Measuring Geographic Preferences in Social Networks Beyond Population Bias

Keitaro Takayanagi
Kyoto University
Kyoto, Japan
takayanagi.keitaro.54h@st.kyoto-u.ac.jp

Shiori Hironaka
Kyoto University
Kyoto, Japan
hironaka@media.kyoto-u.ac.jp

Kazuyuki Shudo
Kyoto University
Kyoto, Japan

***Abstract*—Geographic proximity continues to shape social connections in online networks despite the removal of physical boundaries. Theories of urban sociality offer contrasting expectations: the cosmopolitan hypothesis posits that residents of large cities maintain geographically dispersed and socially diverse relationships, whereas the local segregation hypothesis argues that urban scale reinforces proximity and homogeneity within spatially divided communities. However, conventional measures of geographic homophily may conflate these patterns with population density, since users in dense areas are more likely to connect locally even under random mixing. We introduce a population-bias–corrected measure of geographic inbreeding homophily that normalizes observed proximity by its expected value under a null model. Using Twitter mutual-following networks across ten countries with over 4 million users and 80 million connections, we find that users consistently prefer geographically proximate connections beyond random expectation. Yet after correcting for population bias, the association between population size and geographic preference weakens or reverses in most countries. These results reveal that neither the cosmopolitan nor the segregation hypothesis universally explains urban social connectivity, highlighting the contextual nature of geographic preference in online networks.**

***Index Terms*—homophily, social networks, spatial proximity, cross-country analysis**

I. Introduction

Online social networks have enabled people to overcome geographical boundaries and connect with others globally. However, despite this apparent freedom from spatial constraints, social ties in online environments remain strongly shaped by geography. Empirical studies have shown that the probability and frequency of social interactions decrease with increasing geographical distance, suggesting that spatial structures of the physical world are reflected even in online environments with relatively few geographical constraints [1]–[3].

Homophily, a fundamental concept in social network analysis, refers to the tendency of individuals to associate with others who are similar to themselves in various respects, such as age [4], gender [5], race [6], or occupation [7], [8]. A specific form of this tendency, geographic homophily, describes the preference for forming ties with others who live nearby. Because geographically proximate areas often share similar social and economic environments, geographic homophily can also be interpreted as a spatial manifestation of social or economic similarity.

This expectation aligns with the cosmopolitan hypothesis in urban sociology, which posits that residents of large and dense cities are exposed to diverse social contexts and thus form more geographically dispersed and heterogeneous relationships. Supporting this intuition, a study of MySpace users in the United States examined social media interactions and found that urban users' ties, particularly their strong ties, were more geographically distributed than those of rural users and weak ties [9]. Similarly, an analysis of Facebook networks in the United States found that users in densely populated areas were more likely to maintain long-distance friendships than those in sparsely populated regions, indicating more geographically dispersed social ties [2].

In contrast, other studies have suggested that social relationships in urban areas tend to form among geographically proximate individuals. A recent mobility-based analysis using large-scale mobile phone data from over nine million individuals in the United States revealed that everyday encounters were strongly segregated by socioeconomic status, with this exposure segregation being substantially higher in large metropolitan areas than in smaller ones [10]. This finding challenges the long-standing cosmopolitan mixing hypothesis, suggesting that urban scale can amplify rather than reduce spatial and social segregation. A large-scale cross-country analysis of mutual following relationships in Twitter, in which users' home locations were inferred from geotagged tweets, quantitatively evaluated the proximity between connected users [11]. The study found that both the mean and variance of geographic homophily at the area level increased with local population size, implying that residents in densely populated regions were more likely to connect with geographically closer others.

However, this previous work did not account for population imbalance. In densely populated areas, even random connections are more likely to occur among nearby users, potentially leading to an overestimation of geographic homophily. Moreover, because the existing measure captures only the observed proximity of connected users, it cannot distinguish genuine preferences for nearby connections from apparent proximity caused by population distribution. To more rigorously assess such preferences, a new measure is required that corrects for population bias by comparing observed patterns with their expected values under a null model of random connections.

Homophily can generally be decomposed into two com-

10th Int'l Workshop on Application of Big Data
for Computational Social Science (ABCSS 2025), December 2025

ponents: baseline homophily, which arises from the compositional distribution of categories, and inbreeding homophily, which represents the excess tendency to connect with similar others beyond this baseline [7]. Building on this conceptual framework, we propose a new measure of geographic inbreeding homophily that corrects for population bias by normalizing the observed geographic homophily with respect to its expected value under a null model in which every user connects to others with equal probability.

Using large-scale mutual following networks on Twitter across ten countries, this study addresses the following two research questions:

RQ1 Does geographic preference exist after correcting for population bias?

RQ2 Is geographic preference stronger in more populated areas?

The proposed index yielded positive values in all countries, suggesting that users tended to connect with geographically closer others more often than expected by chance. After adjusting for population bias, this relationship with population size weakened in most cases and even reversed in some countries, including Japan, India, and Turkey.

The main contributions of this study are summarized as follows: (1) We introduce a population-bias-adjusted index for quantifying users' geographic preferences. based on the expected value of homophily under a null model. (2) We empirically show that the conventional claim that geographic homophily is stronger in urban areas requires reconsideration when population bias is removed. (3) We demonstrate that the relationship between geographic preference and population size varies across countries, revealing heterogeneous spatial dynamics in online social networks.

II. Data

The datasets used in this study are identical to those employed in previous works [11]–[13]. These datasets were constructed from more than one billion geotagged tweets collected from Twitter in 2019, covering ten countries. Each dataset contains users who posted at least ten geotagged tweets, along with their location information and social graph data based on mutual following relationships. For each user, the area from which they posted the most geotagged tweets was assigned as their home location. The social graph was constructed using mutual following relationships among users, and only users with at least one mutual connection were included in the analysis. The geographic distance between areas was computed using the `geodesic` function of the `geopy` library, based on the centroid coordinates of the bounding box of each area.

Table I summarizes the number of users, mutual connections and home locations for each country dataset. The datasets correspond to the ten countries with the largest numbers of geotagged Twitter users in our collection. Among them, the United States exhibits the largest network, consisting of over 1.7 million users and approximately 40 million mutual connections. The number of unique home locations varies considerably across countries; for instance, Turkey and Saudi Arabia have relatively few distinct location labels, reflecting differences in administrative granularity and the density of geotagged tweets.

TABLE I
Number of users, relationships, and home locations by countries

Country	Users	Relationships	Locations
United States	1 748 492	40 486 682	13 788
Brazil	591 526	17 829 635	3 988
United Kingdom	395 492	7 417 857	7 176
Japan	325 846	4 652 761	1 847
Philippines	207 758	2 878 156	1 305
India	197 038	1 394 261	2 554
Turkey	188 928	2 928 885	468
Mexico	165 315	1 966 862	1 173
Saudi Arabia	128 700	1 295 182	169
Indonesia	122 686	1 159 229	2 808

III. Geographic Homophily

Geographic homophily refers to the extent to which individuals form social ties with others who are geographically proximate. It captures the spatial aspect of social affinity, reflecting how physical distance shapes online as well as offline interactions. In online social networks, even though geographic constraints are relaxed, users still tend to connect with those located nearby, indicating that geographical proximity continues to play a significant role in shaping social structures. To quantify this tendency, we define a geographic homophily index that measures the spatial closeness between users and their mutual connections. However, when comparing users across areas with different population densities, such as urban and rural regions, this measure may be biased, since users in densely populated areas have more opportunities to connect with nearby individuals even under random linking. To account for this, we introduce both the observed and baseline homophily indices before defining a population-bias-corrected inbreeding homophily index.

The observed homophily [11] of geographic proximity with their mutual followers is defined as follows:

$$\mathrm{H}(u) = \sum_{v \in N(u)} \frac{\mathrm{CCDF}(d_{u,v})}{|N(u)|}. \tag{1}$$

Here, $N(u)$ is the set of mutual followers of user u, and $d_{u,v}$ is the geographic distance between users u and v. $\mathrm{CCDF}(d)$ is the complementary cumulative distribution function, which converts the geographic distance into a similarity measure in the range $[0, 1]$. Because the maximum distance appearing in the dataset varies between datasets, $\mathrm{CCDF}(d)$ was calculated based on the empirical distance distribution within each dataset. When $d = 0$, $\mathrm{CCDF}(d) = 1$, and as d increases, $\mathrm{CCDF}(d)$ approaches 0. Thus, $\mathrm{H}(u)$ takes values in the range $[0, 1]$, increasing toward 1 when user u is geographically closer to their mutual followers, and decreasing toward 0 when they are more distant.

The observed homophily index $\mathrm{H}(u)$ captures the average spatial closeness between a user and their mutual followers in the actual social network. However, comparing $\mathrm{H}(u)$ across users living in urban and rural areas introduces a population bias. Users in densely populated areas tend to have higher $\mathrm{H}(u)$ values because, even under random connection, there are many nearby users available to connect with. Conversely, users in sparsely populated areas are more likely to have lower $\mathrm{H}(u)$ values because few nearby users exist. Therefore, to properly evaluate users' geographic preference, it is necessary to correct $\mathrm{H}(u)$ by the expected value of geographic homophily under random connection.

We define the expected value of geographic homophily under random connections as baseline homophily, denoted as $\omega(u)$. This represents the expected value of geographic homophily when user u connects with all other users in the network with uniform probability, and corresponds to a null model that fixes the spatial distribution of users. Formally, it is expressed as:

$$\omega(u) = \sum_{v \in V \setminus \{u\}} \frac{\mathrm{CCDF}(d_{u,v})}{|V| - 1}, \tag{2}$$

where V is the set of all users in the network. $\omega(u)$ reflects the local population density around user u, that is, the number of nearby users, and represents the baseline level of geographic homophily that arises solely from random connectivity. Therefore, the difference between the observed value $\mathrm{H}(u)$ and $\omega(u)$ contains the inherent geographic preference that cannot be explained by population distribution alone.

Following the definition of inbreeding homophily [14]–[16], we define the corrected geographic homophily (inbreeding homophily) normalized by the expected value based on population distribution as follows:

$$\mathrm{IH}(u) = \frac{\mathrm{H}(u) - \omega(u)}{1 - \omega(u)} \tag{3}$$

This normalization ensures that the maximum value of $\mathrm{IH}(u)$ is 1. When $\mathrm{IH}(u) = 0$, it indicates that the user's connections are equivalent to randomly selecting users (i.e., no geographic preference). When $\mathrm{IH}(u) > 0$, it indicates a tendency to prefer geographically close users, and when $\mathrm{IH}(u) < 0$, it indicates a tendency to prefer distant users. This formulation enables a fair comparison of users' geographic preferences across areas with different population densities, such as urban and rural areas.

We measured the three geographic homophily indices for each user, and the distributions are shown in Figure 1. The left panel represents the proposed inbreeding homophily index $\mathrm{IH}(u)$, the center shows the observed homophily index $\mathrm{H}(u)$, and the right panel shows the baseline homophily index $\omega(u)$. The number of histogram bins was set to 40. The horizontal axis indicates the homophily index value, and the vertical axis indicates the number of users in each bin for each country. The red dashed line represents $\mathrm{H}(u) = \omega(u)$, or equivalently $\mathrm{IH}(u) = 0$, which we refer to as the baseline level. If users follow others randomly without geographic preference, the distribution of homophily is expected to be centered around this baseline level. When the index takes positive values, it indicates that users are connected with geographically close users more than expected under the null model, suggesting a preference for nearby users. Conversely, when it takes negative values, it indicates a tendency to connect with more distant users, demonstrating the formation of long-distance oriented connections. The Spearman's rank correlation coefficients between $\mathrm{IH}(u)$ and $\mathrm{H}(u)$ ranged from 0.88 to 0.99 across all countries, confirming a strong consistency.

IV. Analysis

Using the proposed geographic homophily index, we analyzed the degree of geographic preference in online social networks while accounting for population bias. We also examined how the relationship between geographic preference and population size varies across ten countries.

A. RQ1: Does geographic preference exist after correcting for population bias?

We proposed a geographic homophily index that corrects for population bias. We investigate whether geographic preferences can still be observed on online social networks when using this metric. We calculated three geographic homophily measures for all users in each country's social network. The results are shown in Figure 1.

As shown in Figure 1, the distribution of inbreeding homophily was centered in the positive range across all 10 countries, with most users exceeding the baseline level. This indicates that even after accounting for population heterogeneity, many users still tend to follow geographically close others. A small number of users had $\mathrm{IH}(u) = 1$, meaning that all of their mutual following connections were located in the same area (i.e., zero distance). Conversely, a small fraction of users exhibited negative values, indicating that they formed ties with users located farther away than expected under the null model. These results consistently demonstrate that a preference for geographically proximate connections exists on Twitter mutual following networks, even after correcting for spatial population bias.

B. RQ2: Is geographic preference stronger in more populated areas?

We examined the correlation between population size and geographic preference. Previous research [11] reported positive correlations between the average geographic homophily of each area and their population size. However, users in highly populated areas may have higher observed homophily because they are surrounded by more nearby users, even when connections are random. To properly evaluate the relationship between population size and geographic preference while correcting for such bias, we analyzed the correlation between the mean inbreeding homophily of users located in each area and the user population. Here, population data are required to compare the degree of geographic preference across different population levels. Instead of actual population data, we used

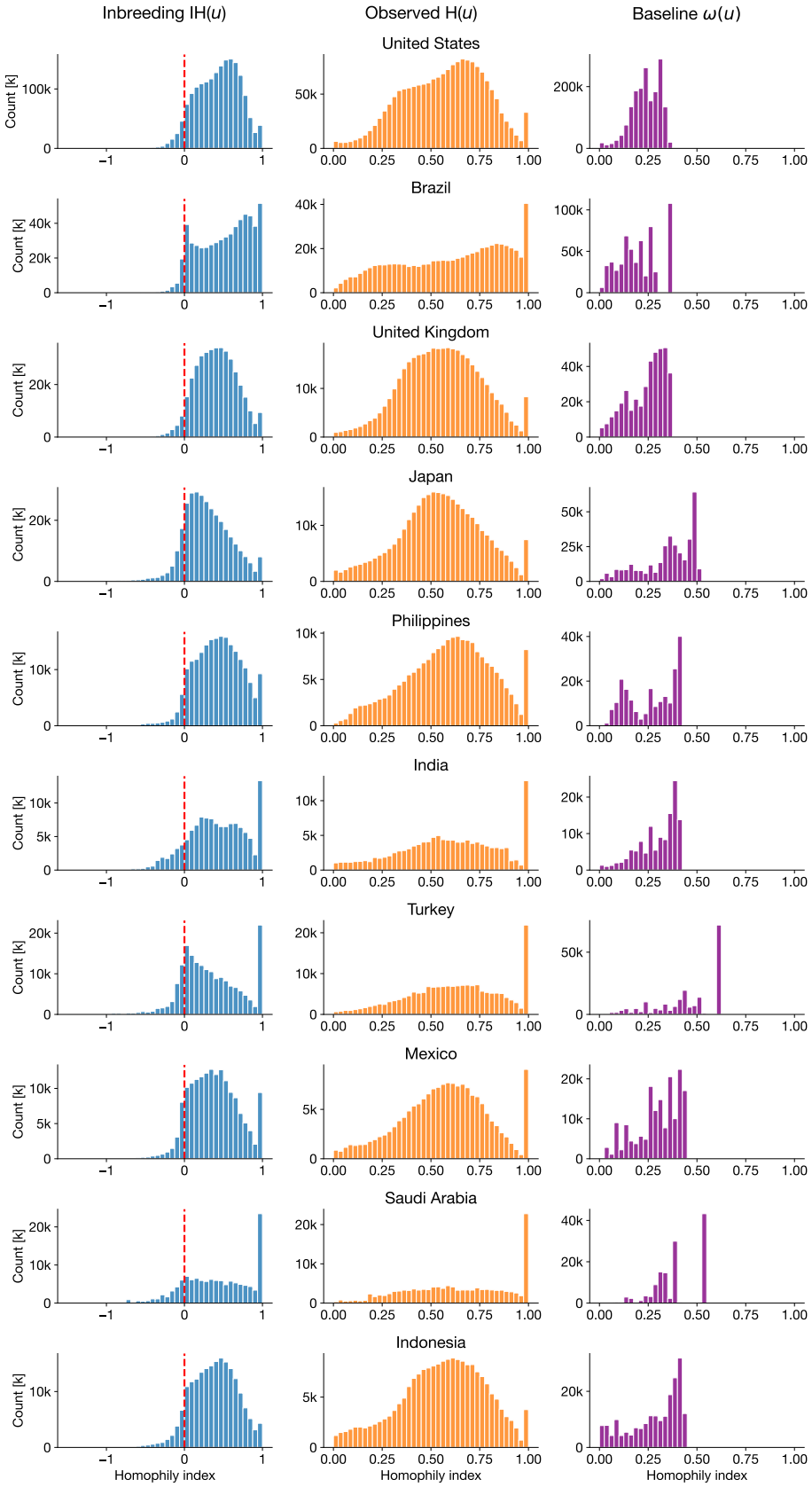


Fig. 1. Distributions of inbreeding $\mathrm{IH}(u)$, observed $\mathrm{H}(u)$, and baseline $\omega(u)$ homophily indices across countries. The red dashed line denotes the baseline level ($\mathrm{IH}(u) = 0$). Positive deviations indicate a preference for geographically close connections, while negative deviations indicate a tendency toward long-distance connections.

the number of users of each area in the dataset. While this approach may introduce bias in population estimation due to varying social media adoption rates, it is assumed to reasonably reflect the actual population ranking of each area.

Figure 2 shows the relationships between area population (on the x-axis, in logarithmic scale) and the average values of inbreeding homophily (left), observed homophily (center), and baseline homophily (right). Spearman's rank correlation coefficients are shown as "Corr." in the lower-right corner of each panel. The correlations between the mean inbreeding homophily of each area and their user populations varied across countries.

Moderate positive correlations between population size and mean inbreeding homophily were observed only in four countries: Brazil, Saudi Arabia, the Philippines, and Mexico. In these countries, users in more populated areas tend to exhibit stronger geographic preference. Among the remaining six countries, weak negative correlations ($\mathrm{Corr.} < -0.2$) were observed in Japan, India, and Turkey, whereas the United Kingdom, the United States, and Indonesia showed little or no correlation ($-0.2 < \mathrm{Corr.} < 0.2$). Thus, in these countries, the strength of geographic preference does not increase with population size.

In summary, a positive relationship between population size and geographic preference was observed only in a subset of countries. In other countries, the relationship was weak, absent, or even reversed, suggesting that the link between urbanization and geographic preference is not universal but context-dependent across countries.

V. Discussion

The correlation between the mean baseline homophily index and the population size of each area suggests that structural bias in the observed homophily indices arises from population distribution. As shown in Figure 2, positive correlations were observed in nine out of ten countries (except Mexico) between the mean baseline homophily and area population. This result is consistent with the design expectation that, in more densely populated regions, random connections are more likely to occur between geographically close users, resulting in higher baseline homophily values. These findings indicate that observed homophily index of geographic homophily may be structurally biased toward higher values in densely populated urban areas.

The proposed inbreeding homophily index effectively reduces this bias while preserving the overall trend of the observed geographic homophily. As shown in Figure 2, the correlations between the mean inbreeding homophily and population size were consistently weaker than those of the observed homophily. Across all countries, the correlations became substantially weaker and in six countries they approached zero or turned negative. The Spearman rank correlation coefficients between $\mathrm{IH}(u)$ and $\mathrm{H}(u)$ remained very high, indicating that the correction maintains consistent ranking among users. These results suggest that the previously reported positive relationship, in which areas with larger populations show higher homophily, may have been largely a reflection of population structure. By correcting for baseline expectations, the proposed index enables a more accurate assessment of geographic preference that reflects actual user behavior rather than demographic imbalance.

To further examine how population bias manifests spatially, we visualized the three homophily indices on maps for each country. Figure 3 illustrates the spatial distribution of the mean baseline, observed, and inbreeding homophily indices in Japan as a representative example. Comparing these maps allows us to assess how population structure influences the apparent geographic homophily and how effectively the proposed correction mitigates this bias. In most countries, baseline homophily was higher in central or metropolitan regions and lower in peripheral areas. This pattern reflects geographical constraints. In dense urban centers, even random connections tend to be short range, whereas in peripheral regions the limited number of neighboring areas reduces such opportunities. Observed homophily exhibited patterns similar to baseline homophily in many countries, suggesting a strong influence of population structure. In contrast, after correction, inbreeding homophily showed much weaker regional variation, indicating that the proposed measure effectively removes population bias and captures genuine geographic preference.

In contrast to Japan, where spatial variation in homophily was limited, Brazil showed clear regional differences, as illustrated in Figure 4. The southern regions, such as Paraná and Rio Grande do Sul, exhibited relatively high inbreeding homophily despite moderate population sizes, suggesting stronger local cohesion beyond population effects. Overall, homophily tended to increase with population size, yet this pattern was not uniform across major cities. We also plotted a detailed map focusing on the São Paulo and Rio De Janeiro area (Figure 5). This figure visualizes the average inbreeding homophily, where the circle size indicates the local user population, and the color represents the mean inbreeding homophily. This close-up view reveals that Rio de Janeiro shows higher inbreeding homophily than its surroundings, whereas São Paulo, the nation's largest and most industrialized city, exhibits lower values even in its dense urban core. These results indicate that Brazil's geographic preferences are shaped by regional social and cultural contexts.

VI. Conclusion

This study proposed a population-bias-corrected measure of geographic inbreeding homophily to evaluate users' spatial preferences in online social networks. Using large-scale Twitter mutual-following networks from ten countries, we demonstrated that users in all countries tend to form geographically proximate ties more frequently than expected under random connection, indicating the persistence of geographic preference even in online settings. However, when comparing areas within each country, the correlation between user population and mean inbreeding homophily varied substantially across countries. Positive correlations were observed only in a subset of cases, such as Brazil and the Philippines, while others

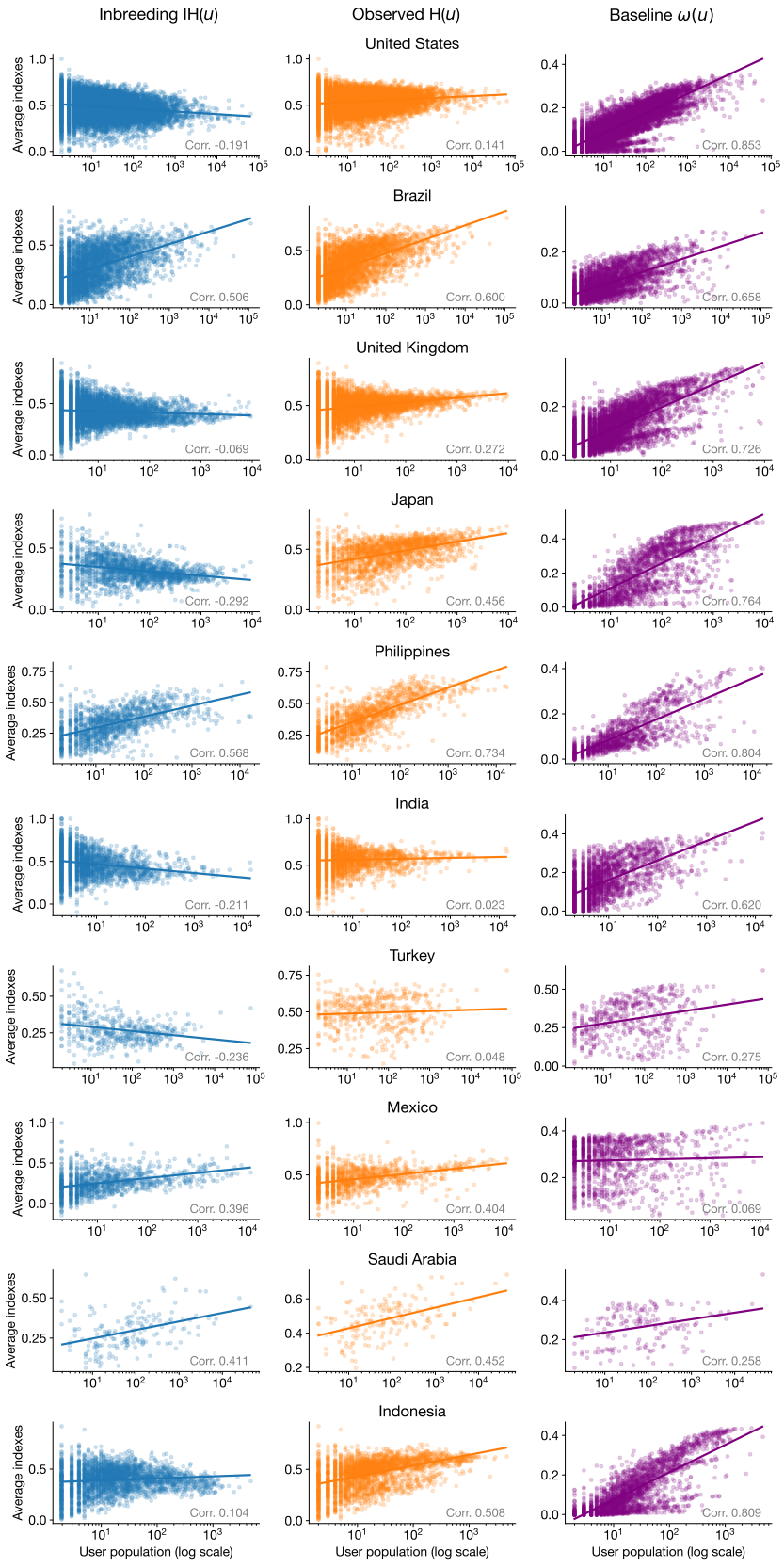
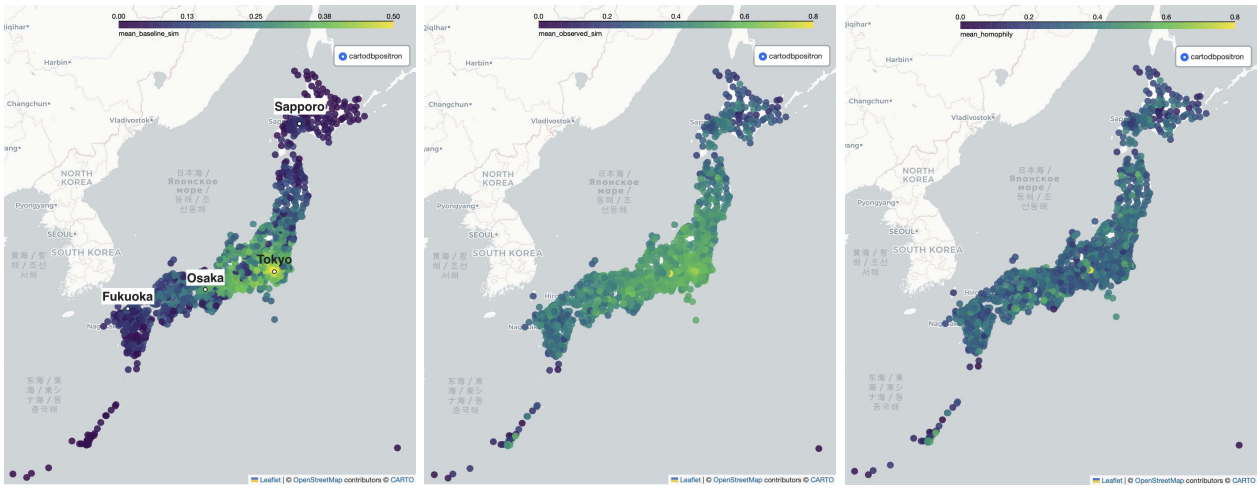
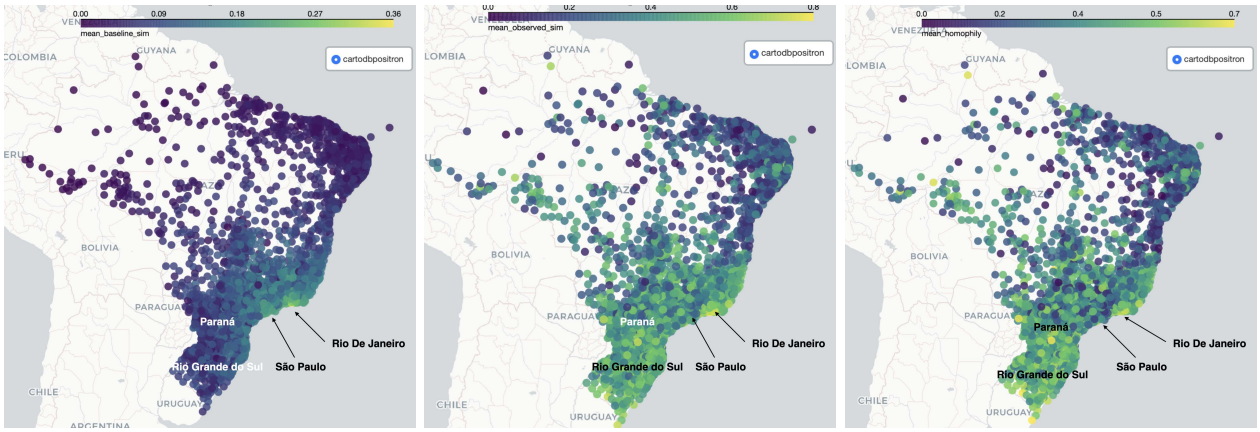


Fig. 2. Correlations between population size and the average degrees of geographic homophily indices in each country. The straight lines indicate regression fits on a logarithmic scale. After adjusting the observed homophily by its expected baseline value, the correlations with area population weakened, suggesting that the previously observed positive associations were largely driven by population bias.



(a) Mean baseline homophily (b) Mean observed homophily (c) Mean inbreeding homophily

Fig. 3. Comparison of three types of homophily maps in Japan.



(a) Mean baseline homophily (b) Mean observed homophily (c) Mean inbreeding homophily

Fig. 4. Comparison of three types of homophily maps in Brazil.

showed weak or even negative associations, including Japan, India, and Turkey. These results suggest that the commonly reported trend that users in densely populated urban areas form more geographically close relationships is not universal once population bias is corrected. By accounting for spatial population heterogeneity, the proposed index enables fairer cross-regional comparisons and reveals how geographic preferences are shaped by local social and cultural contexts. Future research could extend this framework to explore the interplay between geographic preference and socioeconomic or linguistic diversity and to examine how such spatial tendencies evolve over time.

REFERENCES

[1] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," *Proceedings of the National Academy of Sciences*, vol. 102, no. 33, pp. 11623–11628, 2005.

[2] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th international conference on World wide web*, pp. 61–70, 2010.

[3] D. Mok and B. Wellman, "Did distance matter before the internet?: Interpersonal contact and support in the 1970s," *Social Networks*, vol. 29, no. 3, pp. 430–461, 2007. Special Section: Personal Networks.

[4] P. V. Marsden, "Homogeneity in confiding relations," *Social Networks*, vol. 10, no. 1, pp. 57–76, 1988.

[5] D. J. Brass, "Men's and women's networks: A study of interaction patterns and influence in an organization," *The Academy of Management Journal*, vol. 28, no. 2, pp. 327–343, 1985.

[6] P. V. Marsden, "Core discussion networks of americans," *American Sociological Review*, vol. 52, no. 1, pp. 122–131, 1987.

[7] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[8] K. Z. Khanam, G. Srivastava, and V. Mago, "The homophily principle in social network analysis: A survey," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 8811–8854, 2023.

[9] E. Gilbert, K. Karahalios, and C. Sandvig, "The network in the garden: An empirical analysis of social media in rural life," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1603–1612, 2008.

[10] H. Nilforoshan, W. Looi, E. Pierson, B. Villanueva, N. Fishman, Y. Chen, J. Sholar, B. Redbird, D. Grusky, and J. Leskovec, "Human mobility

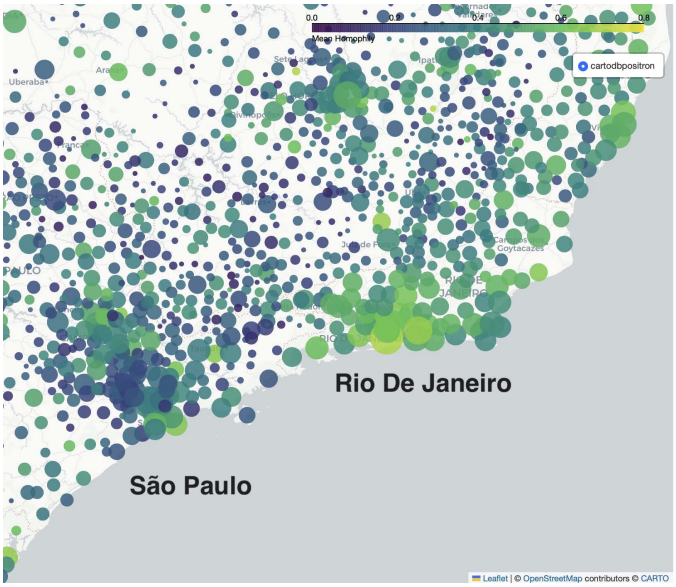


Fig. 5. Mean inbreeding homophily maps in São Paulo and Rio De Janeiro

networks reveal increased segregation in large cities," *Nature*, vol. 624, no. 7992, pp. 586–592, 2023.

[11] M. Ushiba, S. Hironaka, M. Yoshida, and K. Umemura, "Large-Scale Analysis of Rural-Urban Differences in Geographic Homophily," in *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 372–377, 2023.

[12] S. Hironaka, M. Yoshida, and K. Umemura, "Comparison of Indicators of Location Homophily Using Twitter Follow Graph," in *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications*, pp. 1–6, 2021.

[13] S. Hironaka, M. Yoshida, and K. Umemura, "Cross-Country Analysis of User Profiles for Graph-Based Location Estimation," *IEEE Access*, vol. 9, pp. 168831–168839, 2021.

[14] J. S. Coleman, "Relational Analysis: The Study of Social Organizations with Survey Methods," *Human Organization*, vol. 17, no. 4, pp. 28–36, 1958.

[15] S. Currarini, M. O. Jackson, and P. Pin, "An Economic Model of Friendship: Homophily, Minorities, and Segregation," *Econometrica*, vol. 77, no. 4, pp. 1003–1045, 2009.

[16] D. Zuckerman, "Multidimensional homophily," *Journal of Economic Behavior & Organization*, vol. 218, pp. 486–513, 2024.