The Size Distribution of ‘Cities’ Delineated with a Network Theory-based Method and Mobile Phone GPS Data*

Shota Fujishima† Naoya Fujiwara‡
Yuki Akiyama§ Ryosuke Shibasaki§ Ritsu Sakuramachi¶

February 13, 2019

Abstract

We delineate “cities” independent of administrative boundaries in Japan by using a network theory-based method and GPS-based human mobility data. We divide the country into approximately 1×1 km² cells and detect the partition of cells that is optimal from the perspective of information theory. The resulting groups of cells are specified as cities. We find that the combination of two lognormal distributions better fits the city size distribution compared with the distribution with a Pareto upper tail. Moreover, we show that a jump diffusion process is the stochastic process of the city population underlying such a distribution.

JEL classification: R12; C46; C55.

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

*We are grateful to Hiroyasu Inoue, Takaaki Aoki, Daisuke Oyama, 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. 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
‡Graduate School of Information Sciences, Tohoku University.
§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 in terms of the link structure. 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 volume of the flow of people between them. Motivated by information theory, the optimal community structure minimizes the amount of information, or the average code length, necessary to describe the long-run behavior of the random walker. How clustering works to reduce the average code

---

\(^1\)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.
length is in the same spirit as the real address. Specifically, grouping geographical units into communities allows us to assign the same name to multiple geographical units as long as they belong to different communities. In fact, we see many Washington counties in the United States. This reduces the average code length. At the same time, however, it incurs a cost because we must assign names to the communities. This tradeoff gives rise to the optimal grouping, 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 mobile phone apps, with which we obtain about half-a-million commuting trips across Japan. Since we have high-resolution location information for each trip, we are free to choose the scale of geographical unit. We divide the country into approximately 1 km by 1 km cells and aggregate the data to obtain the volumes of commuting flow for each pair of cells, which are then used to compute the switching rates of the random walker. One potential problem with using these types of data is sample selection bias because they contain only people who use the mobile phone apps of a particular mobile service provider. To address this issue, we compare our data with other publicly available data such as the Census and Household Travel Survey. For example, we compute the population shares of each cell from our commuting data and find that they are highly correlated with the population shares from the Grid Square Statistics of the Population Census.

Contrary to Rozenfeld et al. (2011) who use population data, we use only human mobility 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,

---

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

6There 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.
make such an assumption. On the contrary, human mobility data already incorporate the geographical structure because the volume of the flow of people tends to be small as the distance becomes large. Note that unlike typical commuting data such as the Household Travel Survey, GPS data enable us to obtain commuting flows among small geographical units nationwide, which is why we may resort to using only information describing the relationships among geographical units.

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, contrary to most of the previous studies, our cities are not constrained to follow legal boundaries. Moreover, we can examine the size distribution of cities of all sizes. A 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

$$P_{ij} = \frac{N_{ij}}{N_i}, \quad (2.1)$$

\footnote{See Appendix A.2 for more detail.}
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 (2.1). 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 (2.1), it satisfies

\[
pP = p,
\]

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 group cells into several communities which correspond to cities. Formally, let \( \{C^k\} \) be a community partition such that satisfies \( C^k \cap C^\ell = \emptyset \) for any \( k \neq \ell \) and \( \bigcup_k C^k = \{1, 2, ..., S\} \), 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.

As we discussed in the introduction, we seek the community partition that describes the long-run behavior of the random walker in the most effective manner in terms of the amount of information. Appealing to information theory, this reduces to finding the community partition that minimizes the following function:

\[
L^*(C^1, ..., C^K) = qH\left(\frac{q_1}{q}, ..., \frac{q_K}{q}\right) + \sum_{k=1}^{K} (p^k + q^k)H\left(\frac{q^k}{p^k + q^k}, \left\{\frac{p_i}{p^k + q^k}\right\}_{i \in C^k}\right),
\]

where \( H(\pi_1, ..., \pi_I) = \sum_{i=1}^{I} \pi_i \log \frac{1}{\pi_i} \), which is the entropy of a probability distribution \((\pi_1, ..., \pi_I)\), \( q^k \) is the probability of the random walker exiting community \( k \) at one time step, \( p^k = \sum_{i \in C^k} p_i \), \( K \) is the number of communities, and \( q = \sum_{k=1}^{K} q^k \). Rosvall and Bergstrom (2008) call the objective function above the map equation. See Appendix A.1 for detail.

The first term of (2.3) describes the long-run behavior of the random walker across communities whereas the second term of (2.3) describes those inside each community. The key idea here is to reduce the amount of information 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 but we can discriminate between those streets because they belong to different administrative units. However, grouping cells into communities makes it necessary for us to assign codes to each community. This tradeoff yields the optimal community structure.

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 \( \ell \), and so forth. Our objective function is then constructed recursively as

\[
L^*(\{C^k\}_k) = qH\left(\frac{q^1}{q}, \ldots, \frac{q^k}{q}\right) + \sum_k L^*(\{C^{k\ell}\}_\ell), \tag{2.4}
\]

where

\[
L^*(\{C^{k\ell}\}_\ell) = (p^k + q^k)H\left(\frac{q^k}{p^k + q^k}, \left\{\frac{p_i}{p^k + q^k}\right\}_{i \in C^k}\right) + \sum_\ell L^*(\{C^{k\ell m}\}_m), \tag{2.5}
\]

and, at the lowest level of the hierarchy,

\[
L^*(C^{k\ell\cdots r}) = (p^{k\ell\cdots r} + q^{k\ell\cdots r})H\left(\frac{q^{k\ell\cdots r}}{p^{k\ell\cdots r} + q^{k\ell\cdots r}}, \left\{\frac{p_i}{p^{k\ell\cdots r} + q^{k\ell\cdots r}}\right\}_{i \in C^{k\ell\cdots r}}\right). \tag{2.6}
\]

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 human mobility 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 one-half 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. The data we use were collected for one year from January 1, 2012 to December 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 \( \{x_1, x_2, x_3, \ldots\} \), 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.\(^8\)

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.\(^9\) 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.

---

\(^8\)We refer to Horanont et al. (2013) for this method. In this study, this method is applied to the raw GPS data by NTT DOCOMO, INC because the access to the raw GPS data is limited to the data provider for protection of personal information.

\(^9\)The data are similar to those derived from the National Household Travel Survey in the United States. The data are publicly available online at [http://www.cbr.mlit.go.jp/kikaku/chukyo-pt/persontrip/p01.html](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.
4 Detected Communities

By using the commuting data developed above, we first construct the Markov chain defined by (2.1) and find its largest recurrent class. Of the 85,607 cells, the largest recurrent class has 80,926 cells. We then apply the map equation method to our data limited to the largest recurrent class. 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.

![Maps of communities](image_url)

(a) Level-1 communities  
(b) Level-2 communities

Figure 2: Maps of communities

---

10To carry out the map equation method, we use the code provided by Daniel Edler and Martin Rosvall at [http://www.mapequation.org/](http://www.mapequation.org/).
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 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 $\eta = 1$.

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. 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. 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. This distribution, first introduced by Reed (2002), is obtained as a normal variance mixture by the exponential distribution (see Appendix A.3).

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 $x$:

$$f_N(x) = \frac{1}{\sigma} \phi \left( \frac{x - \mu}{\sigma} \right),$$  \hspace{1cm} (5.1)

$$f_{DPLN}(x) = \frac{\alpha \beta}{\alpha + \beta} \phi \left( \frac{x - \nu}{\tau} \right) \left\{ m \left( \alpha \tau - \frac{x - \nu}{\tau} \right) + m \left( \beta \tau + \frac{x - \nu}{\tau} \right) \right\},$$  \hspace{1cm} (5.2)

$$f_{mixN}(x) = \theta \frac{1}{\sigma_1} \phi \left( \frac{x - \mu_1}{\sigma_1} \right) + (1 - \theta) \frac{1}{\sigma_2} \phi \left( \frac{x - \mu_2}{\sigma_2} \right),$$  \hspace{1cm} (5.3)

where $\phi$ 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 $\Phi$ 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 Appendix A.3 for the derivation of the density of the DPLN.

We fit the three distributions above to our data by using the maximum likelihood estimation.

---

11See, for example, Levy (2009), Eeckhout (2009), Giesen et al. (2010), Bee et al. (2013), Ioannides and Skouras (2013), and Giesen and Suedekum (2014).

12Eeckhout (2009) argues that truncating sample results in biased conclusions.

13Ioannides and Skouras (2013) also consider a combination of lognormal and Pareto distributions. They assume that there exists a switching population $S$ such that the distribution is lognormal for $S \leq S$ and is Pareto for $S \geq S$, 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.

14See also Reed and Jorgensen (2004).
Table 1 summarizes the estimation results. 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.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Normal + Normal</th>
<th>DPLN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>3.3472 (0.041)</td>
<td>2.7910 (0.033)</td>
<td></td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>1.8549 (0.029)</td>
<td>1.0635 (0.024)</td>
<td></td>
</tr>
<tr>
<td>$\mu_2$</td>
<td></td>
<td>7.0222 (0.280)</td>
<td></td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td></td>
<td>1.7743 (0.171)</td>
<td></td>
</tr>
<tr>
<td>$\theta$</td>
<td></td>
<td>0.8685 (0.014)</td>
<td>0.5634 (0.013)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td></td>
<td></td>
<td>8.6315 (0.013)</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td>1.6852 (0.006)</td>
</tr>
<tr>
<td>$\nu$</td>
<td></td>
<td></td>
<td>0.5843 (0.002)</td>
</tr>
<tr>
<td>$\tau$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-4,171</td>
<td>-3,778</td>
<td>-3,799</td>
</tr>
<tr>
<td>AIC</td>
<td>8,346</td>
<td>7,566</td>
<td>7,606</td>
</tr>
<tr>
<td>BIC</td>
<td>8,357</td>
<td>7,594</td>
<td>7,628</td>
</tr>
</tbody>
</table>

Standard errors are in parentheses.

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 $\lambda$ and $1 + Y_i$ is iid lognormally distributed with mean $\mu$ and $\delta^2$. $Q$ is called a compound Poisson process. Then, we assume
that $S(t)$ follows the following process:

$$dS(t) = \gamma S(t)dt + \xi S(t)dW(t) + S(t-)dQ(t), \quad (5.4)$$

where $W$ is a Wiener process and $S(t-) = \lim_{s \to t} S(s)$. This is called a jump diffusion process (see, e.g., Shreve, 2004, ch. 11). Given $S(0) = s_0$, the solution to (5.4) is given by the Doleans-Dade exponential:

$$S(t) = s_0 \exp \left\{ \left( \gamma - \frac{1}{2} \xi^2 \right) t + \xi W(t) \right\} \prod_{i=1}^{N(t)} (1 + Y_i). \quad (5.5)$$

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,

$$X(t) = X(0) + \left( \gamma - \frac{1}{2} \xi^2 \right) t + \xi W(t) + \sum_{i=1}^{N(t)} \log(1 + Y_i). \quad (5.6)$$

By discretizing the time with interval $\Delta$, we obtain

$$X(t) = X(t - \Delta) + \left( \gamma - \frac{1}{2} \xi^2 \right) \Delta + \xi (W(t) - W(t - \Delta)) + \sum_{i=N(t-\Delta)+1}^{N(t)} \log(1 + Y_i). \quad (5.7)$$

We consider the probability transition density $p(x; \Delta|y, k)$ that satisfies\(^\text{15}\)

$$\Pr \left[ X(t) \in [x, x + dx] \mid X(t - \Delta) = y, N(t) - N(t - \Delta) = k \right] = p(x; \Delta|y, k)dx. \quad (5.8)$$

By assumption, $W(t) - W(t - \Delta) \sim \mathcal{N}(0, \Delta)$ and $\log(1 + Y_i) \sim \mathcal{N}(\mu, \sigma^2)$. Then, from (5.7),

$$p(x; \Delta|y, k) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ \frac{(x - y - \left( \gamma - \frac{1}{2} \xi^2 \right) \Delta - k\mu)^2}{2 \sigma^2} \right\}. \quad (5.9)$$

Furthermore, because $X(t - \Delta)$ and $N(t) - N(t - \Delta)$ are independent, the variable representing

\(^{15}\)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.
the number of jumps is integrated out from the probability transition density as

\[
p(x; \Delta|y) = \frac{p(x, y; \Delta)}{p(y; \Delta)} = \sum_{k=0}^{\infty} \frac{p(x, y, k; \Delta)}{p(y; \Delta)} \Pr[N(t) - N(t - \Delta) = k] \tag{5.10}
\]

\[
= \sum_{k=0}^{\infty} \frac{p(x, y, k; \Delta)}{p(y, k; \Delta)} \Pr[N(t) - N(t - \Delta) = k] \tag{5.11}
\]

\[
= \sum_{k=0}^{\infty} p(x; \Delta|y, k) \Pr[N(t) - N(t - \Delta) = k]. \tag{5.12}
\]

Because \(N\) is a Poisson process with rate \(\lambda\), 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 \(\Delta\). Then, we have

\[
p(x; \Delta|y) \approx \sum_{z \in \{0, 1\}} p(x; \Delta|y, z) \Pr(Z = z) \tag{5.13}
\]

\[
= \lambda \Delta p(x; \Delta|y, 1) + (1 - \lambda \Delta) p(x; \Delta|y, 0). \tag{5.14}
\]

Therefore,

\[
X(t)|X(t - \Delta) = y \sim \lambda \Delta \times \mathcal{N} \left( y + \left( \gamma - \frac{1}{2} \xi^2 \right) \Delta + \mu, \xi^2 \Delta + \delta^2 \right) + (1 - \lambda \Delta) \times \mathcal{N} \left( y + \left( \gamma - \frac{1}{2} \xi^2 \right) \Delta, \xi^2 \Delta \right). \tag{5.15}
\]

Note that our estimation result is consistent with the regularity conditions behind (5.15). First, for the approximation through the Bernoulli distribution to be accurate, \(\Delta\) 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 \(\theta\) 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 \(\mu_2\) and \(\sigma_2\) are larger than those of \(\mu_1\) and \(\sigma_1\), respectively, this condition is also met. Therefore, we may conclude that the above model is relevant to our data.\(^{16}\)

\(^{16}\)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 mobile phone 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.

In this study, we used only data of human mobility. However, we could also exploit the structure of spatial economy such as transport network. Moreover, 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. In particular, 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 framework here.

Appendix

A.1 The map equation

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 which is specified below. Our objective is to code these as effectively as possible. To make this more precise, let \( \ell_j \) 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:

\[
\sum_j \ell_j p_j,
\]

(A.1)

where \( p_j \) is the long-run probability of state \( j \).

We code by dividing the long-run behavior of the random walker into those inside each community and those across communities. We describe the long-run behavior of the random

---

\(^{17}\)Berliant and Watanabe (2018) consider a spatial economic model with transport network where agents choose their consumptions and locations. They find that the equilibrium city size distribution fits the data well when the degree distribution of transport network follows the power law.
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

$$L^k = \ell^k_e \frac{q^k}{p^k + q^k} + \sum_{i \in C^k} \ell_i \frac{p_i}{p^k + q^k},$$  \hspace{1cm} (A.2)$$

where $\ell_i$ is the length of the binary code assigned to the state of visiting cell $i$, $\ell^k_e$ 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 (2.1), the probability of exiting community $k$ is

$$q^k = \sum_{i \in C^k} \sum_{j \notin C^k} p_i P_{ij} = \sum_{i \in C^k} \sum_{j \notin C^k} p_i \frac{N_{ij}}{N_i}. \hspace{1cm} (A.3)$$

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

$$L = \sum_{k=1}^{K} \frac{\ell^k q^k}{q}, \hspace{1cm} (A.4)$$

where $\ell^k$ 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

$$qL + \sum_{k=1}^{K} (p^k + q^k)L^k. \hspace{1cm} (A.5)$$

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

$$H(\pi_1, ..., \pi_I) = \sum_{i=1}^{I} \pi_i \log \frac{1}{\pi_i}. \hspace{1cm} (A.6)$$

Then, the theorem states that it is possible to make the average code length arbitrarily close
to the entropy of the underlying probability distribution.\footnote{This is a fundamental theorem in information theory. See, e.g., Theorem 5.4.2 of Cover and Thomas (2012).}

In our coding problem, the entropies of the probability distributions in $L^k$ and $L$ are

$$H\left(\frac{q^k}{p^k+q^k}, \{\frac{p_i^k}{p^k+q^k}\}_{i \in C^k}\right) \text{ and } H\left(\frac{q_1^k}{q^k}, \ldots, \frac{q^K}{q^k}\right),$$

respectively. Therefore, our task reduces to finding the community partition that minimizes

$$L^*(C^1, \ldots, C^K) = \sum_{k=1}^{K} (p^k + q^k)H\left(\frac{q^k}{p^k+q^k}, \{\frac{p_i^k}{p^k+q^k}\}_{i \in C^k}\right). \quad (A.7)$$

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

### A.2 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 \footnote{Farmer and Fotheringham (2011) use the modularity method of \footnote{This is a fundamental theorem in information theory. See, e.g., Theorem 5.4.2 of Cover and Thomas (2012).}} 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 $N \times \frac{N_i}{N} \times \frac{N_j}{N} = \frac{N_i N_j}{N^2}$. The modularity is then defined as

$$Q = \frac{1}{N} \sum_{i,j} \left( N_{ij} - \frac{N_i N_j}{N} \right) \delta_{ij}, \quad (A.8)$$

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 $\frac{N_i N_j}{N^2}$ 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.
A.3 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:

$$dS(t) = \gamma S(t)dt + \xi S(t)dW(t),$$  
(A.9)

where $W$ is a Wiener process. Given $S(0) = s_0$, it follows from the standard Itô calculus that the solution to (A.9) is

$$S(t) = s_0 e^{(\gamma - \frac{1}{2} \xi^2) t + \xi W(t)}.$$  
(A.10)

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

$$X(t) = X(0) + \left( \gamma - \frac{1}{2} \xi^2 \right) t + \xi W(t).$$  
(A.11)

Suppose that $S(0)$ is lognormally distributed with mean $\nu$ and variance $\tau^2$ (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

$$X(t) \sim \mathcal{N} \left[ \nu + \left( \gamma - \frac{1}{2} \xi^2 \right) t, \tau^2 + \xi^2 t \right].$$  
(A.12)

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

$$f_X(x) = \int \frac{1}{\sigma_t} \phi \left( \frac{x - \mu_t}{\sigma_t} \right) f_T(t)dt,$$  
(A.13)

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

$$f_X(x) = \frac{\alpha \beta}{\alpha + \beta} \phi \left( \frac{x - \nu}{\tau} \right) \left\{ m \left( \alpha \tau - \frac{x - \nu}{\tau} \right) + m \left( \beta \tau + \frac{x - \nu}{\tau} \right) \right\},$$  
(A.14)

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

$$\frac{\xi^2}{2} z^2 + \left( \gamma - \frac{1}{2} \xi^2 \right) z - \lambda = 0.$$  
(A.15)
References


