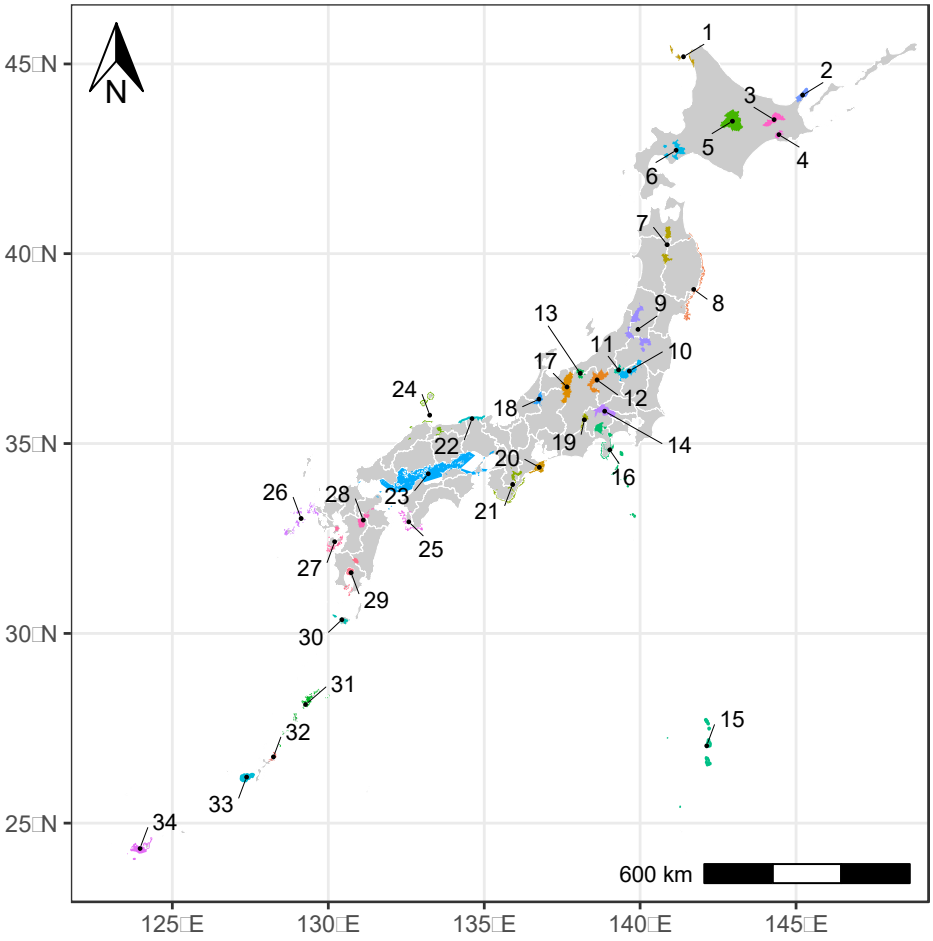
Fig. S1: Map of the 34 national parks in Japan.



1. Rishiri-Rebun-Sarobetsu
2. Shiretoko
3. Akan-Mashu
4. Kushiroshitsugen
5. Daisetsuzan
6. Shikotsu-Toya
7. Towada-Hachimantai
8. Sanriku Fukko
9. Bandai-Asahi
10. Nikko
11. Oze
12. Joshin'etsukogen
13. Myoko-Togakushi renzan
14. Chichibu-Tama-Kai
15. Ogasawara
16. Fuji-Hakone-Izu
17. Chubusangaku
18. Hakusan
19. Minami Alps
20. Ise-Shima
21. Yoshino-Kumano
22. San'inkaigan
23. Setonaikai
24. Daisen-Oki
25. Ashizuri-Uwakai
26. Saikai
27. Unzen-Amakusa
28. Aso-Kuju
29. Kirishima-Kinkowan
30. Yakushima
31. Amamigunto
32. Yambaru
33. Keramashoto
34. Iriomote-Ishigaki

Table S1: Values of the source field of tweets regarded as originating from official Twitter clients and Foursquare apps.

Client type	Source field
Official client	Twitter for iPhone
	Twitter for Android
	Twitter Web Client
	web
	Twitter for iPad
	TweetDeck
	Twitter for Android Tablets
	Twitter for Windows
	Twitter for iAppli
	Twitter for BlackBerry®
	Twitter for Windows Phone
	Twitter for BlackBerry
Foursquare	Foursquare
	Foursquare Swarm

Appendix 1: Development of custom client applications capturing Twitter data

The Twitter data used in this study was obtained by developing small custom client applications. To construct Dataset 1, we developed a small client application which captures target tweets using the 'tweepy' module in the Python programming language. The application sampled tweets using the Streaming API of Twitter (statuses/sample endpoint; the archived documentation can be read at https://web.archive.org/web/20220316014310/https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction) by which we could obtain roughly 1% of tweets in real time. We assumed tweets having Japanese characters were posted by Japanese users and tweets which do not have Japanese characters were discarded using the regular expression ([一-龠]+|[ぁ-ん]+|[ァ-ヴー]+), which matches all three kinds of Japanese characters (i.e. Kanji [一-龠], Hiragana [ぁ-ん], and Katakana [ァ-ヴー]). Due to limited storage capacity, this filtering was done in real time and only tweets having Japanese characters were saved. Captured tweet data were stored in text files using JSON (JavaScript Object Notation) format. We adopted JSON text files because it allowed us to manually inspect and fix data corruption if necessary. Also, to reduce the probability of data corruption, the tweets were saved into new files every hour. To maximize continuity of capturing, the application was programmed to automatically restart immediately after occasional crashes of the application or the computer. When a computer restart was needed, we ran the application in a backup system and continued capturing. After obtaining tweet data, we checked and corrected corruption of JSON data caused by occasional network errors and crashes. The data download was conducted from January 1st, 2017, to December 31st, 2022. Note that despite our efforts to maintain continuous capturing, the system experienced downtimes (Table S2). The longest downtime experienced during October to December 2022 was caused by a migration problem for the API update.

To obtain tweet data for Dataset 2, another application was developed. Originally, we planned to extract georeferenced tweets from Dataset 1. However, after we started capturing Dataset 1, we

34 noticed that the method could not capture enough georeferenced tweets. Therefore, we developed
35 another application which collects georeferenced tweets with location information in Japan by
36 specifying the bounding box of Japan (122.9336, 20.4253, 153.9864, 45.5572) for the Streaming
37 API (statuses/filter endpoint). This application had similar functionality and operated similarly as the
38 one used for Dataset 1, but the data was stored in new files every 10 minutes. The data collection
39 was conducted from March 19th, 2017 to December 31st, 2022, except during short downtimes
40 (Table S2).
41

Table S2: Duration of downtime experienced by the applications capturing tweet data. Because we did not measure actual duration, the durations of downtimes were estimated from the number of files which we expected to be stored every hour or every 10 minutes for Dataset 1 and Dataset 2, respectively.

Dataset	Year-month	Downtime (hours for Dataset 1 and mins for Dataset 2)
Dataset 1	2018-04	1
	2021-11	17
	2022-01	20
	2022-11	183
	2022-12	970
Dataset 2	2017-05	10
	2017-08	20
	2017-10	100
	2018-03	60
	2018-04	10
	2018-07	30
	2018-09	90
	2019-07	40
	2019-10	20
	2020-02	30
	2020-05	20
	2020-06	100
	2020-10	80
	2021-11	1100
	2022-01	70
	2022-05	10
	2022-07	50
	2022-08	40
	2022-10	40
	2022-11	10
	2022-12	10

Appendix 2: Performance evaluation for the estimation of demographic attributes.

To evaluate the reliability of the estimation of demographic attributes, we compared the estimation results with two existing sets of statistics on demographic attributes of Twitter users. The first set of statistics was the results of the Survey on Information and Communication Media Usage Time and Information Behavior conducted by the Institute for Information and Communications Policy, Ministry of Internal Affairs and Communications, Japan (https://www.soumu.go.jp/iicp/research/results/media_usage-time.html, accessed on July 14th, 2023; hereafter, MIC survey). This survey used questionnaires to ask about information and media related activities such as time spent on television, internet, social media, smartphones, etc. It has been conducted every year since 2013 and targeted 1,500 people aged 13-69 in each survey. The respondents of the survey were selected so that the percentages of each sex and age reflect the demographic composition of Japan. We obtained the statistics about demographic attributes (age and sex) and activities (posting and browsing) of Twitter users from 2017 to 2022 and used them for verification (the numbers of respondents for each demographic group are shown in Fig. S2). The second set of statistics used for verification was from the Twitter Usage survey conducted by MyVoice Communications, Inc. (https://myel.myvoice.jp/products/detail.php?product_id=26908, accessed on July 14th, 2023, hereafter MyVoice survey). This survey was conducted every year from 2009 to 2020 and contains age and sex as well as resident prefecture, which is not available from the MIC survey, and activities (posting and browsing). We used the survey results from 2017-2020 for this study (the numbers of respondents for each demographic group are shown in Fig. S3).

The verification was conducted as follows. First, we calculated the proportion of users who were posting tweets for each demographic group and year using data from the two surveys. The MIC survey separately asked respondents whether they post tweets from smartphones and/or computers. Therefore, we calculated two values of the posting user ratio for each demographic group and year:

1) maximum posting user ratio which assumes users who posted tweets using smartphones did not post tweets using computers, that is, the maximum posting user ratio = (number of respondents posting tweets from smartphones + number of respondents posting tweets from computers) / number of respondents; 2) minimum posting user ratio which assumes maximum overlap between those who post on smartphones and computers, that is, the minimum posting user ratio = the larger of the number of respondents posting tweets from smartphones or computers / number of respondents (Fig. S4). For the MyVoice survey, the posting user ratio for each group and year was calculated as the number of respondents who answered that they use Twitter and post tweets (Fig. S5). Next the posting user ratios of each demographic group, year, and survey were multiplied by the population number of the group obtained from the national census conducted in 2020 by the Statistics Bureau, Ministry of Internal Affairs and Communications (https://www.stat.go.jp/data/kokusei/2020/index.html, accessed on July 14th, 2023) and the number of tweet-posting users for each demographic group was estimated. Then, using the estimated number of users, we estimated the proportion of each demographic group in the tweet-posting users.

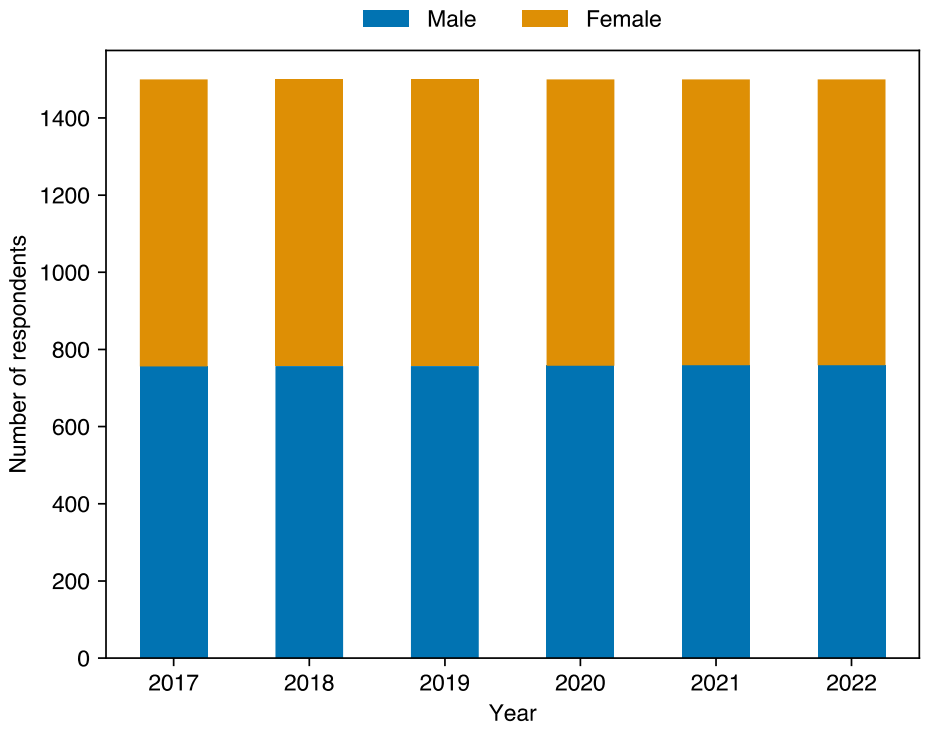
The results were compared with the proportion of each demographic group calculated from the estimation results for Dataset 1 because the demographic attributes of the randomly sampled Twitter user would be comparable to the existing statistics. For the verification of age, only the results of the MIC survey were used as the reference because a very small number of respondents in their 10's in the MyVoice survey cause inter-annual fluctuation of the posting ratio of this age group which causes fluctuation of the estimated proportion of Twitter users for each age group. For age and sex, congruency of the estimations based on Twitter to estimations based on other existing statistics was evaluated by calculating Cramér's V (Cramér, 1946) between the estimations based on Twitter and each statistic, for each year. Cramér's V measures association between variables on rows and columns and ranges from 0 to 1 where 0 means no association (i.e., proportions of groups on

97 rows are not different between estimation methods on columns) and 1 means strong association (i.e.,
98 proportions of groups on rows are different between estimation methods on columns). We calculated
99 Cramér's V by creating 2 × n contingency tables containing the estimated number of
100 users/population in each class in a demographic attribute (n = number of classes in the attribute)
101 estimated by Twitter and an existing statistic. Note that when calculating Cramér's V, the estimated
102 Japanese population who post tweets was rounded into nearest integer. For the estimation of
103 prefecture, Pearson's correlation coefficients were calculated between the estimated proportions of
104 users in each prefecture based on Twitter and those based on other statistics.

105 The estimated proportions of each sex based on Twitter data were congruent with the
106 estimations based on the existing statistics (Fig. S6a): for all the estimation methods, male users
107 were slightly dominant. Also, all estimated Cramér's V were <0.02, indicating the differences were
108 very small (Table S3). On the other hand, although the estimated age showed that Twitter users are
109 dominated by younger age groups, the estimation based on the Twitter data showed a stronger
110 dominance compared to existing statistics (Fig. S6b): the estimation based on the Twitter data
111 showed that Twitter is mostly dominated by users less than 50 years old whereas the MIC survey
112 showed 10–20% of Twitter users were over 50. However, all estimated Cramér's V were <0.07,
113 indicating the differences were not large taken as a whole (Table S4). The estimated proportions of
114 Twitter users in each prefecture based on Twitter data showed strong correlations with those based
115 on the MyVoice survey for all years (Fig. S6c; Pearson's R = 0.952, 0.930, 0.939, 0.947, and 0.952
116 for 2017, 2018, 2019, 2020, and all years, respectively and all p values for the coefficients were <
117 0.001) though the estimation from Twitter showed a higher proportion of users to be in Tokyo (the
118 cluster of points on the upper right side).

119

Fig. S2: Number of respondents of the Survey on Information and Communication Media Usage Time and Information Behavior for each demographic group for each year.



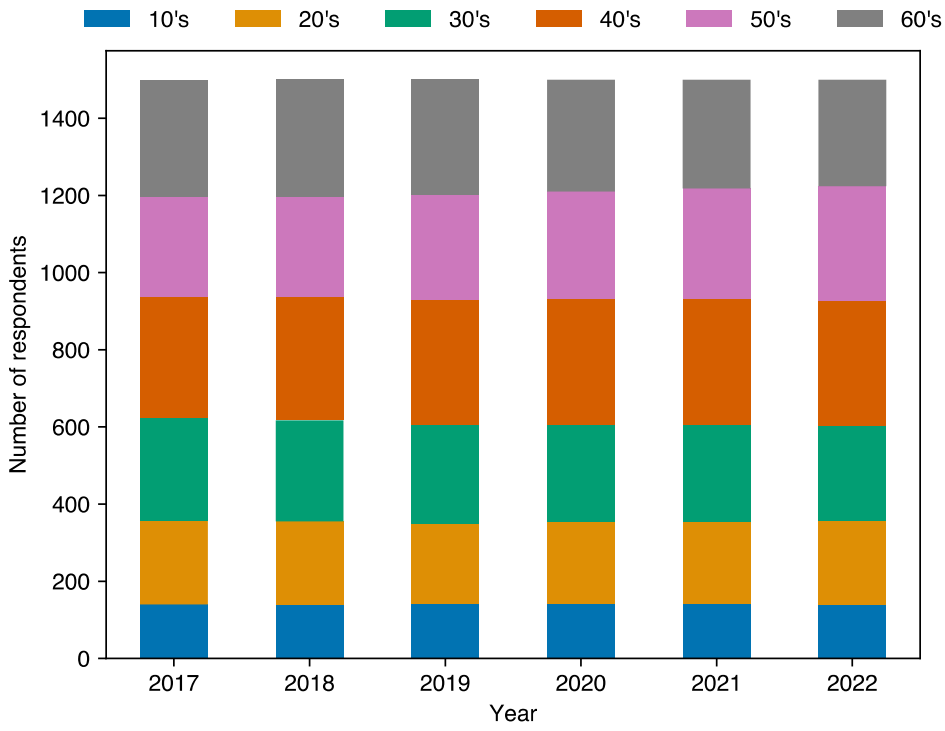
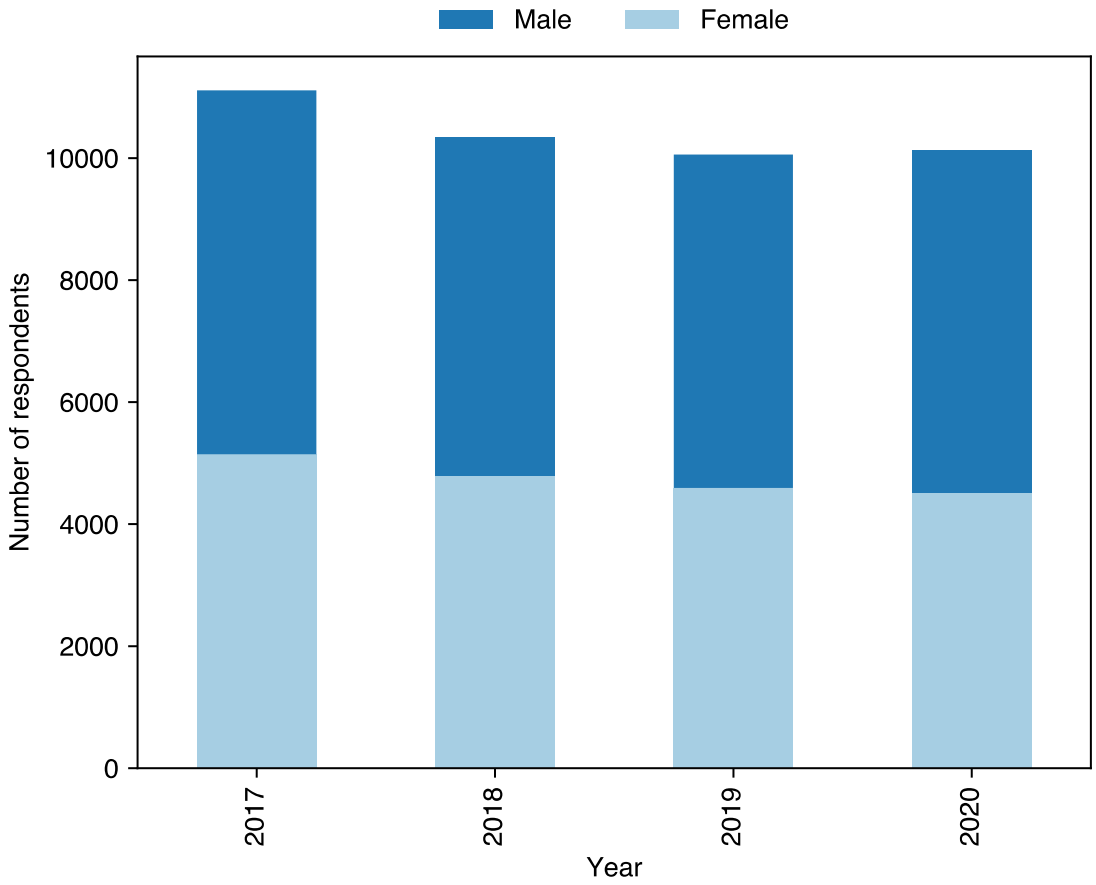
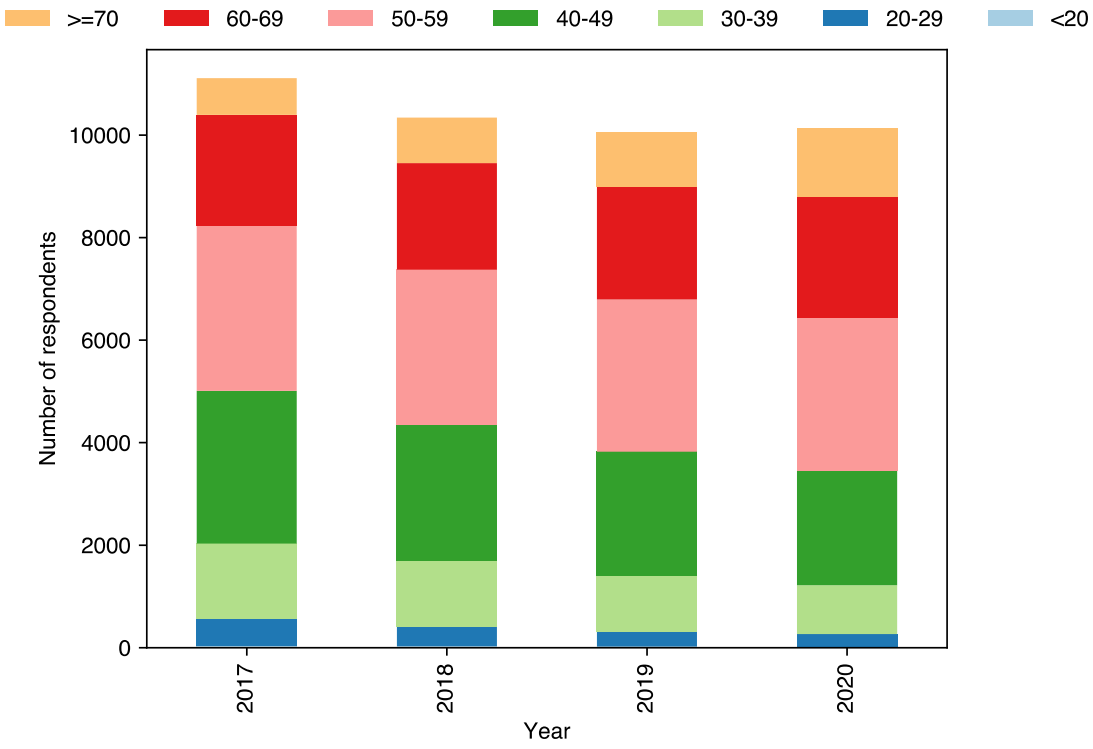
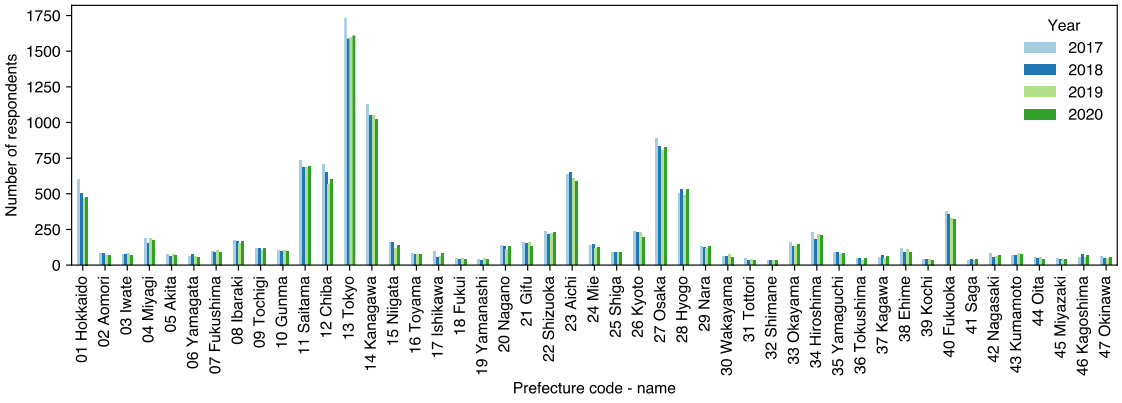


Fig. S3: Number of respondents to the Twitter Usage survey conducted by MyVoice Communications, Inc. for each demographic group for each year.




>=70
60-69
50-59
40-49
30-39
20-29
<20
Number of respondents
0
2000
4000
6000
8000
10000
2017
2018
2019
2020
Year

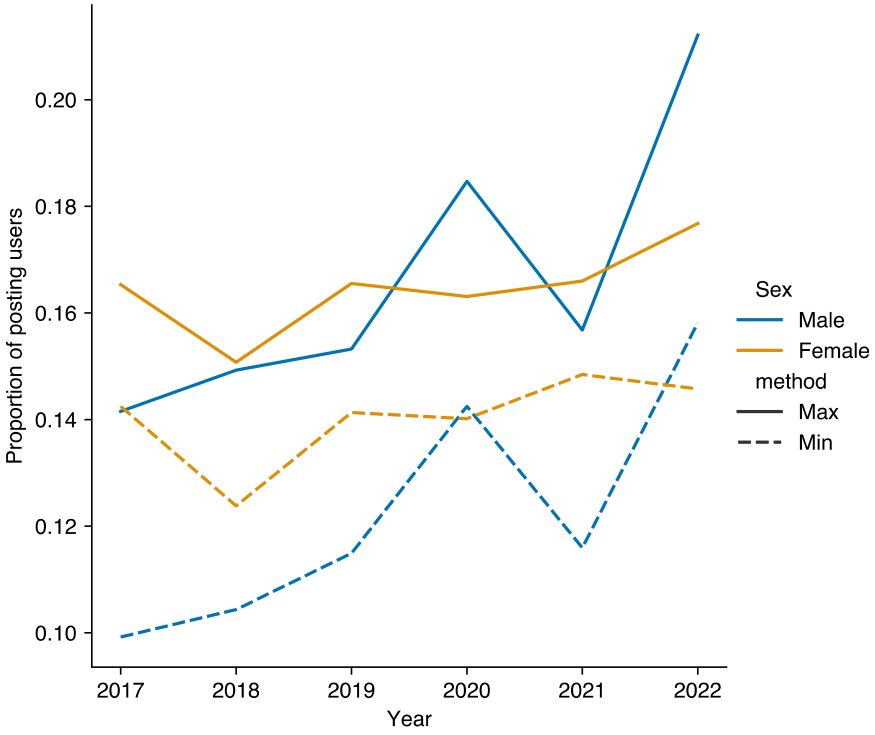
128


Number of respondents
0
250
500
750
1000
1250
1500
1750
Year
2017
2018
2019
2020
01 Hokkaido
02 Aomori
03 Iwate
04 Miyagi
05 Akita
06 Yamagata
07 Fukushima
08 Ibaraki
09 Tochigi
10 Gunma
11 Saitama
12 Chiba
13 Tokyo
14 Kanagawa
15 Niigata
16 Toyama
17 Ishikawa
18 Fukui
19 Yamanashi
20 Nagano
21 Gifu
22 Shizuoka
23 Aichi
24 Mie
25 Shiga
26 Kyoto
27 Osaka
28 Hyogo
29 Nara
30 Wakayama
31 Tottori
32 Shimane
33 Okayama
34 Hiroshima
35 Yamaguchi
36 Tokushima
37 Kagawa
38 Ehime
39 Kochi
40 Fukuoka
41 Saga
42 Nagasaki
43 Kumamoto
44 Oita
45 Miyazaki
46 Kagoshima
47 Okinawa
Prefecture code - name

129

130

Fig. S4: Estimated proportions of users who were posting tweets for each demographic group, method, and year estimated from the MIC survey.



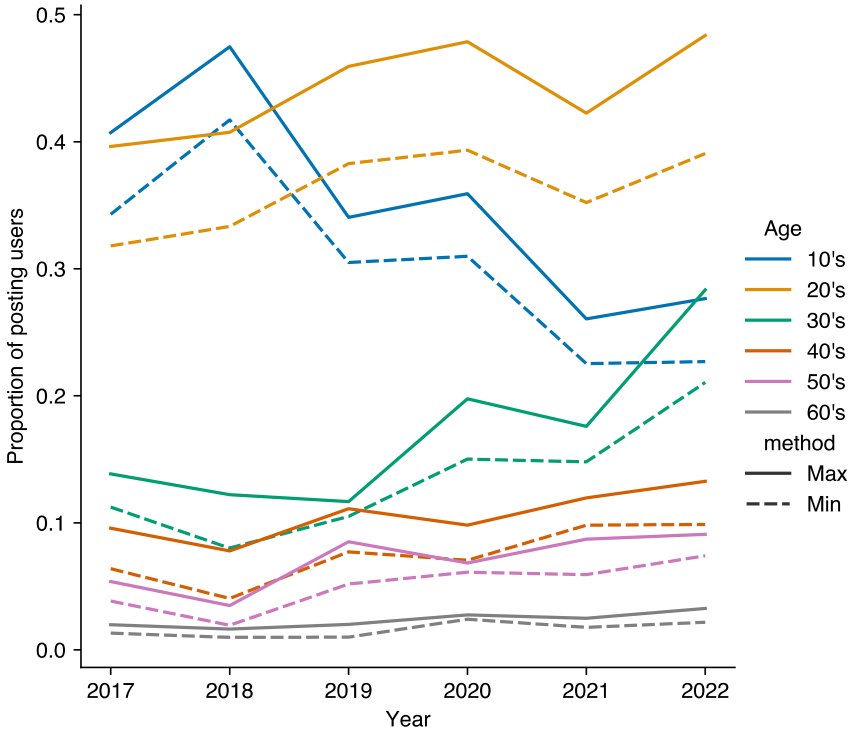
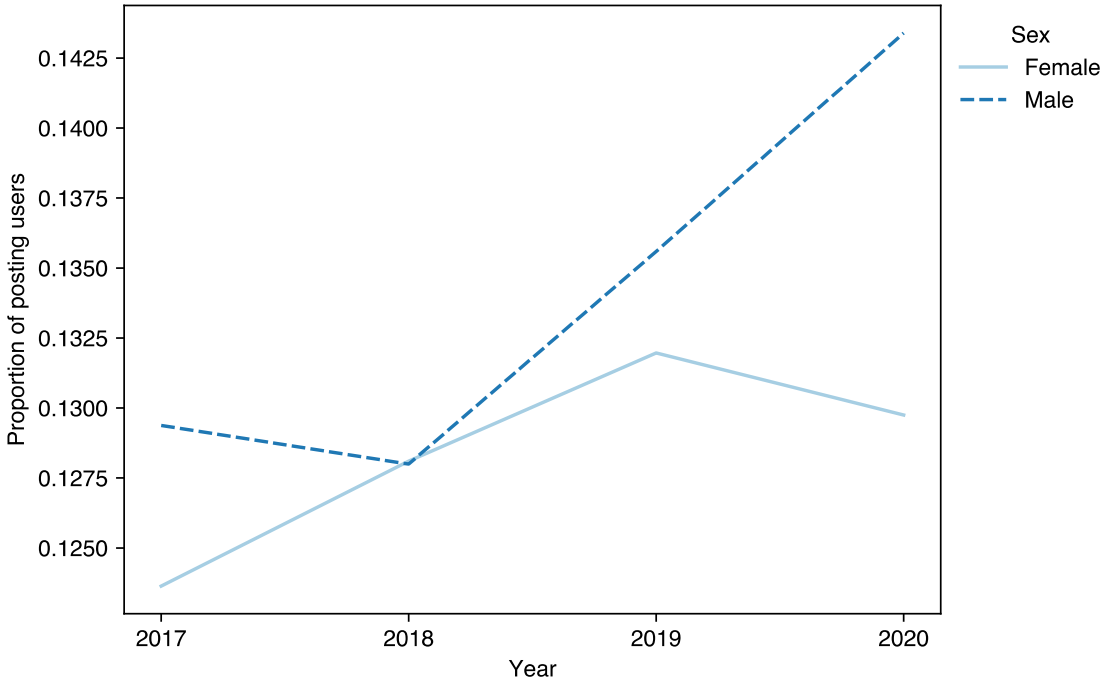
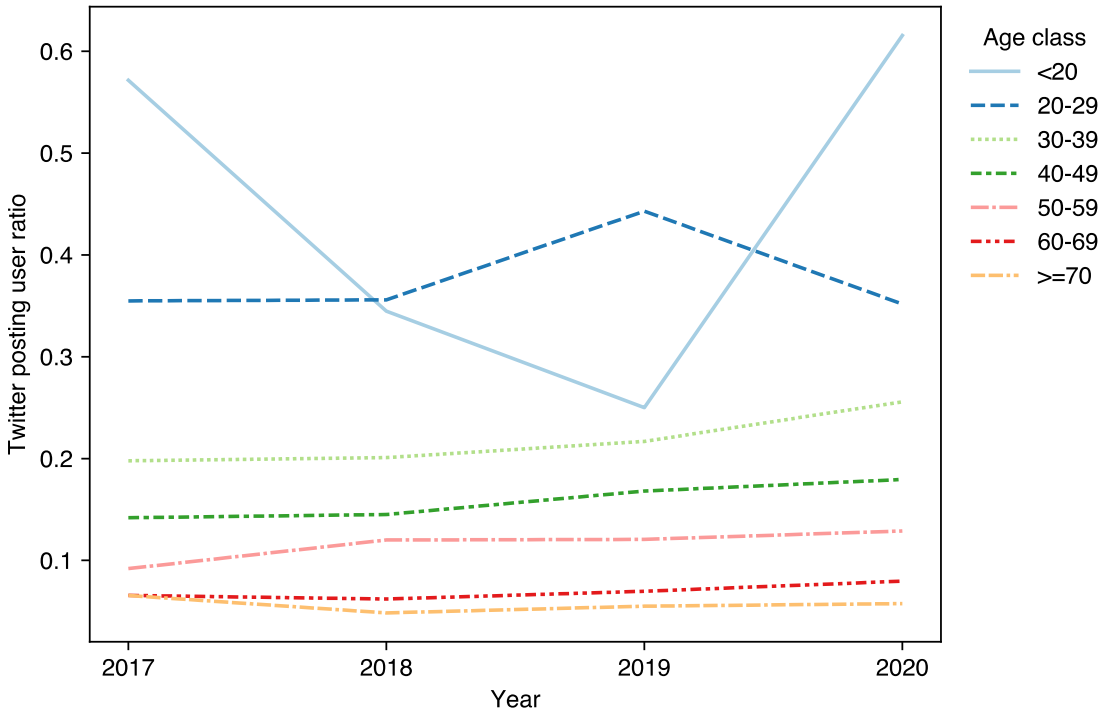
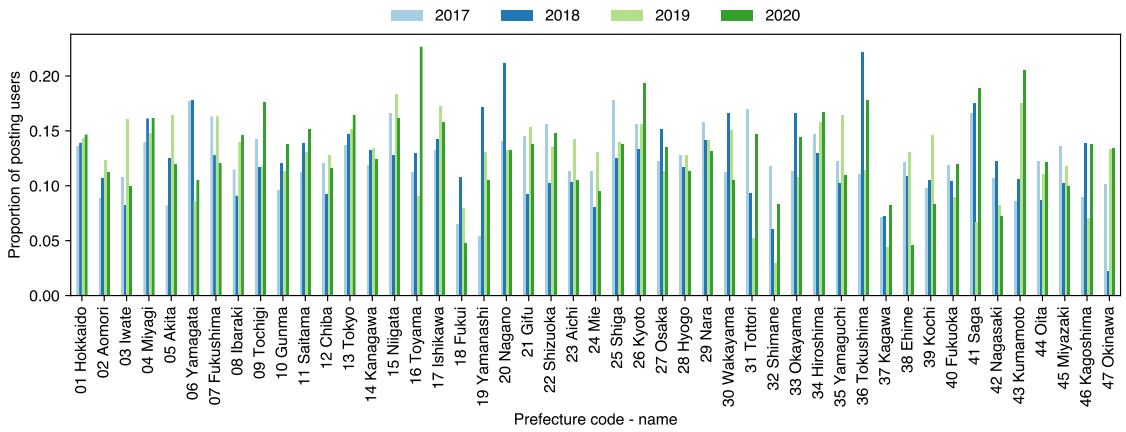


Fig. S5: Proportions of users who were posting tweets for each demographic group and year estimated from the MyVoice survey.





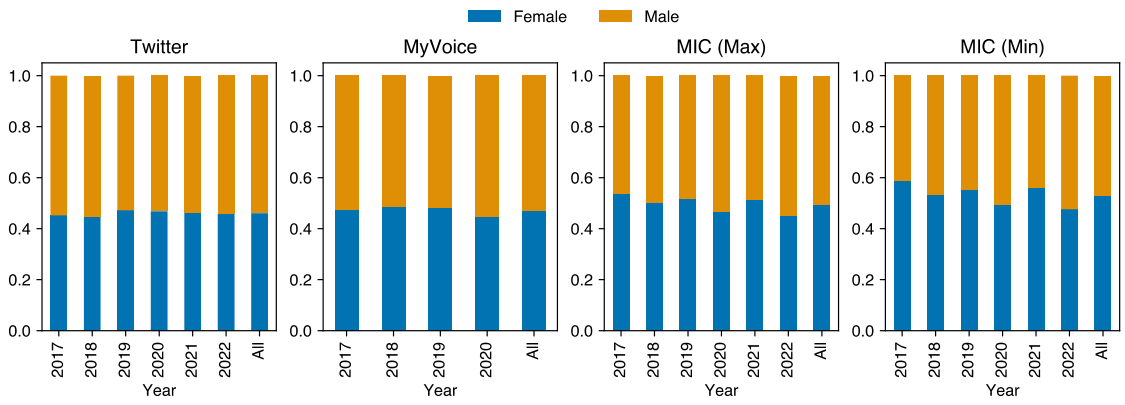

2017
2018
2019
2020
Proportion of posting users
0.00
0.05
0.10
0.15
0.20
01 Hokkaido
02 Aomori
03 Iwate
04 Miyagi
05 Akita
06 Yamagata
07 Fukushima
08 Ibaraki
09 Tochigi
10 Gunma
11 Saitama
12 Chiba
13 Tokyo
14 Kanagawa
15 Niigata
16 Toyama
17 Ishikawa
18 Fukui
19 Yamanashi
20 Nagano
21 Gifu
22 Shizuoka
23 Aichi
24 Mie
25 Shiga
26 Kyoto
27 Osaka
28 Hyogo
29 Nara
30 Wakayama
31 Tottori
32 Shimane
33 Okayama
34 Hiroshima
35 Yamaguchi
36 Tokushima
37 Kagawa
38 Ehime
39 Kochi
40 Fukuoka
41 Saga
42 Nagasaki
43 Kumamoto
44 Oita
45 Miyazaki
46 Kagoshima
47 Okinawa
Prefecture code - name

140

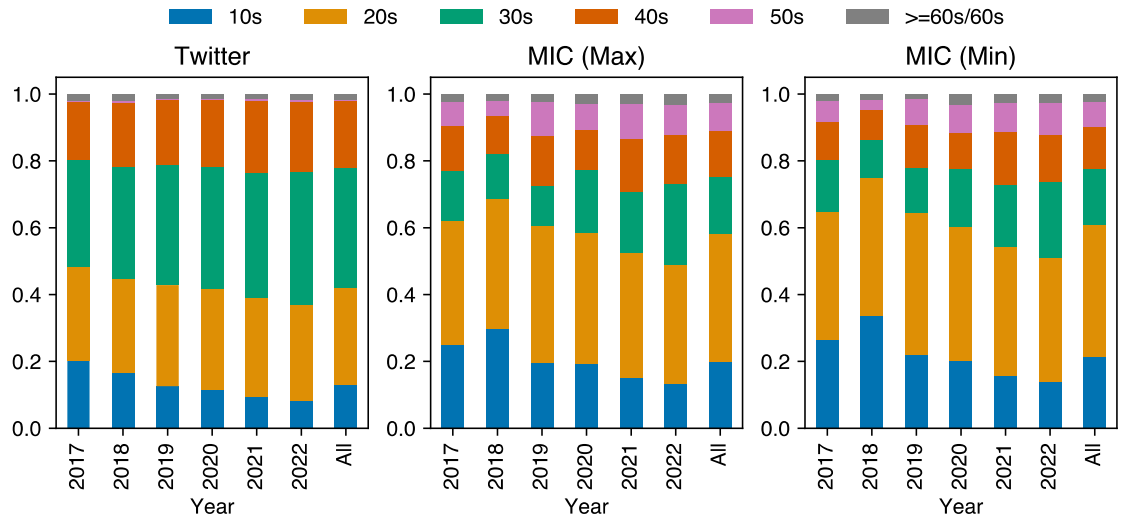
141

Fig. S6: Proportions of Twitter users in each demographic group estimated from Twitter and the existing statistics for each year (a and b) and relationship between proportions of Twitter users in each prefecture estimated using Twitter and the existing statistics (c). The diagonal line shown on the panel (c) denotes y = x.

(a) Sex



(b) Age



(c) Prefecture

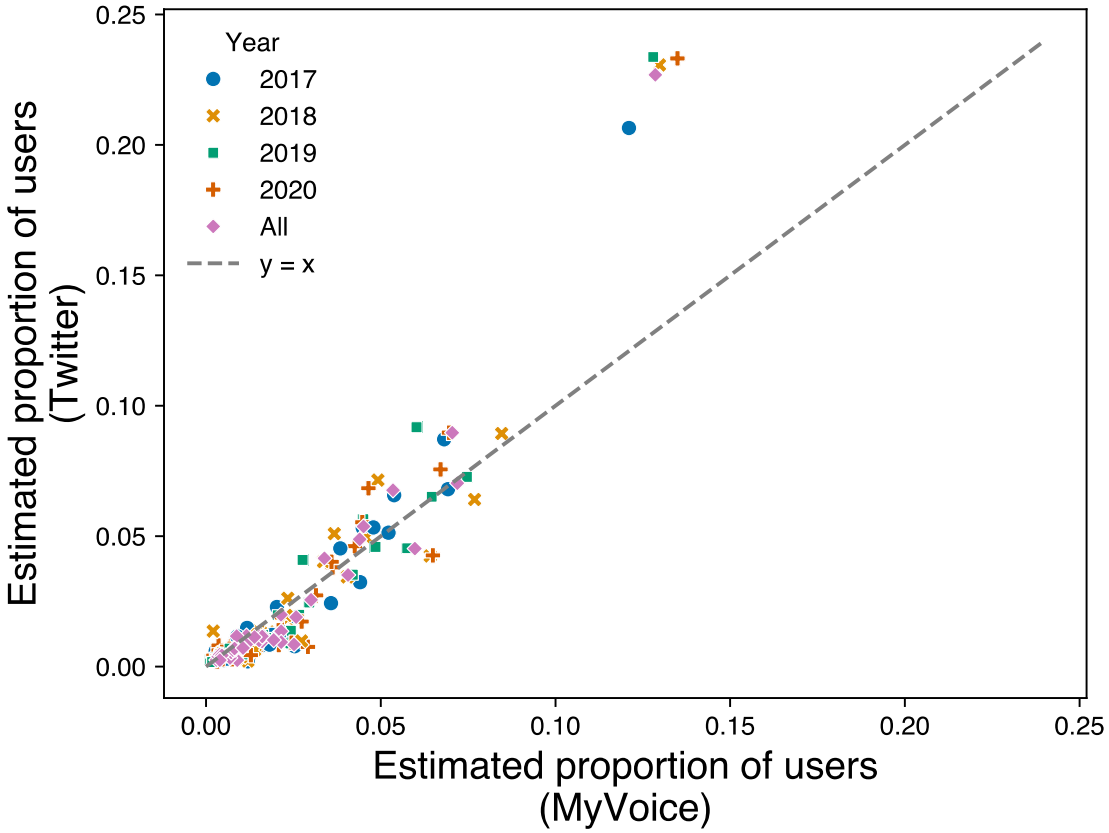


Table S3: Cramér's V statistics for the proportions of each sex estimated based on Twitter and the existing statistics.

Year	MyVoice	MIC (Max)	MIC (Min)
2017	0.0013	0.0062	0.0114
2018	0.0025	0.0042	0.0074
2019	0.0007	0.0034	0.0067
2020	0.0014	0.0001	0.0021
2021	0.0000	0.0036	0.0080
2022	0.0000	0.0005	0.0014
All	0.0015	0.0062	0.0140

Table S4: Cramér's V statistics for the proportions of age groups estimated based on Twitter and the existing statistics.

Year	MIC (Max)	MIC (Min)
2017	0.0202	0.0231
2018	0.0267	0.0368
2019	0.0304	0.0315
2020	0.0214	0.0265
2021	0.0229	0.0243
2022	0.0170	0.0207
All	0.0535	0.0627

Appendix 3: Performance evaluation of sentiment estimation for Japanese short text using IBM Natural Language Understanding.

IBM Watson Natural Language Understanding is a commercially available machine learning service which can estimate sentiment of provided text and returns a sentiment label (positive, neutral, or negative) as well as sentiment score ranging from -1 (negative) to 1 (positive). However, because its estimation performance was not disclosed, we evaluated it by estimating sentiment of an existing corpus with manually attached sentiment labels. Also, we compared its performance with that of other machine learning models and a dictionary-based method.

For the machine learning models, we employed *autonlp-japanese-sentiment-59362* (Thakur, 2021a) and *autonlp-japanese-sentiment-59363* (Thakur, 2021b) which estimate sentiment of Japanese text and return a sentiment label of the text in binary form (negative/positive) along with its confidence score. As for the dictionary-based method, we used python binding of ML-Ask (Ikegami, 2019; Ptaszynski et al., 2009), which assigns a sentiment label of five classes (positive, mostly positive, neutral, mostly negative, and negative) for the analyzed text.

For the test data, we employed the *Japanese realistic textual entailment corpus* (Hayashibe, 2020) which includes a corpus of 5,553 hotel reviews with manually attached sentiment labels (positive, neutral, and negative). We estimated sentiment of all texts in the corpus and evaluated congruence of estimations and manually attached labels for each method using the Matthews correlation coefficient (MCC; Matthews, 1975), The MCC is a correlation coefficient for discrete predictions ranging from -1 to 1 where 1 indicates complete prediction, 0 indicates random prediction, and -1 indicates complete inverse prediction.

Because *autonlp-japanese-sentiment-59362* and *autonlp-japanese-sentiment-59363* do not return neutral labels, we adjusted the threshold value for the confidence score under which the

analyzed text was regarded as neutral from 0.5 to 1, and obtained the best threshold which brings the highest value of MCC. As for the ML-Ask, the labels of *mostly positive* and *mostly negative* were regarded as positive and negative, respectively, and MCC was calculated. IBM Natural Language Understanding failed to estimate sentiments of 76 texts which were too short or contained only symbols. Also, ML-Ask could not assign sentiment to 4,019 texts because of their dictionary-based nature. For these cases, texts without sentiment labels were omitted from the calculations of MCC. Python programs and data are available at Zenodo (doi:10.5281/zenodo.11366621) and Mendeley Data (doi:10.17632/2hvh5yzcbd.1).

The results showed that IBM Natural Language Understanding had the highest value of MCC among the methods investigated (Table S5). Although ML-Ask also showed a relatively high value of MCC, it could estimate only 27.6% of the text. Therefore, we concluded IBM Natural Language Understanding was the best method for this study.

Table S5: Matthews correlation coefficient (MCC) for each method/model.

Model/Method	MCC
IBM Watson Natural Language Understanding	0.550
autonlp-japanese-sentiment-59362	0.402
autonlp-japanese-sentiment-59363	0.390
ML-Ask	0.516

Appendix 4: Text preprocessing applied before the estimation of sentiment.

Before the estimation, we applied a preprocessing algorithm to the text data in each tweet. First, because Foursquare clients add automatically generated text in a fixed format (e.g., "(@ LOCATION in CITY, PREFECTURE) https://t.co/xxxxxxxxxx") at the end of the *text* attribute for tweets posted by clients, this text was removed. Next, invalid characters in tweet text which are rarely used in usual sentences and cannot be stored in Microsoft Excel files were removed using the *openpyxl* module in Python. Finally, URLs in the text were removed. If a tweet contained no characters after the preprocessing, it was omitted from the analysis.

Appendix 5: Evaluation and adjustment of spatial auto correlation

The spatial autocorrelations of the distributions of the number of posts were evaluated using Moran's I. Moran's I measures spatial autocorrelation using a spatial weight matrix ranging from -1 (negative spatial autocorrelation, i.e., dispersion) to 1 (positive spatial autocorrelation, i.e., clustering) and 0 indicates no correlation. The calculation used the *moran.test* function in the *spdep* (Bivand & Wong, 2018) package in R. We calculated a spatial weights matrix as an inverse of pairwise geographical distance among the centers of all the grids with the diagonal of the matrix set to 0 using the *spdep* and *geosphere* (Hijmans, 2024) packages in R. To reduce computational requirements, only grids within 5km were considered to have spatial relatedness. The result showed that the spatial distribution of the number of tweets showed weak but significant spatial autocorrelation (Table S6).

Because we found significant autocorrelations, we removed them by using a spatial filtering approach using spatial lag variables before calculating the correlation coefficients as follows. 1) We calculated a spatial weights matrix as an inverse of pairwise geographical distance among the centers of all the grids with the diagonal of the matrix set to 0 using the *spdep* (Bivand & Wong, 2018) and *geosphere* (Hijmans, 2024) packages in R. Note that to reduce computational requirements, only grids within 5km were considered to have spatial relatedness. 2) We calculated spatial lag (i.e., the weighted average of values at neighboring locations) for count variables using the weight matrix with the *lag.listw* function in the *spdep* package. 3) Then each count was regressed on its own spatial lag using a linear model and the residuals were extracted. These residuals represent variation that was not explained by local spatial autocorrelation. 4) Using the residuals, correlation coefficients were calculated. Note that the analysis was integrated into a Python script using the *rpy2* module.

231 **Table S6**: Moran's I and its two-sided *p*-value for the spatial distributions of the total number of
232 tweets as well as the numbers of tweets for each sex and age group (a) without or (b) with removing
233 massively tweeting users.

234

Category	Group	(a) Moran's I	*p*	(b) Moran's I	*p*
Total	Total	0.104	<0.001	0.103	<0.001
Sex	Male	0.105	<0.001	0.103	<0.001
	Female	0.089	<0.001	0.105	<0.001
Age	<20	0.088	<0.001	0.097	<0.001
	20-29	0.096	<0.001	0.098	<0.001
	30-39	0.108	<0.001	0.109	<0.001
	40-49	0.099	<0.001	0.102	<0.001
	50-59	0.065	<0.001	0.078	<0.001
	>=60	0.038	<0.001	0.071	<0.001

235

236

Table S7: Number of tweets and users obtained for each dataset and condition.

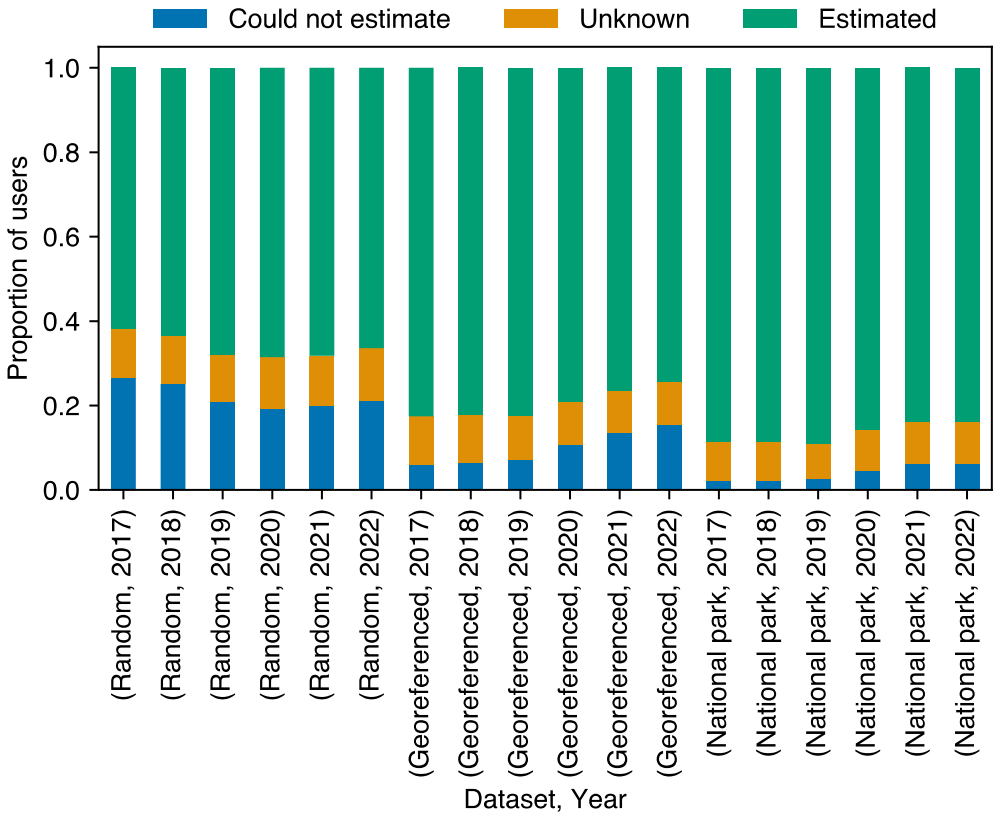
Dataset	Year	Tweets	Users
Randomly sampled Japanese tweets (Dataset 1)	2017	128,182,687	11,698,048
	2018	98,887,809	10,911,155
	2019	83,057,028	10,954,135
	2020	86,980,084	12,561,231
	2021	81,819,279	12,156,577
	2022	74,404,113	11,934,367
	Total	553,331,000	70,215,513
All georeferenced tweets with coordinate or POI	2017	21,464,748	722,011
	2018	23,377,945	729,537
	2019	19,217,940	610,168
	2020	15,384,985	582,433
	2021	15,251,789	538,931
	2022	18,189,046	645,832
	Total	112,886,453	3,828,912
Georeferenced tweets having Japanese characters posted from official Twitter or Foursquare clients having coordinates on the standard 2nd mesh (Dataset 2)	2017	4,331,931	345,129
	2018	5,485,961	386,304
	2019	4,778,818	326,105
	2020	5,533,190	383,847
	2021	6,044,083	371,094
	2022	8,032,706	456,544
	Total	34,206,689	2,269,023
All national park tweets	2017	311,734	72,895
	2018	353,763	78,305
	2019	291,113	64,854
	2020	308,700	64,871
	2021	336,587	61,631
	2022	411,269	76,164
	Total	2,013,166	418,720

Dataset	Year	Tweets	Users
National Park tweets with Japanese characters posted from the official the Twitter or Foursquare clients (Dataset 3)	2017	68,518	26,454
	2018	98,269	35,231
	2019	84,432	29,516
	2020	144,446	40,536
	2021	171,512	41,559
	2022	226,544	53,669
	Total	793,721	226,965

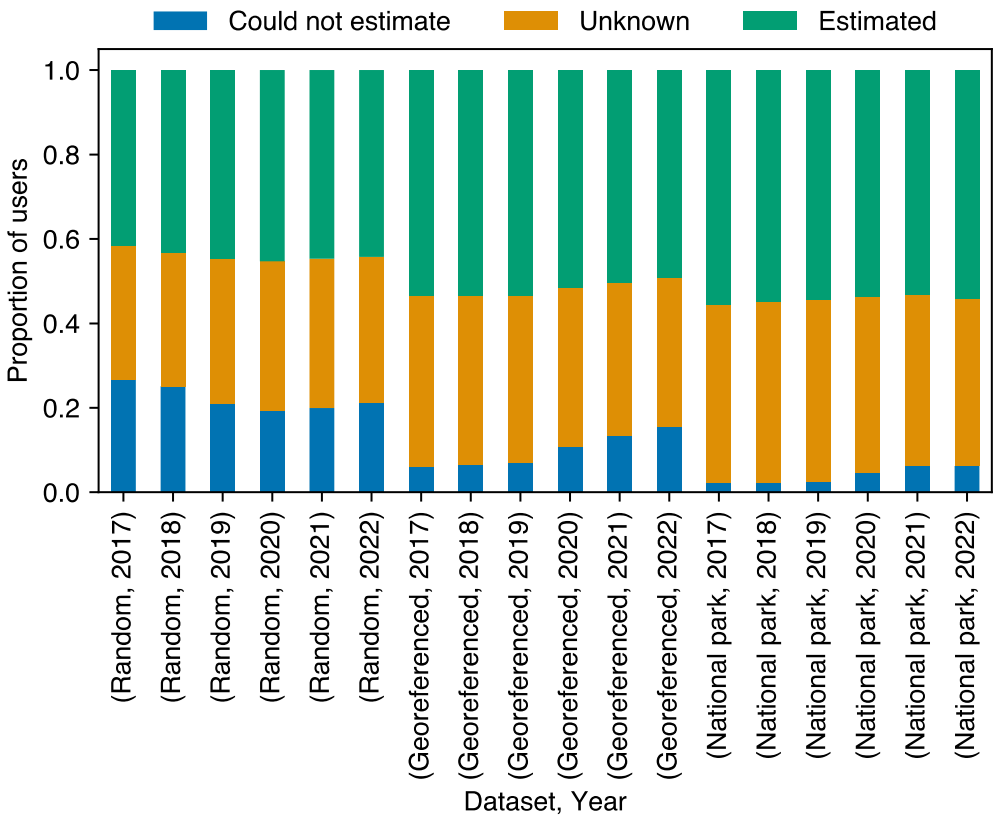
240

Fig. S7: Proportion of users whose attributes (a: sex, b: age, and c: prefecture) could not be estimated (e.g., deleted or private users), users whose attributes could be estimated but the attributes were unknown (i.e., public users but their attributes could not be estimated), and users whose attributes were successfully estimated.

(a) Sex



(b) Age


Could not estimate
Unknown
Estimated
Proportion of users
1.0
0.8
0.6
0.4
0.2
0.0
(Random, 2017)
(Random, 2018)
(Random, 2019)
(Random, 2020)
(Random, 2021)
(Random, 2022)
(Georeferenced, 2017)
(Georeferenced, 2018)
(Georeferenced, 2019)
(Georeferenced, 2020)
(Georeferenced, 2021)
(Georeferenced, 2022)
(National park, 2017)
(National park, 2018)
(National park, 2019)
(National park, 2020)
(National park, 2021)
(National park, 2022)
Dataset, Year

(c) Residential prefecture

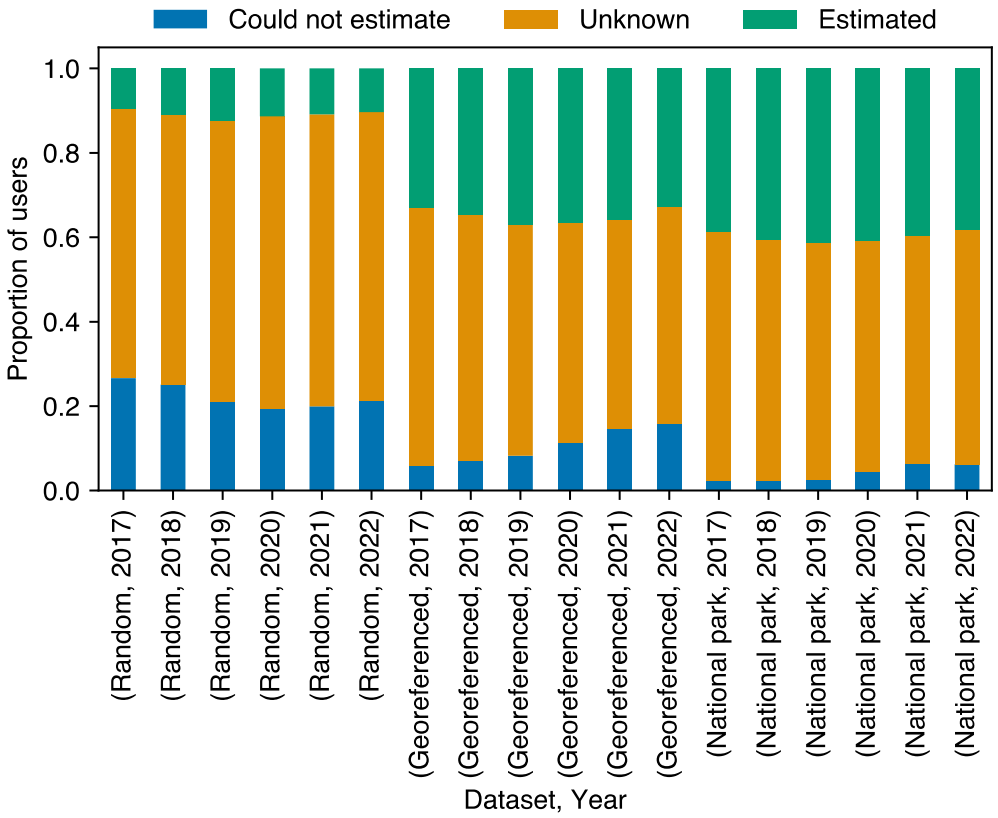
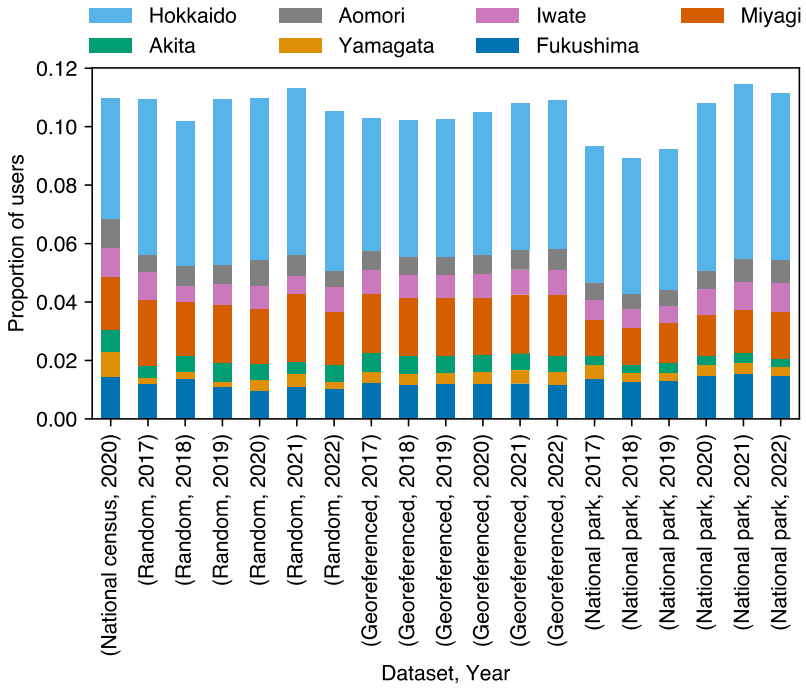
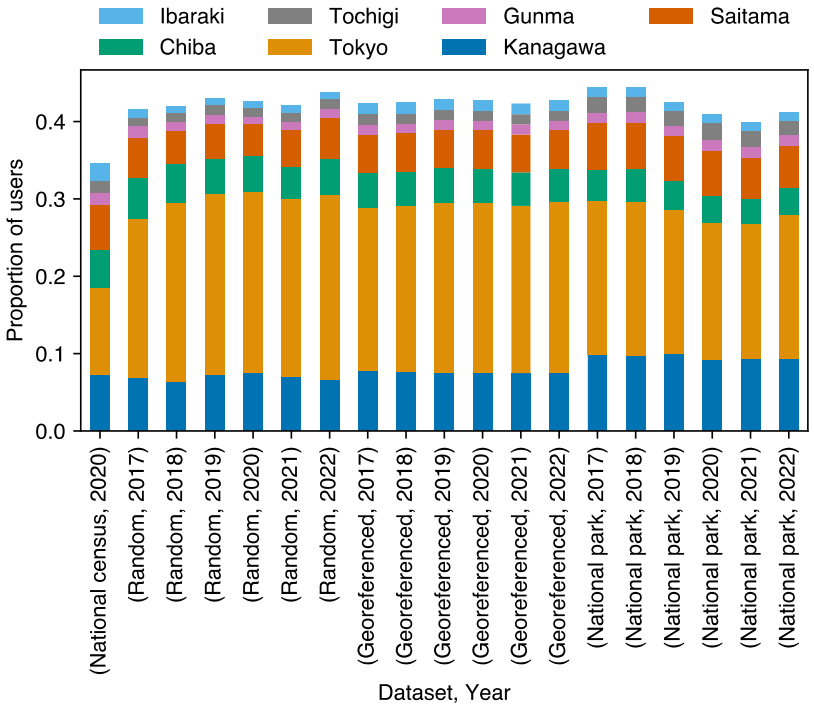


Fig. S8: Estimated proportions of residential prefectures of the Twitter users for each dataset in each year for (a) Tohoku and Hokkaido, (b) Kanto, (c) Chubu, (d) Kinki, (e) Chugoku, (f) Shikoku, and (g) Kyushu and Okinawa regions.

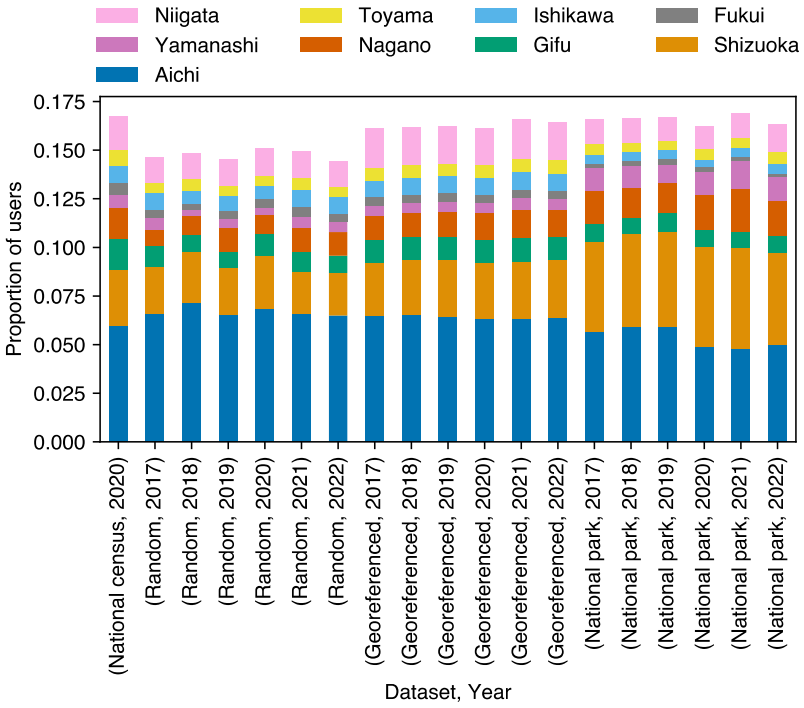
(a) Tohoku and Hokkaido region



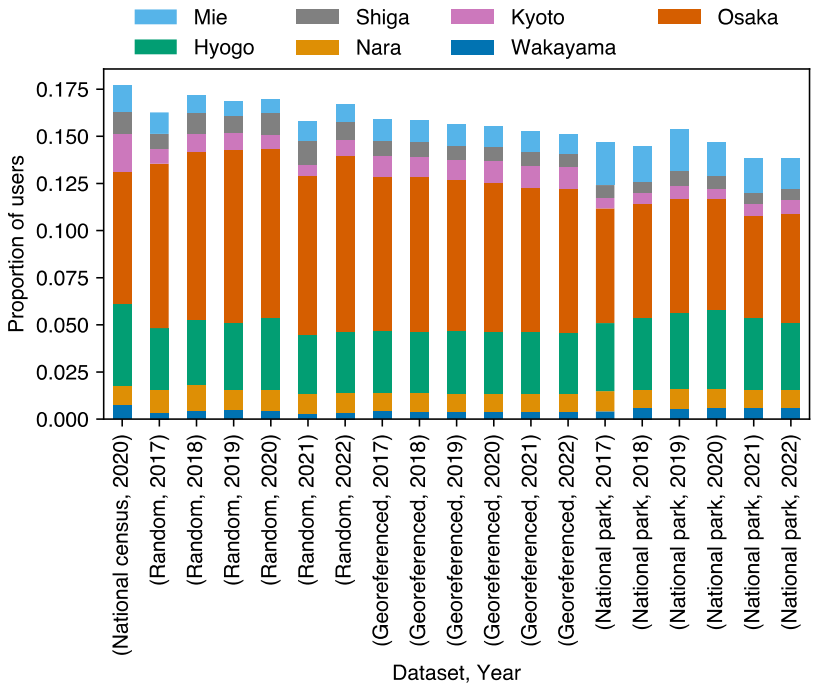
(b) Kanto region



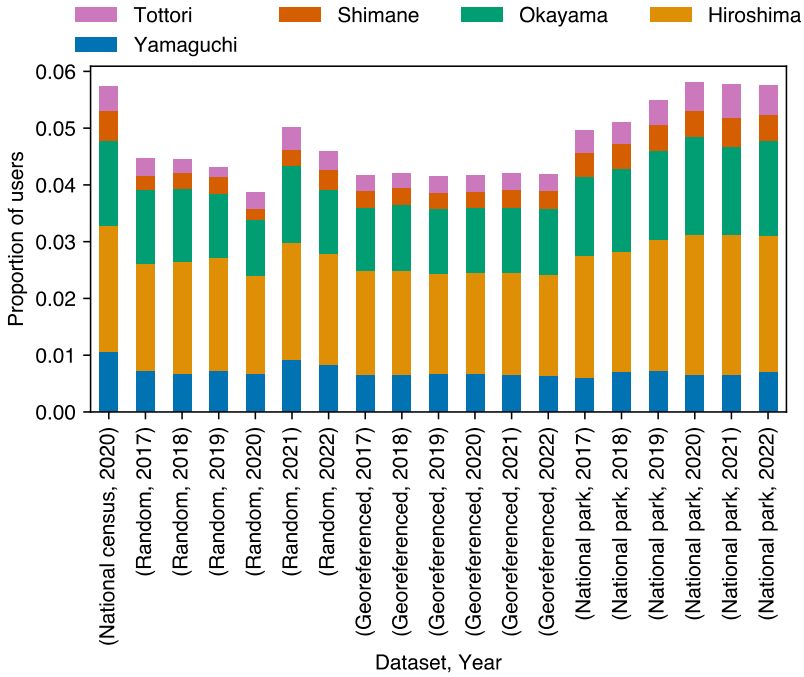
(c) Chubu region



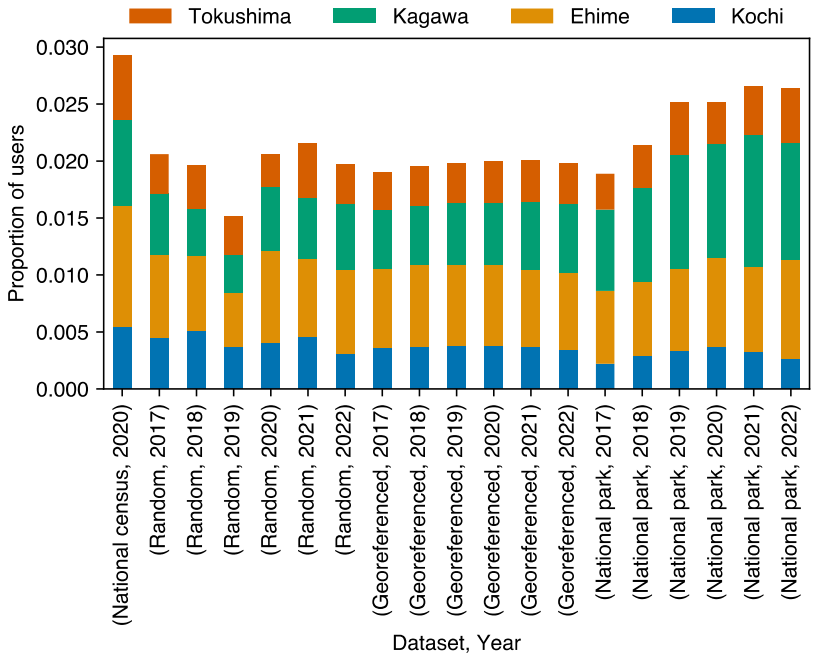
(d) Kinki region



(e) Chugoku region



(f) Shikoku region



(g) Kyushu and Okinawa region

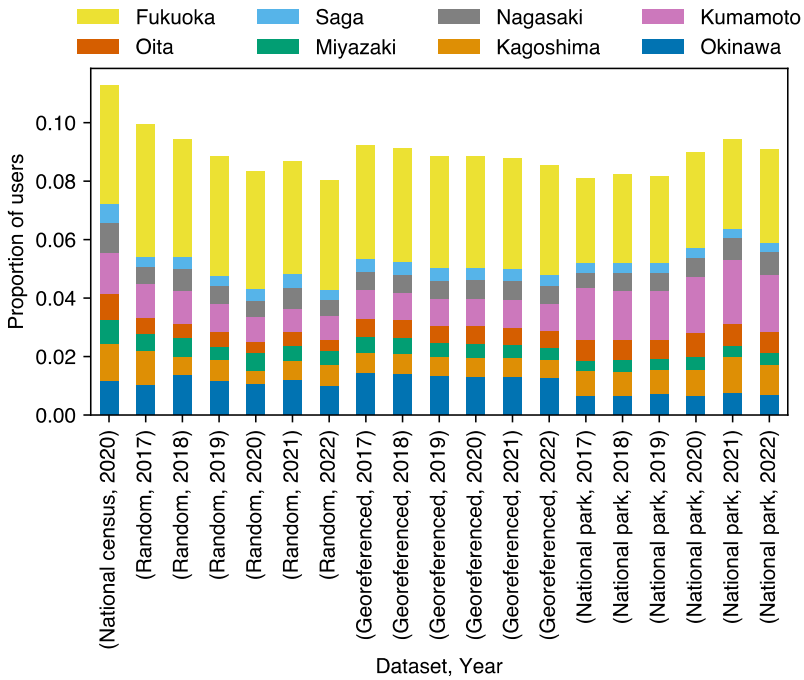
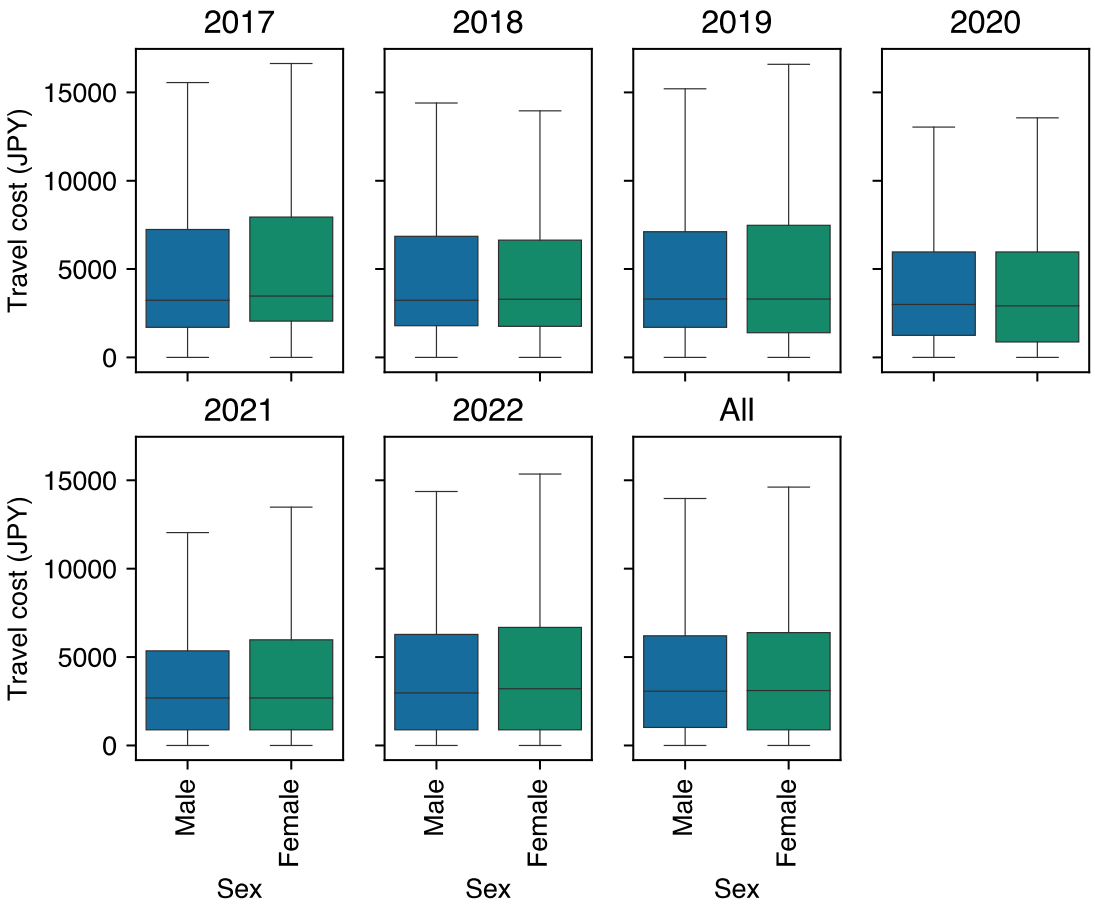


Table S8: Result of the type II analysis of deviance for the GLMM testing differences in travel cost among client type (Twitter/Foursquare), age, sex, and year. For this model, 155,659 tweets from 34,147 users were used. The marginal R^2 (i.e., performance of explanatory variables) of the model was 0.001.

	χ^2	df	*p*	
Foursquare	0.789	1	0.374404	
Age	19.7569	5	0.001388	**
Sex	0.2882	1	0.591376	
Year	85.6179	5	<0.001	***

Fig. S9: Box-whisker plots showing distributions of estimated travel cost for (a) each sex and (b) each age group for each year. Note that to reduce the dominance of massively tweeting users, only one tweet for one day for one grid for each user was randomly selected and used.

(a)



(b)

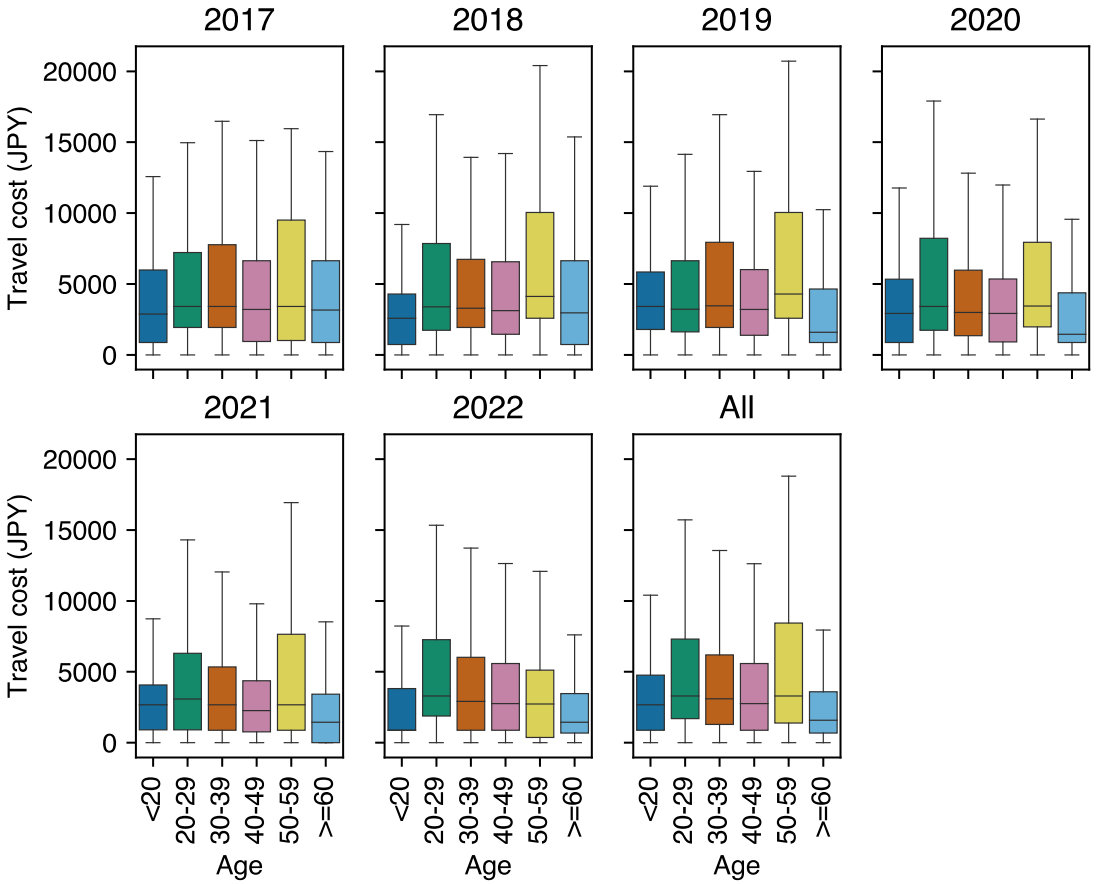
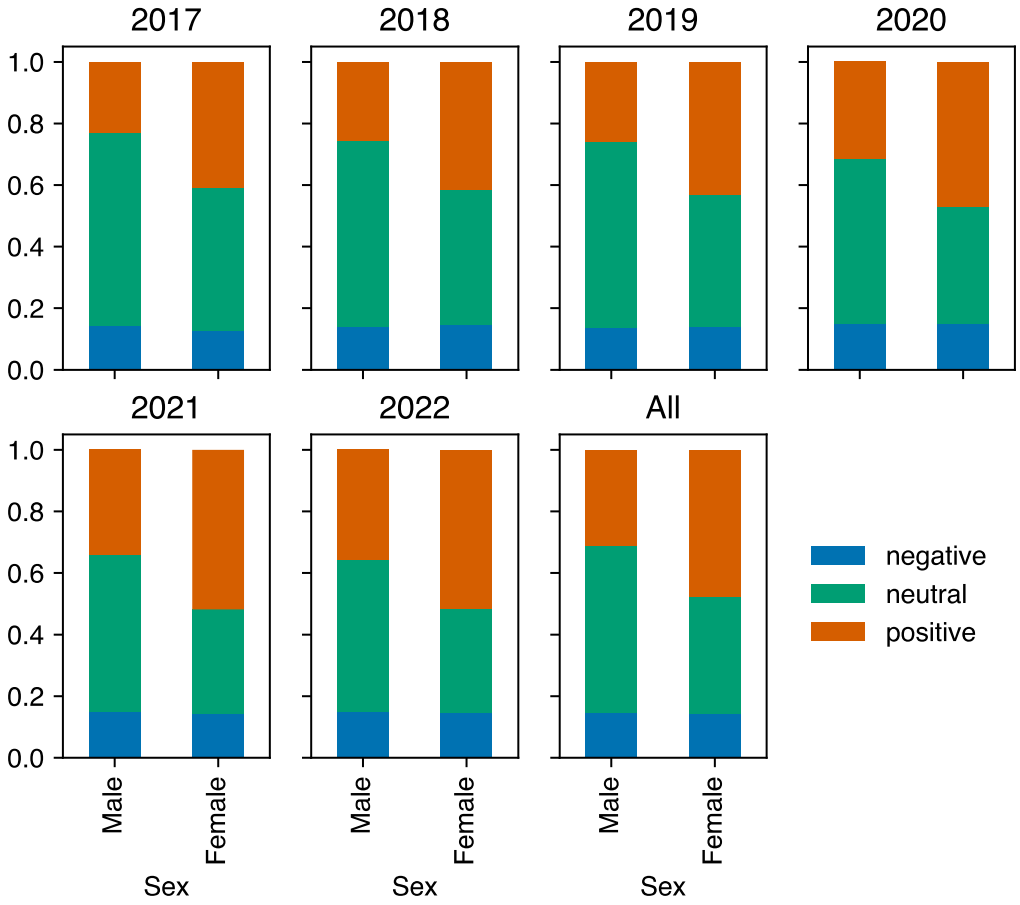

2017
2018
2019
2020
2021
2022
All
Travel cost (JPY)
0
5000
10000
15000
20000
<20
20-29
30-39
40-49
50-59
>=60
Age

Table S9: Results of the type II analysis of deviance for the GLMM testing differences in sentiment score among client type (Twitter/Foursquare), age, sex, residential prefecture, and year. For this model, 150,646 tweets from 33,558 users were used. The marginal R^2 of the model was 0.025.

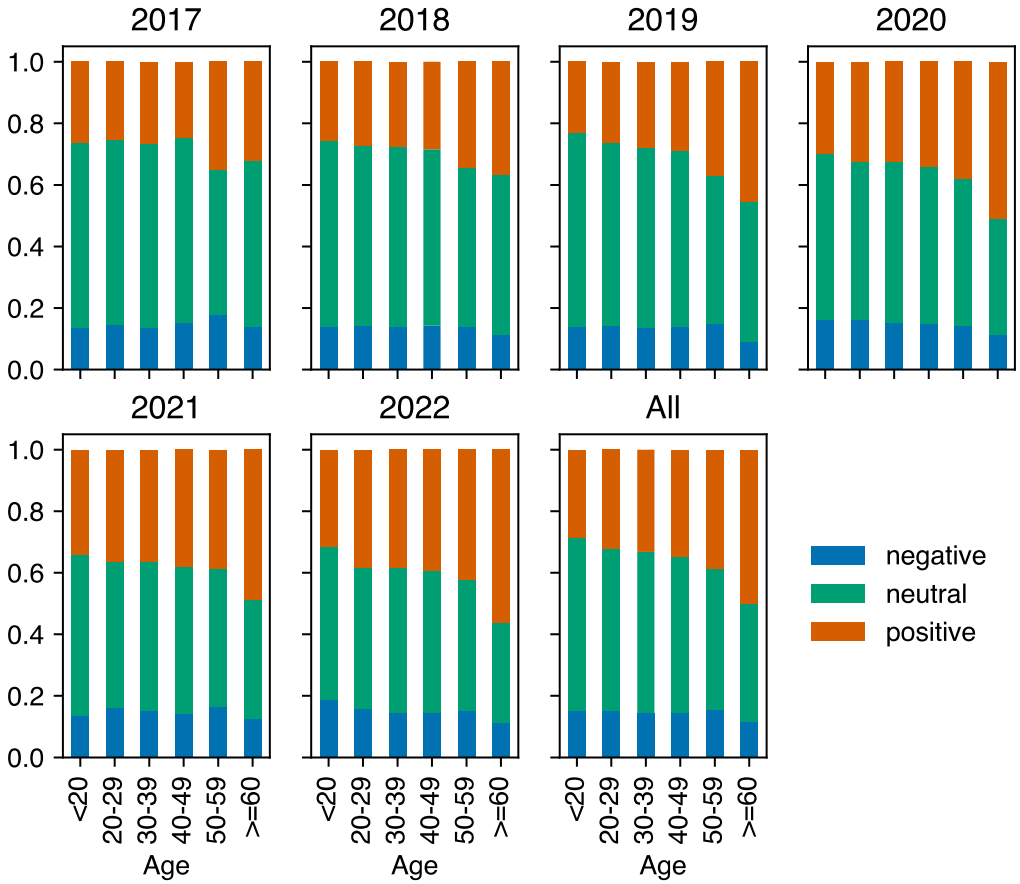
	χ^2	df	*p*	
Foursquare	797.383	1	<0.001	***
Age	119.57	5	<0.001	***
Sex	491.066	1	<0.001	***
Prefecture	68.644	46	0.01684	**
Year	97.371	5	<0.001	***

Fig. S10: Proportion of tweets having positive, neutral, or negative sentiment labels for (a) each sex and (b) each age group for each year. Note that to reduce the dominance of massively tweeting users, only one tweet for one day for one grid for each user was randomly selected and used.

(a)



(b)


2017
2018
2019
2020
2021
2022
All
1.0
0.8
0.6
0.4
0.2
0.0
<20
20-29
30-39
40-49
50-59
>=60
Age
negative
neutral
positive

Appendix 6: Prediction of the sentiment labels using a machine learning model.

In addition to the model showing effects of the demographic attributes of users on sentiment score, we implemented a machine learning model predicting sentiment labels (negative/neutral/positive) from the demographic attributes. We used the random forest which can automatically incorporate interactions among the explanatory variables implemented in the *randomForest* package (Liaw & Wiener, 2002) with R. For the explanatory variables, we used sex, age, residential prefecture, client type, and year. For the hyperparameters of the model, we used default values.

The resultant model could not predict sentiment labels from the explanatory variables. The value of the Cohen's kappa (Cohen, 1960) for the prediction of the model was 0.152, indicating the sentiment labels could not be predicted from the explanatory variables well (Landis & Koch, 1977).

Table S10: Result of the type II analysis of deviance for the GLMM testing effects of age, sex, and their interaction on the number of tweets per user. For this model 565,641 tweets from 126,494 users were used. The marginal R^2 of the model was 0.03.

	χ^2	df	*p*	
Sex	616.791	1	<0.001	***
Age	902.238	5	<0.001	***
Sex × Age	8.555	5	0.128	

Table S11: Result of the type II analysis of deviance for the GLMM testing effects of age, sex, and their interaction on the number of tweets per user after removing the top 1% of users with the highest number of tweets. For this model 421,789 tweets from 125,036 users were used. The marginal R^2 of the model was 0.02.

	χ^2	df	*p*	
Sex	569.9959	1	<0.001	***
Age	653.3263	5	<0.001	***
Sex × Age	5.0596	5	0.4086	

References

Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, *27*(3), 716–748. https://doi.org/10.1007/s11749-018-0599-x

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton university press.

Hayashibe, Y. (2020). Japanese realistic textual entailment corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6827–6834. https://aclanthology.org/2020.lrec-1.843

Hijmans, R. J. (2024). *geosphere: Spherical Trigonometry* (Version R package version 1.5-20) [Computer software]. https://CRAN.R-project.org/package=geosphere

Ikegami, Y. (2019). *Pymlask* (Version 0.3.2) [Computer software]. https://github.com/ikegami-yukino/pymlask/

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R., & Araki, K. (2009). A System for Affect Analysis of Utterances in Japanese Supported with Web Mining. **知能と情報**, *21*(2), 194–213. https://doi.org/10.3156/jsoft.21.194

Thakur, A. (2021a). *Autonlp-japanese-sentiment-59362* [Dataset]. https://huggingface.co/abhishek/autonlp-japanese-sentiment-59362

Thakur, A. (2021b). *Autonlp-japanese-sentiment-59363* [Dataset]. https://huggingface.co/abhishek/autonlp-japanese-sentiment-59363