Exploratory potential analysis of Mapillary for streetscape monitoring in Japan

Xinrui ZHENG and Mamoru AMEMIYA

Abstract: Mapillary is the world's largest crowdsourced-based open landscape photo-sharing service, launched in Sweden in 2014. By July 2022, more than 50 million images throughout Japan have been contributed by Mapillary users. In this study, we conducted an exploratory analysis of data from Mapillary in Tokyo from the perspective of road coverage and attempted to get a deep insight to the contribution behavior pattern by examining the effect of reginal characteristics.

Keywords: crowdsourcing, Volunteered Street View Imagery, Mapillary, completeness

1. Introduction

The surge of SVI (Street View Imagery) as an important data source for urban analytics has been catalyzed by the proliferation of imagery platforms, advances in computer vision, machine learning, and availability of computing resources (Biljecki and Ito, 2021). However, the limitations of the mainstream data (e.g., Google Street View, Baidu Total View) provided by government agencies or private companies are pointed out in data application such as barriers set up to free data download, history data fetching, and free data usage (d'Andrimont et al., 2018; Inoue et al., 2022; Google, 2022). These restrictions present issues when trying to conduct the study at a large scale or temporal analysis.

The emergence of the Web2.0 era has fostered the potential of individuals to contribute and access information through multiple resources which also facilitated the collection of massive Volunteered Street View Imagery (VSVI) at a large scale and in a short time (Goodchild, 2007; Zhen et al., 2014). The data application in the field of research has been promoted by the free data usage but also restricted by the incompleteness of data in some regions (Ma et al., 2019). To grasp the VSVI data distribution pattern, this study conducted a completeness evaluation analysis in Tokyo and examined the effect of regional characteristics to further understand the mechanism of the crowdsourced data contribution behaviors by citizens. Specifically, we explored the imagery data on the Mapillary platform (launched in 2014) which is the first website to share geotagged photos contributed by the crowd. Along with the exponential growth of data, until now, more than 1.5 billion street-level images from over 190 counties are available on this platform (https://www.mapillary.com/). To note, Mapillary provides images under CC BY-SA 4.0 license, which means that it is free to use the data even for commercial purposes.

2. Literature Review

Previous research concludes the generally lower spatial coverage of Mapillary than GSV, with the exception of data along the pedestrian/cycle paths and

Xinrui Zheng

Doctoral Program in Policy and Planning Sciences, University of Tsukuba

s2230131@s.tsukuba.ac.jp

indicates the large difference in data coverage across different regions (Juhasz and Hochmair, 2015; Juhasz and Hochmair, 2016; Mahabia et al., 2020). The detailed data distribution in Japan remains unknown and more research is needed. As for the effect of regional characteristics, population density has been correlated with VGI (Volunteered Geographic Inforamtion) data contribution (Mullen et al., 2015), and the similar trend is also found in VSVI data contribution owing to the more potential users and road segments (Mahabie et al., 2020). However, the moderate correlations with population density (0.33-0.52, p<0.01) also imply the existence of other factors (Mahabie et al., 2020).

3. Research Method

3.1. The Overview of the Analysis

The study analyzed Mapillary data contributed since the inception of the Mapillary project in 2014 in Tokyo. There are 23 special wards and 39 municipalities (26 cities, 5 towns, and 8 villages) in Tokyo which is divided into 23-ku area, Tama-area, and islands. We conducted the analysis at the municipality level to be 3.5 km^2 in the average in Tokyo. The first part of the analysis is the Mapillary data completeness evaluation to compute road coverage for each municipality unit and the second part focuses on correlations between regional characteristics and data distribution.

3.2. Data description and extraction

3.2.1 Reference data for determining completeness

The center line data of railways and other roads (including highways, national roads, prefectural roads, municipal roads, and pedestrian paths) in 2021 published by the Geospatial Information Authority of Japan are used as reference datasets to evaluate the completeness of Mapillary data. To facilitate detailed analysis based on road segments, the original road data were split at intersections. The reference road data of Tokyo were extracted and uploaded into ArcGIS.

3.2.2 Mapillary data

Imagery point data downloaded via API provided by the Mapillary platform were used in this study for the completeness analysis. To be specific, bounding box covering the whole area of Tokyo was created to download complete imagery data contributed until the retrieving date (July 18, 2022). GeoJSON files containing the coordinates of image nodes, the id of sequence the image belongs to, the image id, and timestamp were then retrieved, and transferred into point shapefiles in ArcGIS software for further analysis.

3.2.3 Reginal characteristics data

We assumed that data distribution pattern at the level of municipality would be affected by the features of population density, landscape attractiveness, and the city size. Firstly, unlike other types of VGI data (such as OSM which can be contributed by volunteers from anywhere), VSVI data collection requires contributors to be in site to take photos and higher population density tends to provide more potential contributors. Both the residential density and the daytime population density extracted from the 2020 Population Census data collected by the Statistics Bureau of Japan were included for analysis.

In addition, landscape attractiveness was assumed to encourage photograph behavior since Mapillary was found more complete in park or along river (Juhasz & Hochmair, 2016). To do that, the variables of green space ratio and blue space ratio were calculated based on the land cover map from JAXA (2018-2020) for analysis. The sealine ratio to the length of district boundary was calculated using the sealine data in 2006 obtained from the digital national land information. In addition, the tourist spots density was based on the tourism resource data in 2014 from the national digital land information. The data sources of the tourism resources are the list of tourist resources published by Japan Travel Bureau and the list of tourist spots from Japan Tourism Agency.

Finally, we also included the variables of the total length of roads and the total area calculated with ArcGIS to reflect the city size considering the ease for a small area to reach high coverage. For example, a study of Mapillary completeness evaluation over the world found that the highest completeness value was in a small country of Barbados where mapping 11 km of main roads already leads to a completeness value of over 50% (Juhasz & Hochmair, 2016).

3.3. Completeness evaluation

Completeness evaluation is to reflect the degree of data completeness using the ratio of reference road segments with overlapping Mapillary data. Since the Mapillary points and the road center line are usually not coincident owning to the road width, this study applied the tool of closest facility in the network analysis to find the closest road segments for each point data. Furthermore, to avoid including points that are not taken along roads, the search tolerance (the maximum search distance to locate points on the road network) was set to 50 meters taking into account road width and GPS drift. After that, the total number of imagery points located was calculated for each road segment via Python and the data density (number of points per meter d_j) for each road segment (j) is calculated by dividing the number of imagery points (n_j) with the length of road segment (l_j) as shown in equation (1).

$$d_j = \frac{n_j}{l_j} \tag{1}$$

Roads with density more than 1 point per 50m were selected since the largest interval distance for automatic capture is 50 meters. Finally, completeness values (D_i) for the area i were defined as the ratio of the length of roads containing data (L_{road_data}) to the total length of roads (L_{total}) as equation (2).

$$D_i = \frac{L_{road\_data}}{L_{total}} \tag{2}$$

4. Results

4.1. Completeness evaluation

Totally 5,630,683 image points in Tokyo were downloaded and 5,604,853 were located on the nearest road segments after removing points far from road centerlines. Among 761,146 road segments, 184,793 (24.28%) of them was found containing at least one Mapillary imagery point. After removing segments with low point density (less than 1 point per 50m), 177,465 (23.32%) segments were defined as roads containing Mapillary data.

Figure 1 shows data completeness among 62 municipalities in Tokyo. The mean value is 28.93% (SD: 19.38%) and ranges between 0.00% and 80.83%. Both the highest and the lowest values are in two villages on islands. These are Aogashima-mura (8.75 km^2) with a road coverage of 80.08% and Mikurajima-mura (20.54 km^2) where no data exist. As for data distribution in the mainland portion of Tokyo, the completeness values are generally high in the 23-ward area and low in the Tama-area. Among the 23 wards of Tokyo, the highest coverage values are in wards near the center which are Chiyoda-ku (77.76%), Chuo-ku (74.30%), Minato-ku (68.83%), and Shibuya-ku (63.22%). The values decrease as farther away from the city center and relatively low in the northeast. In the Tama-area, units closer to Tokyo center show higher values and most of them are lower than 40% while the highest value of 59.39% is in Kunitachi-shi (8.15km^2).

4.1. Regional characteristics

Table 1 presents the eight variables of regional characteristics assumed to be related to data contribution. There is a large difference in population density (e.g., daytime population density ranges from 0.0 to 78.5) and city size (e.g., total area value ranges from 4.1 to 225.5 km^2) among the 62 municipalities in Tokyo. The values of sealine ratio are extremely high for 9 islands in Tokyo and low in the mainland portion. Low-urbanization areas such as towns and villages tend to be covered by vegetation and show higher green space ratio, while the blue space ratio is generally low in Tokyo ranging from 0.0% to 10.1%. In terms of the tourist spot density, the highest values are in some special wards such as Taito-ku (1.09 spots per km^2) and Chiyoda-ku (0.87 spots per km^2).

Table 2 shows the Pearson's correlation coefficients between completeness values and regional variables. Among the 8 dependent variables, only 5 show significant correlations. Both the residential density and the daytime population density are significantly correlated with coverage ratio, while the latter one has a stronger correlation of 0.687 ($p<0.001$). Green space ratio is negatively related to completeness level and no significant correlation is found with blue space ratio and the sealine ratio. Furthermore, in line with the assumption, tourist point density shows a strong correlation with coverage ratio. As for the city size, the total area shows a negative correlation while the road length represents no significant correlation.

5. Discussion

As a preliminary study for Mapillary data analysis across Japan, this study conducted completeness evaluation based on the value of road coverage using the closest facility tool in the network analysis with ArcGIS

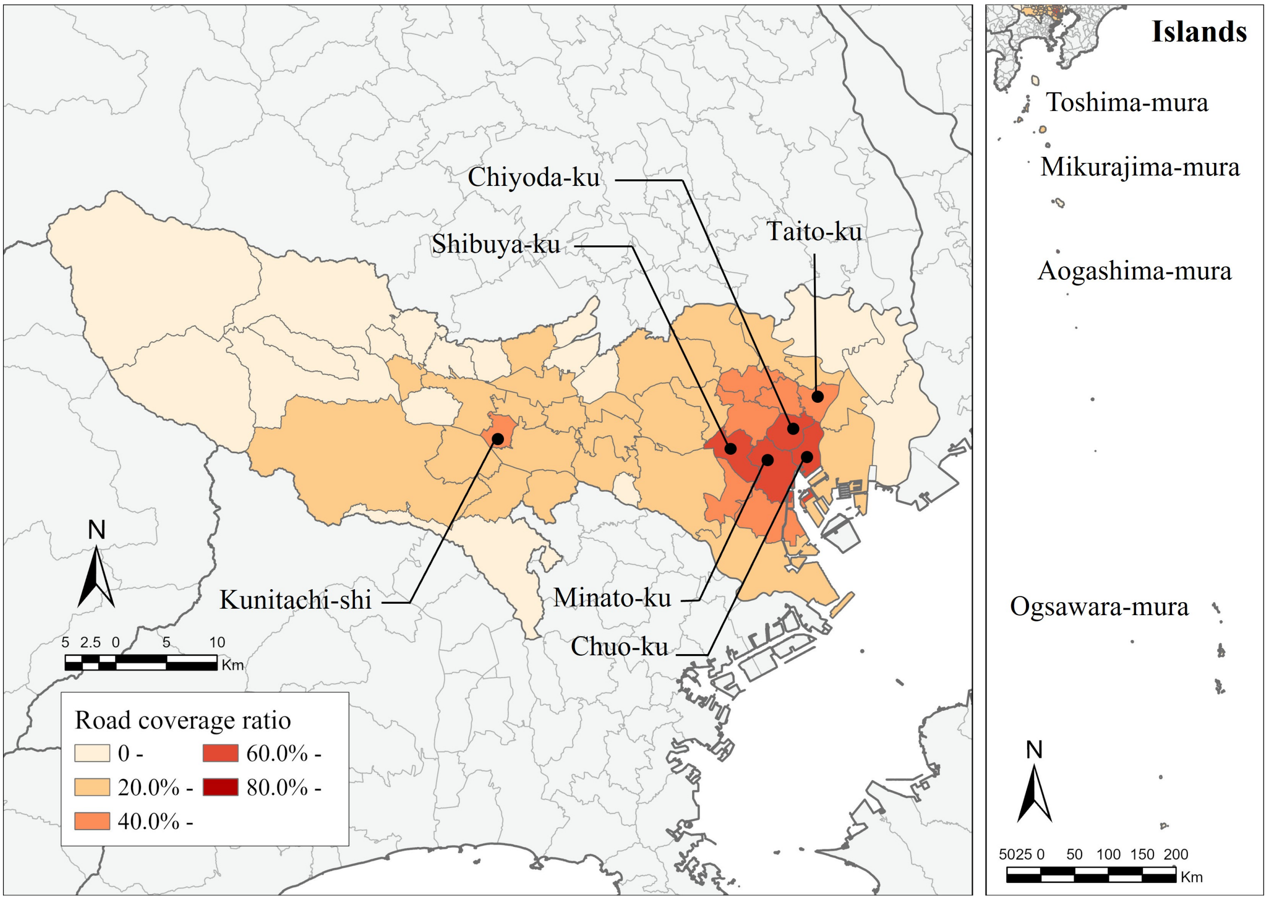

Figure 1. Road coverage ratio in Tokyo

Pro which help to solve the problem of data duplication in previous methods (Mahabia et al., 2020; Seto and Nishimura, 2022). The correlations with regional features were also examined to explore the mechanism of contribution pattern. Tokyo is the capital of Japan with the highest population density and is also one of the hotspots of Mapillary activity. Results based on data from 2014 to July 2022 show that totally 24.28% roads in Tokyo contain available imagery data which is higher than the main road coverage value of Japan (less than 4.65%) at the country level in 2016 represented in a study by Juhasz and Hochmair. Even though the highest coverage ratio reaches 80.0%, the large difference between different municipalities also indicates the limitation of data usage.

To understand the factors of contribution behavior, we discussed the correlations with the coverage ratio of 8 regional feature variables. As we assumed, the completeness of the volunteered data is related to the number of people who are also the potential sensors in this area. Moreover, the daytime population density shows stronger correlation than the residential density which may imply the effect of human activity pattern on data distribution. Taking street photos

Table 1. Regional characteristics

	Variables	Mean	SD	Range
population density	1. Residential density (1000 people /km^2)	9.6	7.1	0.0- 23.2
	2. Daytime pop density (1000 people /km^2)	12.7	15.2	0.0-78.5
Landscape attractiveness	3. Green space ratio (%)	34.3	31.5	2.8 - 99.3
	4. Blue space ratio (%)	1.45	2.56	0.0 - 10.1
	5. Sealine ratio (%)	14.5	28.5	0.0 - 98.7
	6. Tourist spots density (/km^2)	0.1	0.2	0.0 - 1.1
City size	7. Total Area (km^2)	3.5	40.7	4.1 - 225.5
	8. Total length of roads (km)	655.8	490.7	24.8-3254.3

Table 2. Correlations between coverage ratio and regional characteristics

Variables	Road coverage	1	2	3	4	5	6	7	8
Dependent variable									
Road coverage	1.00								
Independent variables									
1. Residential density	.350**	1.00							
2. Daytime pop density	.687**	.589**	1.00						
3. Green space ratio	-.315*	-.893**	-.593**	1.00					
4. Blue space ratio	.122	.289*	.342**	-.373**	1.00				
5. Sealine ratio	.128	-.475**	-.268*	.570**	-.093	1.00			
6. Tourist spots density	.652**	.235	.722**	-.215	.247	.018	1.00		
7. Total Area	-.325**	-.392**	-.305*	.468**	-.080	.112	-.189	1.00	
8. Total length of roads	-.143	.157	-.034	-.169	.137	-.224	-.167	.521**	1.00

* $p < 0.05$, ** $p < 0.01$

requires users to shoot while on the move. The higher correlation with daytime population density may be explained by more travel and contribution behavior at day (Mahabia et al., 2020). In addition, the coverage ratio is higher in areas dominated by commercial and business use than residential area. This may be explained by the fact that collecting photos is often not the purpose for traveling and the shooting path tend to be consistent with the daily route (from home to destinations).

Among the four landscape-related variables, only two show significant correlations. Contrary to expectation, green space ratio is negatively related to coverage ratio which is probably due to the high negative correlation with population density. In other words, a high green space ratio is usually accompanied by a lower level of urbanization. Furthermore, the blue space ratio shows no significant correlation which may be attributed to the low difference between different areas and so as to sealine ratio which is zero for 47 units. The tourists point density is highly correlated with the independent variable which may imply the impact of activity pattern on data distribution. On the one hand, people may be more willing to share photos on tour. On the other hand, unlike the daily fixed routes which are often the closest path to destinations, the trajectories of tourists are more inclined to cover the whole area thus causing a high coverage rate.

Finally, the variables of city area and the total road length are included considering the effect of the difficulty to cover an area. Results show that even though there is a significant negative correlation between coverage ratio and the values of total area as expected, the road length shows no correlation. This might be explained by the fact that people tend to determine their activity zone in terms of straight-line distances rather than the road length. The examples of small area with high coverage are Kunitachi-shi ($8.1km^2$) where the data is found to be contributed mostly by one power user ("esophagus") during April 2021 (Figure2). Another power user named "mura" is found to has been contributing large amount of island and seashore pictures (Figure 2), who has covered more than 80.8% roads in Aogashima-mura ($5.96\ km^2$) in three days from Feb. 9 to Feb. 11 in 2018, and more than 60% roads in Toshima-mura ($4.12\ km^2$) only within one day. The contribution pattern of power users is more related to individual factors than regional factors.

5. Conclusion

This study conducted a completeness analysis using the road coverage ratio in Tokyo and explore the related regional characteristics. There is a generally high coverage ratio in Tokyo while the values range widely across different municipalities. By examining the effects of regional features, the data distribution within an area is found to be related to not only the number of people but also the dominant activity pattern in this area. As a result, there is higher completeness of data in residential areas than in visited areas and in destinations for tour than those for daily activities.

Therefore, it can be concluded that it is difficult to achieve a uniform distribution among different regions like GSV only depending on the free activities of users. Popularity increasing or user base expanding may be more to the data volume increase in existing data

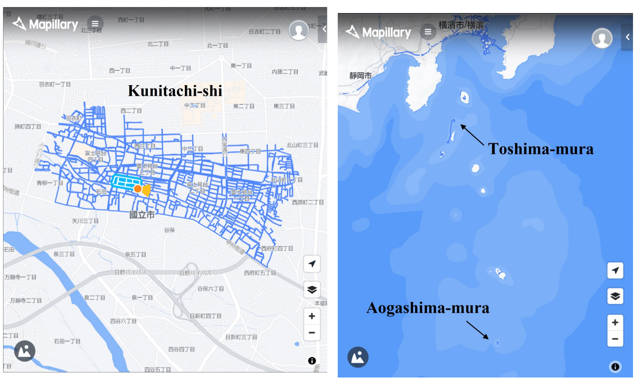

Figure 2. Mapped roads by the power users of "esophagus" (left) and "mura" (right)

locations. In August 2022, Mapillary launched a challenge in Japan to encourage users to contribute data in specified meshes. Only 8 users have completed more than one mission until 30th August which may be due to the difficulty (it is required to cover every road and path within a mesh). Mission challenges with lower extra time should take during daily life or with more fun like a game should be considered to improve data coverage.

References

Biljecki, F. and Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, **215**, 104217.

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, **69**(4), 211-221.

Google: Geoguideline (n.d.). Retrieved April 20, 2022 from https://www.google.com/intl/ja_ALL/permissions/geoguidelines/

Inoue, T., Manabe, R., Murayama, A. and Koizumi, H., 2022. Landscape value in urban neighborhoods: A pilot analysis using street-level images. *Landscape and Urban Planning*, **221**, 104357.

Juhasz, L. and Hochmair, H., 2015. Exploratory completeness analysis of Mapillary for selected cities in Germany and Austria. *Gi_Forum J. Geogr. Inf. Sci*, 535-545.

Juhász, L. and Hochmair, H. H., 2016. User contribution patterns and completeness evaluation of Mapillary, a crowdsourced street level photo service. *Transactions in GIS*, **20**(6), 925-947.

Ma, D., Fan, H., Li, W. and Ding, X., 2019. The state of mapillary: An exploratory analysis. ISPRS International *Journal of Geo-Information*, **9**(1), 10.

Mahabir, R., Schuchard, R., Crooks, A., Croitoru, A. and Stefanidis, A., 2020. Crowdsourcing street view imagery: a comparison of mapillary and OpenStreetCam. ISPRS *International Journal of Geo-Information*, **9**(6), 341.

Seto, T. and Nishimura, Y., 2022. Analysis of the spatio-temporal accumulation process of Mapillary and OSM road data: A case study in Japan. The International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences*, IV/4, III/10, IV/7, V/1 & ICWG IV/III. (in press)

Zhen, F., Wang, B. and Wei, Z., 2015. The rise of the internet city in China: Production and consumption of internet information. *Urban Studies*, **52**(13), 2313-2329.

d'Andrimont, R., Yordanov, M., Lemoine, G., Yoong, J., Nikel, K. and Van der Velde, M., 2018. Crowdsourced street-level imagery as a potential source of in-situ data for crop monitoring. *Land*, **7**(4), 127.