptyset$ for any $k \neq \ell$ and $\bigcup_k C^k = S$ where $C^k \subseteq \{1, 2, ..., S\}$ is the set of cells that belong to community k. Observe that the number of communities is endogenously determined here. We describe the long-run behavior of the random walker inside community k by assigning binary codes to the states of visiting each cell in the community and the state of exiting the community. Then, the average code length for describing the long-run behavior of the random walker inside community k is

$$L^k = \ell_e^k \frac{q^k}{p^k + q^k} + \sum_{i \in C^k} \ell_i \frac{p_i}{p^k + q^k}, \tag{4}$$

where ℓ_i is the length of the binary code assigned to the state of visiting cell i, ℓ_e^k is the length of the binary code assigned to the state of exiting community k, q^k is the probability of exiting community k, and $p^k = \sum_{i \in C^k} p_i$. From (1), the probability of exiting community

k is

$$q^k = \sum_{i \in C^k} \sum_{j \notin C^k} p_i P_{ij} = \sum_{i \in C^k} \sum_{j \notin C^k} p_i \frac{N_{ij}}{N_i}. \tag{5}$$

On the contrary, we describe the long-run behavior of the random walker across communities by assigning binary codes to the states of visiting each community. Note that because our Markov chain is stationary,

$$\sum_{i \in C^k} \sum_{j \notin C^k} p_i P_{ij} = \sum_{i \in C^k} \sum_{j \notin C^k} p_j P_{ji}. \tag{6}$$

Hence, q^k also represents the probability of visiting community k from another community. Therefore, the average code length for describing the long-run behavior of the random walker across communities is

$$L = \sum_{k=1}^{K} \ell^k \frac{q^k}{q}, \tag{7}$$

where ℓ^k is the length of the binary code assigned to the state of visiting community k, K is the number of communities, and $q = \sum_{k=1}^{K} q^k$.

We consider the weighted sum of L and $\{L^k\}$, where the weight of L is q whereas the weight of L^k is $p^k + q^k$. Therefore, what we try to minimize is given by

$$qL + \sum_{k=1}^{K} (p^k + q^k) L^k. \tag{8}$$

It might seem a daunting task to find the community structure and coding that minimize the average code length above. However, *Shannon's source coding theorem* simplifies our task. To state the theorem, we define that the *entropy* of a probability distribution $(\pi_1, ..., \pi_I)$ is

$$H(\pi_1, ..., \pi_I) = \sum_{i=1}^{I} \pi_i \log \frac{1}{\pi_i}. \tag{9}$$

Then, the theorem states that it is possible to make the average code length arbitrarily close to the entropy of the underlying probability distribution.[8]

In our coding problem, the entropies of the probability distributions in L^k and L are $H(\frac{q^k}{p^k+q^k}, \{\frac{p_i}{p^k+q^k}\}_{i \in C^k})$ and $H(\frac{q^1}{q}, ..., \frac{q^K}{q})$, respectively. Therefore, our task reduces to finding

[8] This is a fundamental theorem in information theory. See, e.g., Theorem 5.4.2 of Cover and Thomas (2012).

the community partition that minimizes

$$L^*(C^1, ..., C^K) = qH\left(\tfrac{q^1}{q}, ..., \tfrac{q^K}{q}\right) + \sum_{k=1}^{K}(p^k + q^k)H\left(\tfrac{q^k}{p^k+q^k}, \left\{\tfrac{p_i}{p^k+q^k}\right\}_{i\in C^k}\right). \tag{10}$$

Rosvall and Bergstrom (2008) call the objective function above the *map equation.*

Moreover, this method can be extended to allow for the hierarchical structure of communities, as is done by Rosvall and Bergstrom (2011). Let $C^{k\ell} \subseteq C^k$ be a subcommunity of community k, $C^{k\ell m} \subseteq C^{k\ell}$ be a subsubcommunity of subcommunity ℓ, and so forth. Our objective function is then constructed recursively as

$$L^*(\{C^k\}_k) = qH\left(\tfrac{q^1}{q}, ..., \tfrac{q^k}{q}\right) + \sum_k L^*(\{C^{k\ell}\}_\ell), \tag{11}$$

where

$$L^*(\{C^{k\ell}\}_\ell) = (p^k + q^k)H\left(\tfrac{q^k}{p^k+q^k}, \left\{\tfrac{p_i}{p^k+q^k}\right\}_{i\in C^k}\right) + \sum_\ell L^*(\{C^{k\ell m}\}_m), \tag{12}$$

and, at the lowest level of the hierarchy,

$$L^*(C^{k\ell\cdots r}) = (p^{k\ell\cdots r} + q^{k\ell\cdots r})H\left(\tfrac{q^{k\ell\cdots r}}{p^{k\ell\cdots r}+q^{k\ell\cdots r}}, \left\{\tfrac{p_i}{p^{k\ell\cdots r}+q^{k\ell\cdots r}}\right\}_{i\in C^{k\ell\cdots r}}\right). \tag{13}$$

Note that the depth of the hierarchy is also a choice variable. We exploit the flexibility of this method, and detect the hierarchical structure of metropolitan areas.

3 Data

We use *Konzatsu-Tokei*® (Congestion Statistics), a GPS dataset of the flow of people in Japan provided by ZENRIN DataCom Co., Ltd. *Konzatsu-Tokei*® refers to the people flow data collected from the individual location data sent from mobile phones with the Auto-GPS function, enabled under users' consent, through the "docomo map navi" service provided by NTT DOCOMO, INC (Japan's primary mobile service provider). Those data are processed collectively and statistically to conceal private information. The number of users, who live across Japan, is around half-a-million. The information on people's trips is so detailed that the time interval between the acquisitions of location information via GPS is five minutes at its the shortest interval. Moreover, the data are panel data that record each individual's trips every day.[9] The data we use were collected for one year from January 1, 2012 to December

[9] In generating *Konzatsu Tokei*®, NTT Docomo, INC performed an overall and statistical processing of GPS data as per order of ZENRIN DataCom CO., LTD. This applies to all the figures presented in this

31, 2012. Each data entry includes information such as the unique user ID, location (latitude and longitude), and time stamp. The attributes of users such as age and sex are not available.

Although detailed information on people's trips is available, the purpose of each trip is not specified. Moreover, each user's place of residence and work place, if he or she works, are not known. Hence, to delineate metropolitan areas by using these data, we need to extract commuting trips. To this end, we start by detecting the *stops* in each trip. Let us represent the trip of a user by a sequence of locations $\{x_1, x_2, x_3, ...\}$, where x_i is the i-th location of the user. The location is composed of latitude y, longitude z, and time t. Hence, $x_i = (y_i, z_i, t_i)$. Then, given the threshold values S and T, we define the *stop* as a set of locations $\{x_k, x_{k+1}, x_{k+2}, ...x_{k+m}\}$ such that the distance between (y_i, z_i) and (y_j, z_j) is less than S for any $i, j \in \{k, k+1, k+2, ..., k+m\}$ and $t_{k+m} - t_k > T$. That is, we regard a set of locations as a stop if all the locations in the set are close to each other and the user stayed there for a reasonably long time. We set S to 200 meters and T to five minutes. Given these stops, we then identify the residential and working zones of users. Specifically, the residential zone is identified as the zone to which the most frequent stop during the night (10pm-6am) belongs, whereas the working zone is identified as the zone to which the most frequent stop during daytime hours (9am-5pm) belongs.[10]

By following the procedure outlined above, we extract around 540,000 commuting trips. To assess the reliability of our extracted data, we conduct two reliability checks. First, we aggregate the data to calculate the residential densities for each grid of cell size 1 km by 1km and compare those with the residential densities from the Grid Square Statistics of the 2010 Population Census. As shown in Figure 1(a), we obtain a correlation coefficient of around 0.9393. Second, we compute the trip volumes for each pair of municipal districts and compare them with those from the 2011 Person Trip Survey conducted in the Chukyo metropolitan area.[11] As shown in Figure 1(b), we obtain a correlation coefficient of around 0.9369. These results show that our data are consistent with aggregated public data.

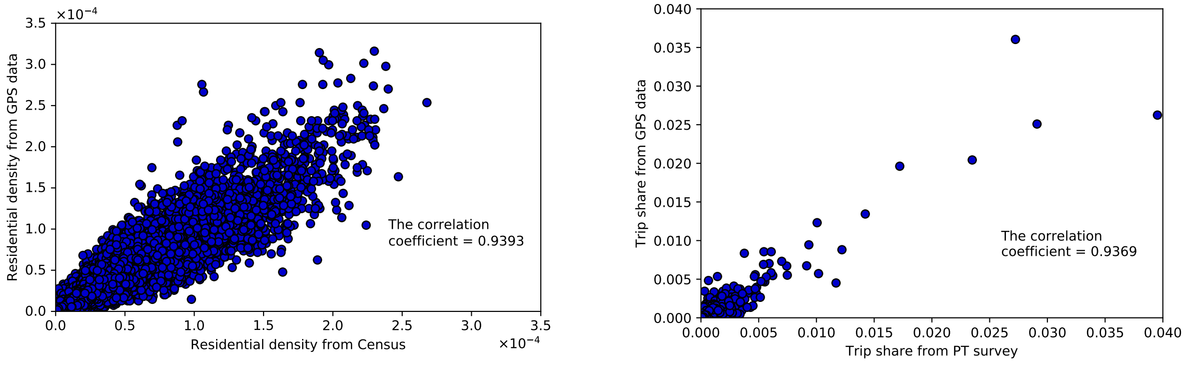
4 Detected Communities

By using the commuting data developed above, we first construct the Markov chain defined by (1) and find its largest recurrent class. Of the 85,607 cells, the largest recurrent class has

paper.

[10] We refer to Horanont *et al.* (2013) for this method.

[11] The data are similar to those derived from the National Household Travel Survey in the United Sates. The data are publicly available online at `http://www.cbr.mlit.go.jp/kikaku/chukyo-pt/persontrip/p01.html` (in Japanese). The Chukyo metropolitan area covers parts of the Aichi, Gifu, and Mie prefectures. The largest city in this metropolitan area is Nagoya, which is located between Tokyo and Osaka.



「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

(a) GPS data vs. Census

(b) GPS data vs. Person Trip Survey

Figure 1: Reliability checks of GPS data

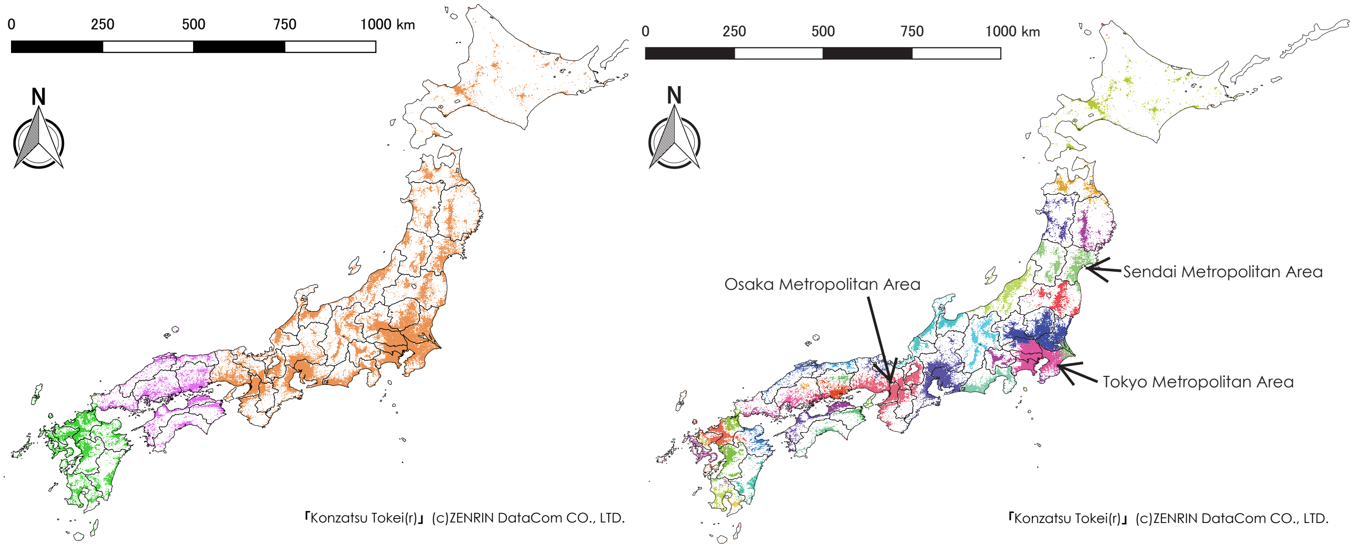
80,926 cells. We then apply the map equation method to our data limited to the largest recurrent class.[12] As a result, the cells are grouped into three level-1 (i.e., the highest level of the hierarchy) communities. As shown in Figure 2(a), the rightmost community is prominent: it has major cities such as Tokyo, Osaka, and Nagoya and, moreover, the Hokkaido and the whole of the Tohoku region are included in the community.

However, as the first level is too coarse to obtain insights for urban agglomerations, we proceed to the lower level. Here, we detect 55 level-2 communities, which are depicted in Figure 2(b). This level has an intuitively relevant scale for metropolitan areas. For example, we can think of the rose community in the Kanto region as the Tokyo metropolitan area, the salmon pink community in the Kansai region as the Osaka metropolitan area, the lime community in the Tohoku region as the Sendai metropolitan area, and so forth. These communities are divided into subcommunities (level-3 communities). The total number of level-3 communities is 2,048, although 30% are composed of fewer than 10 cells. Figure 3 depicts the level-3 communities in the Tokyo metropolitan area. We detect up to the sixth level for the hierarchy of communities, although the depth of the hierarchy is generally different among parent communities.

5 City Size Distribution

In this section, we study the size distribution of the detected communities in terms of population, where the population of community k is computed as $\sum_{i\in C^k} N_i$. To simplify the

[12] To carry out the map equation method, we use the code provided by Daniel Edler and Martin Rosvall at `http://www.mapequation.org/`.



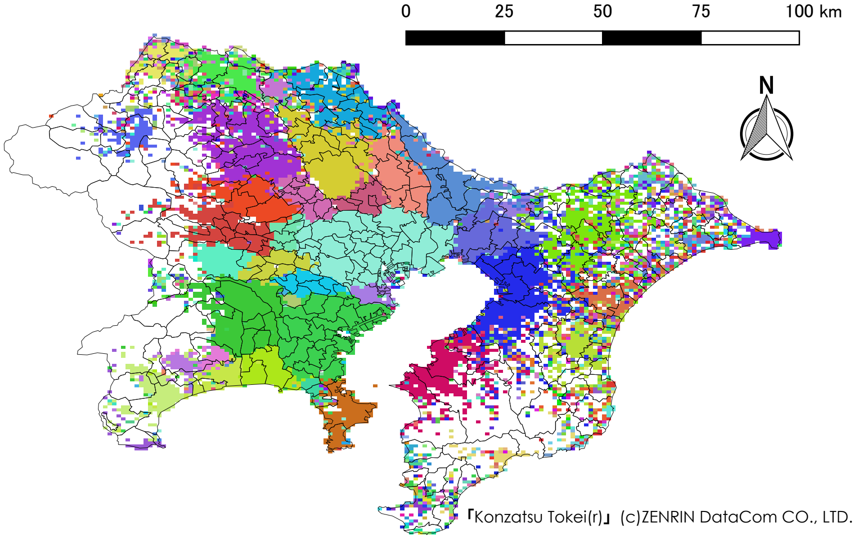
「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

(a) Level-1 communities

(b) Level-2 communities

Figure 2: Maps of communities



「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

Figure 3: Map of level-3 communities in Tokyo area

expositions, communities, or metropolitan areas, are hereafter interchangeably called cities. Studying the city size distribution, which has been a major research topic in the urban economics literature, is particularly interesting here because our cities are independent of administrative boundaries and, by definition, the populations of *all* cities are available.

In the context of the city size distribution, *Zipf's law* has been regarded as an important regularity condition for urban economics models (Gabaix and Ioannides, 2004). Let us review several concepts. A *power law* is a distribution function of the form $\Pr(X \geq x) \propto a/x^{\eta}$ where $a, \eta > 0$. Alternatively, we can say that X follows the *Pareto distribution*. Then, Zipf's law corresponds to the case where $\eta = 1$. When visualizing a power law, it is convenient to rank the realized values of the variable (population in our context) by their sizes. Let $x_{(i)}$ be the i-th largest city's population and K be the total number of cities. Under Zipf's law, $x_{(i)}^{\eta} \Pr(X \geq x_{(i)}) = a$, where, as long as K is sufficiently large, $\Pr(X \geq x_{(i)})$ may be regarded as the share of cities larger than city i. Thus,

$$\Pr(X \geq x_{(i)}) \approx i/K \Rightarrow \log x_{(i)} \approx c - \eta^{-1} \log i, \tag{14}$$

where $c = \eta^{-1} \log(aK)$. Therefore, under a power law, plotting log-population against log-rank should yield the slope of $-1/\eta$ as long as the sample size is sufficiently large.[13] See Figure 4 for this type of graph plotted with our data.

Although Zipf's law has been documented for the city size distributions in many countries, Eeckhout (2004) points out that it appears to be important because only large cities are considered. By using population data on "places," which were newly introduced geographical units in the U.S. Census at that time, he claims that the city size distributions of *all cities* are best fit by using a lognormal distribution. Observe that we also have the whole sample of cities (i.e., detected communities).

Some authors, however, claim that a single lognormal distribution is insufficient to describe the city size distribution and that the power law is important for, at least, the upper tail of the distribution. It seems that this debate has not yet been settled.[14] However, even if the power law is relevant to the upper tail, it is still sound to consider the whole sample because it is difficult to truncate the sample in a convincing way.[15] Then, a sensible suggestion to this debate would be to consider a distribution such that its body is characterized by the lognormal distribution, whereas its tails are characterized by the Pareto distributions. In fact, Giesen *et al.* (2010) and Giesen and Suedekum (2014) consider the *double Pareto lognormal distribution* (DPLN) for U.S. city size distributions, concluding that the DPLN has a better fit with the data than the lognormal distribution.[16] This distribution, first

[13] Note that when (14) holds with equality and $\eta = 1$, we have $x_{(i)} = \frac{1}{i} x_{(1)}$ because $\log x_{(1)} = c$. Thus, the size of the second largest city is half that of the largest city, the size of the third largest city is two-thirds that of the second largest city, and so forth. Therefore, Zipf's law is also called the *rank-size rule*.

[14] See, for example, Levy (2009), Eeckhout (2009), Giesen *et al.* (2010), Bee *et al.* (2013), Ioannides and Skouras (2013), and Giesen and Suedekum (2014).

[15] Eeckhout (2009) argues that truncating sample results in biased conclusions.

[16] Ioannides and Skouras (2013) also consider a combination of lognormal and Pareto distributions. They assume that there exists a switching population $\bar{S}$ such that the distribution is lognormal for $S \leq \bar{S}$ and is Pareto for $S \geq \bar{S}$, subject to the regularity condition that the density integrates to one. Their estimation results also indicate that such a mixed distribution is better than the single lognormal distribution.

introduced by Reed (2002),[17] is obtained as a normal variance mixture by the exponential distribution (see the Appendix).

Thus, we also consider the lognormal distribution and DPLN. However, although the result that the combination of lognormal and Pareto distributions is better than the lognormal distribution appears to be reasonable, it might be unfair to require a *single* lognormal distribution to describe a city size distribution for the whole sample. Therefore, we additionally consider a combination of two lognormal distributions. We present the stochastic process of the city population behind this type of distribution in the next section.

In sum, the density functions we fit to our data are given as follows. For convenience, we consider the densities of log-population which is denoted by x:

$$f_N(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \tag{15}$$

$$f_{DPLN}(x) = \frac{\alpha\beta}{\alpha+\beta}\phi\left(\frac{x-\nu}{\tau}\right)\left\{m\left(\alpha\tau-\frac{x-\nu}{\tau}\right)+m\left(\beta\tau+\frac{x-\nu}{\tau}\right)\right\}, \tag{16}$$

$$f_{mixN}(x) = \theta\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)+(1-\theta)\frac{1}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right), \tag{17}$$

where ϕ is the standard normal density and m is the Mills ratio of the standard normal law (i.e., $m(x) = \frac{1-\Phi(x)}{\phi(x)}$, where Φ is the cumulative standard normal distribution). f_N corresponds to the case where the city population follows the lognormal distribution, f_{DPLN} corresponds to the case where the city population follows the DPLN, and f_{mixN} corresponds to the case where the city population follows the combination of two lognormal distributions. See the Appendix for the derivation of the density of the DPLN.

We fit the three distributions above to our data by using the maximum likelihood estimation. Table 1 summarizes the estimation results.

[17] See also Reed and Jorgensen (2004).

Table 1: Estimated parameters of the three distributions fitted to the level-3 communities

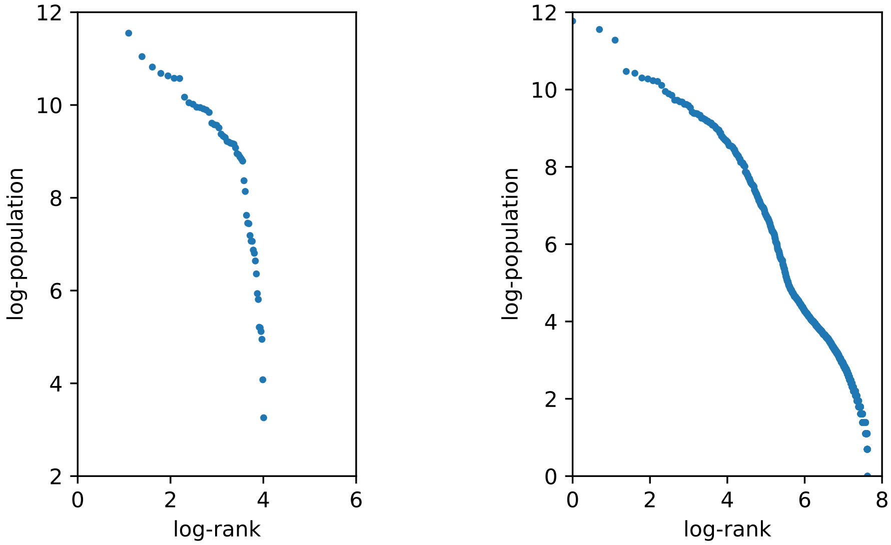
	Normal	Normal + Normal	DPLN
μ_1	3.3472 (0.041)	2.7910 (0.033)	
σ_1	1.8549 (0.029)	1.0635 (0.024)	
μ_2		7.0222 (0.280)	
σ_2		1.7743 (0.171)	
θ		0.8685 (0.014)	
α			0.5634 (0.013)
β			8.6315 (0.013)
ν			1.6852 (0.006)
τ			0.5843 (0.002)
Log-Likelihood	-4,171	-3,778	-3,799
AIC	8,346	7,566	7,606
BIC	8,357	7,594	7,628

Standard errors are in parentheses.

These estimation results indicate that the DPLN is a better fit to our data than the lognormal distribution in terms of any of likelihood, AIC, and BIC, which is in line with the findings of Giesen *et al.* (2010) and Giesen and Suedekum (2014). However, the results also show that the combination of two lognormal distributions is better than the DPLN in terms of any of likelihood, AIC, and BIC. This finding implies that although the lognormal distribution is sufficient to describe the city size distribution, we must consider a combination of such distributions. Hence, our results indicate that Pareto distributions are not relevant, even for the tails.

We can also confirm this result graphically. Figure 4 plots the log-populations against the log-ranks for the level-2 and level-3 communities. We can hardly see from Figure 4(b) that Zipf's law, or more generally a power law, holds even for the upper tail. In the upper left of Figure 4(b), the graph seems to be concave rather than a straight line; moreover, there is a jump between the log-populations of the third and fourth largest cities.[18]

[18] On the contrary, the upper left of Figure 4(a) seems to be convex. This finding implies that the power law does not hold for level-2 communities either.



「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

「Konzatsu Tokei(r)」 (c)ZENRIN DataCom CO., LTD.

(a) Level-2 Communities

(b) Level-3 Communities

Figure 4: The size distributions of communities in terms of population

5.1 Stochastic process behind the Gaussian mixture

One of the important motivations behind studying the city size distribution is to provide regularity for the underlying theoretical model because the relevant model should yield the city size distribution observed in the real world as an equilibrium. Although Eeckhout (2004) and Giesen and Suedekum (2014) present the stochastic processes of the city population for the lognormal distribution and DPLN, respectively, we are not aware of work that does so for a combination of lognormal distributions. In this section, we show that adding a compound Poisson process, where the jump amplitude is stochastic, to a geometric Brownian motion yields the desired distribution. The resulting process has been commonly used in finance since the seminal work of Merton (1976).

Let $Q(t) = \sum_{i=1}^{N(t)} Y_i$, where N is a Poisson process with rate λ and $1+Y_i$ is iid lognormally distributed with mean μ and δ^2. Q is called a *compound Poisson process*. Then, we assume that $S(t)$ follows the following process:

$$\mathrm{d}S(t) = \gamma S(t)\mathrm{d}t + \xi S(t)\mathrm{d}W(t) + S(t-)\mathrm{d}Q(t), \tag{18}$$

where W is a Wiener process and $S(t-) = \lim_{s \nearrow t} S(s)$. This is called a *jump diffusion process* (see, e.g., Shreve, 2004, ch. 11). Given $S(0) = s_0$, the solution to (18) is given by the Doleans-Dade exponential:

$$S(t) = s_0 \exp\left\{\left(\gamma - \frac{1}{2}\xi^2\right)t + \xi W(t)\right\}\prod_{i=1}^{N(t)}(1+Y_i). \tag{19}$$

Note that because Y_i is lognormally distributed, $Y_i > -1$. This assures that $S(t)$ does not jump to a negative value. Let $X(t) = \log S(t)$. Then,

$$X(t) = X(0) + \left(\gamma - \frac{1}{2}\xi^2\right)t + \xi W(t) + \sum_{i=1}^{N(t)}\log(1+Y_i). \tag{20}$$

By discretizing the time with interval Δ, we obtain

$$X(t) = X(t-\Delta) + \left(\gamma - \frac{1}{2}\xi^2\right)\Delta + \xi\big(W(t) - W(t-\Delta)\big) + \sum_{i=N(t-\Delta)+1}^{N(t)}\log\big(1+Y_i\big). \tag{21}$$

We consider the probability transition density $p(x;\Delta|y,k)$ that satisfies[19]

$$\Pr\big[X(t) \in [x, x+\mathrm{d}x]\,\big|\,X(t-\Delta) = y, N(t) - N(t-\Delta) = k\big] = p(x;\Delta|y,k)\mathrm{d}x. \tag{22}$$

By assumption, $W(t) - W(t-\Delta) \sim \mathcal{N}(0,\Delta)$ and $\log(1+Y_i) \sim \mathcal{N}(\mu,\delta^2)$. Then, from (21),

$$p(x;\Delta|y,k) = \frac{1}{\sqrt{\xi^2\Delta + k\delta^2}}\phi\left(\frac{x - y - \left(\gamma - \frac{1}{2}\xi^2\right)\Delta - k\mu}{\sqrt{\xi^2\Delta + k\delta^2}}\right). \tag{23}$$

Furthermore, because $X(t-\Delta)$ and $N(t) - N(t-\Delta)$ are independent, the variable representing the number of jumps is integrated out from the probability transition density as

$$\begin{aligned} p(x;\Delta|y) = \frac{p(x,y;\Delta)}{p(y;\Delta)} &= \sum_{k=0}^{\infty}\frac{p(x,y,k;\Delta)\Pr[N(t)-N(t-\Delta)=k]}{p(y;\Delta)\Pr[N(t)-N(t-\Delta)=k]} && (24)\\ &= \sum_{k=0}^{\infty}\frac{p(x,y,k;\Delta)\Pr[N(t)-N(t-\Delta)=k]}{p(y,k;\Delta)} && (25)\\ &= \sum_{k=0}^{\infty}p(x;\Delta|y,k)\Pr[N(t)-N(t-\Delta)=k]. && (26)\end{aligned}$$

[19] We focus on the transition density here because the stationary distribution does not exist, as in the case of the geometric Brownian motion. However, it is possible to "stabilize" the process so that the stationary distribution exists. One possible way is to allow for the possibility of people's death. In such a model, Gabaix *et al.* (2016) show in Proposition 8 that if the initial distribution and distribution of jump amplitudes both have Pareto tails, the stationary distribution also has a Pareto (upper) tail.

Because N is a Poisson process with rate λ,

$$\Pr[N(t) - N(t-\Delta) = 0] = 1 - \lambda\Delta + o(\Delta), \tag{27}$$

$$\Pr[N(t) - N(t-\Delta) = 1] = \lambda\Delta + o(\Delta), \tag{28}$$

$$\Pr[N(t) - N(t-\Delta) \geq 2] = o(\Delta). \tag{29}$$

Hence, we approximate $N(t) - N(t-\Delta)$ by the random variable Z that follows the Bernoulli distribution with parameter $\lambda\Delta$ for a short time interval Δ. Then, we have

$$\begin{aligned} p(x;\Delta|y) &\approx \sum_{z\in\{0,1\}} p(x;\Delta|y,z)\Pr(Z=z) && (30)\\ &= \lambda\Delta p(x;\Delta|y,1) + (1-\lambda\Delta)p(x;\Delta|y,0). && (31) \end{aligned}$$

Therefore,

$$\begin{aligned} X(t)|X(t-\Delta) = y \sim \lambda\Delta \times \mathcal{N}\left(y + \left(\gamma - \frac{1}{2}\xi^2\right)\Delta + \mu, \xi^2\Delta + \delta^2\right) \\ + (1-\lambda\Delta) \times \mathcal{N}\left(y + \left(\gamma - \frac{1}{2}\xi^2\right)\Delta, \xi^2\Delta\right). \end{aligned} \tag{32}$$

Note that our estimation result is consistent with the regularity conditions behind (32). First, for the approximation through the Bernoulli distribution to be accurate, Δ must be sufficiently small. Thus, one of the two normal distributions must have a sufficiently small weight. This means that $\theta \geq \frac{1}{2}$ must be sufficiently large, and our estimate of θ is around 0.87. Second, because $\mu, \delta > 0$, the normal distribution having the smaller weight must have the larger mean and variance. Because the estimates of μ_2 and σ_2 are larger than those of μ_1 and σ_1, respectively, this condition is also met. Therefore, we may conclude that the above model is relevant to our data.[20]

[20] Note that if we do not approximate $N(t) - N(t-\Delta)$ by using the Bernoulli distribution, the transition density is an infinite mixture of the normal distributions according to the Poisson distribution. In fact, because $\Pr[N(t) - N(t-\Delta) = k] = \frac{(\lambda\Delta)^k}{k!}e^{-\lambda\Delta}$, $p(x;\Delta|y) = \sum_{k=0}^{\infty} \frac{(\lambda\Delta)^k}{k!}e^{-\lambda\Delta}\frac{1}{\sqrt{\xi^2\Delta + k\delta^2}}\phi\left(\frac{x - y - \left(\gamma - \frac{1}{2}\xi^2\right)\Delta - k\mu}{\sqrt{\xi^2\Delta + k\delta^2}}\right)$. Yu (2007) proposes a method of approximating the likelihood function based on this density.

6 Conclusion

By using the network theory-based map equation method following Rosvall and Bergstrom (2008) and smartphone GPS data, we delineated "cities" in Japan that are independent of administrative boundaries and then examined the size distribution of the delineated cities. Contrary to previous observations in which the power law has been regarded as an important regularity, we found that mixing lognormal distributions is sufficient to describe the city size distribution and hence the Pareto distribution does not play a role. We argued that in such a case, the jump diffusion process is relevant to the stochastic process of the city population.

Having said that, a fundamental problem in this literature is that how our cities look like is decided by our definition of city. Hence, it is important to connect the definition of city to economics. One possible way of doing so might be to consider a coalition formation game of local governments (Weese, 2015). One conceptual advantage of the map equation method is that it is explicitly formulated as an optimization problem. Moreover, it is well known that minimizing the entropy, as in the map equation method, is related to finding the equilibrium of logit-type discrete choice models. Hence, there would be potential for doing this in the map equation method.

Appendix

A1 Comparison with the modularity method

Several methods have been proposed for community detection as summarized in Fortunato and Hric (2016). Among other things, Farmer and Fotheringham (2011) use the modularity method of Newman and Girvan (2004) to delineate metropolitan areas. This method searches for a community partition such that the volume of flows within a community is large, whereas that of flows between communities is small. The performance of a community partition is evaluated relative to the case where cells are placed completely at random. Specifically, if cells are placed at random, the expected number of commuting trips from cells i to j is $N \times \frac{N_i}{N} \times \frac{N_j}{N} = \frac{N_i N_j}{N}$. The *modularity* is then defined as

$$Q = \frac{1}{N} \sum_{i,j} \left(N_{ij} - \frac{N_i N_j}{N} \right) \delta_{ij}, \tag{33}$$

where $\delta_{ij} = 1$ if cells i and j belong to the same community and $\delta_{ij} = 0$ otherwise. The modularity method finds the community partition that maximizes the modularity, which does not depend on the free parameters. However, because $\frac{N_i N_j}{N}$ only depends on the numbers of workers in each cell, it can take a large value even if the two cells are far apart as long as they host large numbers of workers. Owing to this, Farmer and Fotheringham (2011), who use commuting flow data in Ireland, needed to discount the flow volumes by geographical distance to obtain communities of reasonable sizes. However, this means that the resulting community structure depends on the discounting method. The map equation method, on the contrary, does not use information on the population of each cell and hence spatial discounting is not necessary. In fact, the volume of the flow of people, which is the only information used by the map equation method, naturally tends to be small as the distance between origin and destination becomes large. Moreover, the map equation method can accommodate the hierarchical structure of communities and hence detect various scales. Hence, we do not have to explicitly take the distance into account.

A2 DPLN

As demonstrated by Eeckhout (2004), conditional on the city age and population in the previous period, the city size distribution is given by a lognormal distribution if the logarithm of the city population follows a geometric Brownian motion. Specifically, let $S(t)$ be the city population at time t, and suppose that it obeys the following process:

$$\mathrm{d}S(t) = \gamma S(t)\mathrm{d}t + \xi S(t)\mathrm{d}W(t), \tag{34}$$

where W is a Wiener process. Given $S(0) = s_0$, it follows from the standard Itô calculus that the solution to (34) is

$$S(t) = s_0 e^{\left(\gamma - \frac{1}{2}\xi^2\right)t + \xi W(t)}. \tag{35}$$

Let $X(t) = \log S(t)$. Then,

$$X(t) = X(0) + \left(\gamma - \frac{1}{2}\xi^2\right) t + \xi W(t). \tag{36}$$

Suppose that $S(0)$ is lognormally distributed with mean ν and variance τ^2 (i.e., $X(0) \sim \mathcal{N}(\nu, \tau^2)$). Because W is a Wiener process (i.e., $W(t) \sim \mathcal{N}(0, t)$), we have

$$X(t) \sim \mathcal{N}\left[\nu + \left(\gamma - \frac{1}{2}\xi^2\right) t, \tau^2 + \xi^2 t\right]. \tag{37}$$

The distribution above depends on t, which is the city age in our context. Giesen and Suedekum (2014) assume that the city age is exponentially distributed with rate parameter

λ. Hence, the density function of t is $f_T(t) = \lambda e^{-\lambda t}$. Then, the unconditional density of X is

$$f_X(x) = \int \frac{1}{\sigma_t}\phi\left(\frac{x-\mu_t}{\sigma_t}\right) f_T(t)\mathrm{d}t, \tag{38}$$

where ϕ is the standard normal density, $\mu_t = \nu + \left(\gamma - \frac{1}{2}\xi^2\right)t$, and $\sigma_t^2 = \tau^2 + \xi^2 t$. Let $m(z) = \frac{1-\Phi(z)}{\phi(z)}$ where Φ is the standard normal cumulative distribution. It is shown that

$$f_X(x) = \frac{\alpha\beta}{\alpha+\beta}\phi\left(\frac{x-\nu}{\tau}\right)\left\{m\left(\alpha\tau - \frac{x-\nu}{\tau}\right) + m\left(\beta\tau + \frac{x-\nu}{\tau}\right)\right\}, \tag{39}$$

where α and $-\beta$ $(\alpha, \beta > 0)$ are the roots of the following equation:

$$\frac{\xi^2}{2}z^2 + \left(\gamma - \frac{1}{2}\xi^2\right)z - \lambda = 0. \tag{40}$$

References

Bee, M., Riccaboni, M. and Schiavo, S. (2013). The size distribution of us cities: Not pareto, even in the tail. *Economics Letters*, **120** (2), 232–237.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *The American Economic Review*, **94** (5), 1429–1451.

— (2009). Gibrat's law for (all) cities: Reply. *The American Economic Review*, **99** (4), 1676–1683.

Farmer, C. J. and Fotheringham, A. S. (2011). Network-based functional regions. *Environment and Planning A*, **43** (11), 2723–2741.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, **486** (3), 75–174.

— and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, **659**, 1–44.

Gabaix, X. and Ioannides, Y. M. (2004). The evolution of city size distributions. *Handbook of regional and urban economics*, **4**, 2341–2378.

—, Lasry, J.-M., Lions, P.-L. and Moll, B. (2016). The dynamics of inequality. *Econometrica*, **84** (6), 2071–2111.

Giesen, K. and Suedekum, J. (2014). City age and city size. *European Economic Review*, **71**, 193–208.

—, Zimmermann, A. and Suedekum, J. (2010). The size distribution across all cities–double pareto lognormal strikes. *Journal of Urban Economics*, **68** (2), 129–137.

Horanont, T., Phithakkitnukoon, S., Leong, T. W., Sekimoto, Y. and Shibasaki, R. (2013). Weather effects on the patterns of people's everyday activities: a study using gps traces of mobile phone users. *PloS one*, **8** (12), e81153.

Ioannides, Y. and Skouras, S. (2013). Us city size distribution: Robustly pareto, but only in the tail. *Journal of Urban Economics*, **73** (1), 18–29.

Levy, M. (2009). Gibrat's law for (all) cities: Comment. *The American Economic Review*, **99** (4), 1672–1675.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of financial economics*, **3** (1-2), 125–144.

Mori, T. and Smith, T. E. (2015). On the spatial scale of industrial agglomerations. *Journal of Urban Economics*, **89**, 1–20.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, **69** (2), 026113.

Reed, W. J. (2002). On the rank-size distribution for human settlements. *Journal of Regional Science*, **42** (1), 1–17.

— and Jorgensen, M. (2004). The double pareto-lognormal distribution - a new parametric model for size distributions. *Communications in Statistics-Theory and Methods*, **33** (8), 1733–1753.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105** (4), 1118–1123.

— and — (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, **6** (4), e18209.

Rozenfeld, H. D., Rybski, D., Gabaix, X. and Makse, H. A. (2011). The area and population of cities: new insights from a different perspective on cities. *The American Economic Review*, **101** (5), 2205–2225.

Shreve, S. E. (2004). *Stochastic calculus for finance II: Continuous-time models*, vol. 11. Springer Science & Business Media.

Weese, E. (2015). Political mergers as coalition formation: an analysis of the heisei municipal amalgamations. *Quantitative Economics*, **6** (2), 257–307.

Yu, J. (2007). Closed-form likelihood approximation and estimation of jump-diffusions with an application to the realignment risk of the chinese yuan. *Journal of Econometrics*, **141** (2), 1245–1280.