An evaluation of geo-located Twitter data for measuring human migration

Junjun Yin\textsuperscript{a}, Yizhao Gao\textsuperscript{b,c} and Guangqing Chia\textsuperscript{a,d}

\textsuperscript{a}Social Science Research Institute and Population Research Institute, The Pennsylvania State University, University Park, PA, USA; \textsuperscript{b}CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign, Urbana, IL, USA; \textsuperscript{c}Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA; \textsuperscript{d}Department of Agricultural Economics, Sociology and Education, The Pennsylvania State University, University Park, PA, USA

ABSTRACT
This study evaluates the spatial patterns of flows generated from geo-located Twitter data to measure human migration. Using geo-located tweets continuously collected in the U.S. from 2013 to 2015, we identified Twitter users who migrated per changes in county-of-residence every two years and compared the Twitter-estimated county-to-county migration flows with the ones from the U.S. Internal Revenue Service (IRS). To evaluate the spatial patterns of Twitter migration flows when representing the IRS counterparts, we developed a normalized difference representation index to visualize and identify those counties of over-/under-representations in the Twitter estimates. Further, we applied a multidimensional spatial scan statistic approach based on a Poisson process model to detect pairs of origin and destination regions where the over-/under-representativeness occurred. The results suggest that Twitter migration flows tend to under-represent the IRS estimates in regions with a large population and over-represent them in metropolitan regions adjacent to tourist attractions. This study demonstrated that geo-located Twitter data could be a sound statistical proxy for measuring human migration. Given that the spatial patterns of Twitter-estimated migration flows vary significantly across the geographic space, related studies will benefit from our approach by identifying those regions where data calibration is necessary.

1. Introduction

Human migration is an important and complex social phenomenon. Accurate estimation of human migration flows typically requires large-scale censuses and surveys but these data do not permit timely estimation of migration flows, such as county-to-county migration patterns. With its increasing popularity, location-based social media...
provides a tremendous opportunity for studying human migration at a fine-grained spatial scale. Twitter provides a publicly accessible geo-located data that can continuously track the movements of a large group of users over time and across space. Demographers have already used geo-located Twitter data to track population movements, suggesting that Twitter data could be used to infer migration patterns (Zagheni et al. 2014). Complex and time-sensitive studies of mobility activities can also benefit from using such data. For example, geo-located Twitter data have been used to explore human mobility dynamics during the COVID-19 pandemic (Huang et al. 2020) and estimate refugee migration patterns following natural disasters (Hübl et al. 2017).

The validity of using geo-located Twitter data for estimating long-term migration remains unclear because Twitter users are not currently representative of the population (Kim et al. 2013). Almost all the existing migration studies that use Twitter data document this limitation (Weber et al. 2014, Zagheni et al. 2014, Jurdak et al. 2015, Fiorio et al. 2017, Hübl et al. 2017). A direct consequence is that migration flows estimated from geo-located tweets can significantly over- or under-represent reality. For example, the Twitter-estimated migration flows between two regions with dense populations may better reflect the real-world situation than those between two regions with sparse populations. If using geo-located Twitter data can be proven to provide a useful measure of human migration events, the potential applications can be significant. The availability and release time of official migration statistics varies dramatically across countries but often lags by at least a year. The spatial units of the migration flows are bound to certain administrative units, such as counties or states. For example, in the observed population of migrants from the U.S. Internal Review Service (IRS) tax return data, the smallest spatial unit of migration flows generated is a county, and the latest available data as of January 2022 is for 2018–2019. Despite these issues, this data is considered to be a gold standard and is used by the U.S. Census Bureau in its estimates of county-to-county annual migration flows (DeWaard et al. 2018). As geo-located Twitter data has worldwide coverage and can be obtained continuously in near real-time, it has the possibility of providing more timely and fine-grained migration flows across regions and countries. Few studies have yet to examine the detailed spatial patterns of the representativeness of Twitter-estimated migration flows or to quantitatively model and assess the spatial patterns.

This paper aims to evaluate the spatial patterns of the representativeness of Twitter-estimated migration flows by comparing county-to-county migration flows of Twitter users in the U.S. with the IRS county-to-county annual migration flows. The county-to-county migration flows of Twitter users are estimated by identifying Twitter users whose residency county changed over one year. The initial results suggest that geo-located Twitter data could provide a sound statistical proxy for measuring human migration. The results show that the aggregated county inbound and outbound migration flows estimated from Twitter data have a high correlation with the IRS estimates (with $R^2 > 0.79$ and $R^2 > 0.88$, respectively), which suggest that the Twitter-estimated migration flows using our method do provide a reasonably good prediction of the IRS estimates at the county level.
To evaluate how the Twitter-estimated migration flows are representative of the IRS estimates across the geographic space, we first developed a normalized difference representation index to visualize and identify counties where Twitter-estimated migration flows over- or under-represent the IRS counterparts. We applied a multidimensional spatial scan statistic based on a Poisson process model to detect the pairs of origin and destination regions where the over- or under-representativeness occurred. The existence, spatial extents, and statistical properties of these region pairs, which are represented by statically significant high and low spatial clusters, describe how well Twitter can capture the spatial patterns of the IRS estimates. This approach provides important and detailed insights into the spatial patterns of Twitter-estimated migration flows. The developed approach can assist future migration research using geo-located Twitter data by identifying the regions where data calibration is necessary.

The remainder of this paper is organized as follows. Section 2 reviews the published literature related to human migration patterns using geo-located Twitter data. Section 3 describes (1) the detailed methods for extracting county-to-county migration flows using geo-located Twitter data (2) a normalized difference representation index to visualize and identify counties where Twitter migration flows over- or under-represent the IRS counterparts, and (3) a multidimensional spatial scan statistic used to detect the pairs of origin and destination regions where the over- or under-representativeness occurred. Section 4 presents the analytical results. Section 5 concludes the paper by discussing the remaining issues and future directions.

2. Background and related work

2.1. Geo-located twitter data for capturing human movements

Migration statistics are estimated in several ways including, longitudinal residential address registries, population stocks, and census surveys (Jensen 2013). However, these data sources can be inconsistent over time, are collected by different agencies, and are not measured consistently. For example, two major U.S. Census Bureau surveys collect migration statistics: the American Community Survey (ACS) and the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS), but their migration rates differ at the national, state, and county levels (Kaplan and Schulhofer-Wohl 2012). In terms of international migration statistics, the data comparability can be an issue considering that different countries have different approaches to performing migration estimations (Willekens et al. 2016). Additionally, the data collection process is labor-intensive and expensive, and the release of migration statistics is often delayed by a year or more. Consequently, researchers have been advocating for using more flexible supplementary data sources (Smith and Leigh 1997).

With the increasing availability of digital geo-located data sources, such as mobile phone call data (Blumenstock 2012), geo-located Twitter data (Hawelka et al. 2014, Zagheni et al. 2014), web log-in/IP records (Zagheni and Weber 2012), and Internet search queries (Lin et al. 2019), researchers have started to use these data sources for population and migration studies. For example, researchers have used mobile phone data for dynamic population mapping and estimation (Deville et al. 2014) and have
performed near-real-time assessments of population displacement following disasters (Wilson et al. 2016). By continuously observing the movements of individuals using digital geo-located data, studies have examined short-term human mobility patterns (González et al. 2008) based on mobile phone data and event-oriented travel patterns based on geo-located Twitter data (Xin and MacEachren 2020). However, mobile phone data is (1) difficult to access (2) only available from certain service providers in a few cities or countries, and (3) rarely covers long-term observations.

Geo-located Twitter data may serve as a good compromise. It is publicly accessible and has nearly comprehensive global coverage (Hawelka et al. 2014, Weber et al. 2014, Belyi et al. 2017). Geo-located Twitter data are becoming increasingly popular to study both short-term mobility and long-term migration patterns. Geo-located Twitter data are regular tweets tagged with real-world locations derived from Twitter users’ smartphones which have integrated GPS or Wi-Fi positioning. Note that geotagged tweets make up only a small portion of the overall Twitter data volume. Around 3% of all tweets worldwide were geotagged in 2012 (Leetaru et al. 2013). With the capability of continuously monitoring a group of Twitter users, this data offers great potential for short-term and long-term human behavior studies. Complex and time-sensitive studies of migration activities can also benefit from such data. For example, several studies have used geo-located Twitter data to explore migration flow patterns (Weber et al. 2014), with applications ranging from integrating the dimensions of internal and international migration (Zagheni et al. 2014), to estimating refugee migration patterns (Hüb l et al. 2017), and testing existing migration theories, such as the relationships between short-term mobility and long-term migration (Fiorio et al. 2017). Some studies have suggested that Twitter data can serve as a barometer for migration flows, providing timely migration information before the availability of official statistics (Zagheni et al. 2014).

However, one major drawback of Twitter data (and indeed, of all social media data), is that Twitter users are not representative of the population (Kim et al. 2013). When using social media data for population-related research, it is important to generalize the results to the whole population, but this is difficult for Twitter data. When researchers collect Twitter data, sampling is controlled by Twitter, and little is known about its sampling methodology. Certain demographic groups are known to be over- or under-represented in Twitter data. Researchers have found that Twitter users skew towards young, urban, minority individuals (Mislove et al. 2011). Other research suggests that social media users tend to be younger (Nguyen et al. 2016), socioeconomically advantaged (Duggan and Smith 2014, Jones et al. 2016), and more likely to be male, leaving the results biased and invalid without any calibration (Yildiz et al. 2017). Age bias was found to affect attempts to predict political elections from Twitter sentiments (Gayo-Avello et al. 2011). Almost all the studies that use Twitter data for migration study purposes also documented this limitation (Weber et al. 2014, Zagheni et al. 2014, Jurdak et al. 2015, Fiorio et al. 2017, Hüb l et al. 2017).

Multiple researchers have begun addressing the sampling bias issues by proposing methods to estimate the characteristics of Twitter user demographics, such as gender and race/ethnicity (Longley et al. 2015, Longley and Adnan 2016, Luo et al. 2016, Yildiz et al. 2017), as well as the representativeness in the geographic distribution of
Twitter user demographics (Yin et al. 2018). Our study focuses on the spatial patterns of the representativeness of Twitter-estimated migration flows over the established migration estimates and is carried out in the following steps. First, we aim to establish that geo-located Twitter data could be a useful statistical proxy for measuring human migration. Second, we illustrate the spatial patterns of counties where Twitter-estimated migration flows over- or under-represent the IRS counterparts. Third, we propose a multidimensional spatial scan statistic based on a Poisson process model to quantitatively model and identify the pairs of origin and destination regions where the over- or under-representativeness occurred.

3. Materials and methods

3.1. Geo-located Twitter data and data processing

The geo-located tweets used in this study were collected continuously from 01 January 2013 to 31 December 2015 using the publicly accessible Twitter Streaming API (https://developer.twitter.com/). Spatially, data collection includes all 50 states and Puerto Rico. To ensure it does not exceed the 1% policy data quota in the Twitter data stream as mentioned in Hawelka et al. (2014), several sub-regions are created, and data collection is performed simultaneously at each. The information fields of each data record include user ID, location in the form of latitude/longitude, timestamp, and other ones that are irrelevant to this study. We examined the ‘geo’ attribute in each raw tweet and kept the ones with location information derived from GPS recording rather than from geocoding. By filtering out tweets with geo-locations outside of the study boundaries, the data collection contains over 733 million tweets in 2013, 887 million in 2014, and 265 million geo-located tweets in 2015, which are collected from approximately 6.3 million (2013), 6.5 million (2014), and 4.7 million (2015) Twitter users in the U.S. Due to changes Twitter made to geo-locations on tweets in the second half of 2015, there was a noticeable decrease in the total number of geo-located tweets collected in 2015. Nevertheless, the geo-located tweets collected in all three years have reasonable spatial coverage over the entire U.S. See Supplement Materials S1 for an example of the locations of all the collected tweets in the U.S. from 2015. The collective point visualization reveals the geography of human settlements as clusters with higher densities of tweets correspond to the locations of major cities. The overall spatial coverage of the collected geo-located tweets provides strong evidence for using such data sources to investigate migration patterns at the national scale.

With location and time information in each tweet, we reconstruct the location history of individuals by sorting the temporal sequence in chronological order. Each location in a user’s location history is mapped to a corresponding county. Geo-located tweets may be generated by non-human Twitter users (e.g. bots). Because we are interested in the county-to-county migration patterns stationary bot tweets were automatically excluded. To minimize the potential impact of relocating bot users being included in the migration estimates, we examined all the consecutive locations in each user’s location history and excluded those with relocating speeds over the threshold of 240 m/s, as suggested by previous research (Hawelka et al. 2014, Jurdak
et al. 2015). To ensure the migration patterns of residents rather than tourists, we imposed a strict condition that a user must be observed to have stayed in the U.S. more than 30 days a year as described in Yin et al. (2017). Although we defined this temporal constraint subjectively, it does provide a strict criterion for selecting a stable Twitter user population.

### 3.2. U.S. County-to-County migration flows

Currently, there are two major data sources for providing migration flow estimates at the county level in the U.S.: The 5-year average county-to-county migration flow estimates from the ACS and the yearly estimates from the IRS. As our study focuses on annual migration, we chose the IRS annual estimates. The IRS migration flow estimates are based on year-to-year changes of residence as reported on individual income tax returns and are considered a gold standard for studying migration patterns in the U.S. (DeWaard et al. 2018). IRS migration estimates are available by state or by county for the entire U.S. from 1991 to 2019. The data are presented as inflows (the number of tax filers moved to a state or county and from where they moved) and outflows (the number of tax filers moved from a state or county and to where they moved). IRS estimates include two additional fields of information: (1) number of tax returns filed, and (2) number of personal exemptions claimed. Because the first is a direct reference to the migrants that filed tax returns and the values of the two fields are highly positively correlated, we used the number of tax returns filed as the IRS migration estimates.

In regards to our Twitter data collection, this study uses the 2013–2014 and 2014–2015 IRS estimates, which aims to ensure validity with a longer time range. A total of 3141 U.S. counties are included in this study. There were a total of 46,128 county-to-county pairs with 5,379,236 migration events from the 2013 to 2014 estimates and 35,770 county-to-county pairs with a total of 4,065,923 migration events from the 2014 to 2015 estimates.

### 3.3. Estimation of county-to-County migration flows

In the IRS county-to-county migration estimates, a migration event occurs when an individual's county-of-residence in the current filing year differs from the previous year. To generate county-to-county Twitter user migration estimates, we needed to assign each Twitter user with a county-of-residence. Indeed, home location identification is one of the fundamental steps for studying human mobility using digital footprint data (locations obtained from GPS logs, geo-located social media, and mobile phone positioning). The applied methods vary and include spatial clustering algorithms (Lin and Cromley 2018), linking tweet locations to individual land use parcels (Yin and Chi 2021) for geo-located tweets, inferring home locations of individuals by leveraging Twitter message content with a machine learning-based classifier (Mahmud et al. 2014) and performing sentiment analysis (Mostafa et al. 2022). These methods are computation-intensive and are difficult to apply to a large collection of tweets with national coverage. Since the spatial unit in this study is a county, a Twitter user's
county-of-residence is defined by using the most frequently tagged county as a user’s county-of-residence (McNeill et al. 2017, Jiang et al. 2018). The resident county of each user is identified for each year; 3.2 million in 2013, 3.8 million in 2014, and 2.5 million in 2015.

Although studies have shown the most tweeted locations are homes or workplaces (Soliman et al. 2015, Yin and Chi 2021), many situations can affect the accuracy of assigning county-of-residence. For example, the number of tweets in a year for defining the county-of-residence varies significantly among Twitter users. Figure 1 shows a log-log plot of the probability density function (PDF) for the number of tweets in each user’s most tweeted county. The PDFs in all three years are heavy-tailed indicating strong variation in Twitter users’ behavior.

Therefore, we conducted a sensitivity test to find a suitable threshold number (i.e. the number of minimum tweets needed) to assign a Twitter user’s county-of-residence. First, the threshold numbers are set to 1 (no restriction), 5, 10, 15, 20, 30, 40, and 50. We then assigned a county-of-residence to each Twitter user for each year if the maximum number of tweets in a county is greater or equal to the threshold numbers. Further, a migration event is determined when the resident county changes in the following year. An origin-destination (OD) county migration matrix was generated based on the collective Twitter user changes at the U.S. county level. Based on this definition, the county-to-county migration inflows and outflows of Twitter users are estimated. We then tested how Twitter-estimated migration flows generated by different threshold values are correlated to IRS estimates. In terms of pairwise county-to-county migration flows, Spearman’s rank-order correlation test, which is a nonparametric measure of association, suggests that the two flows are not correlated. By aggregating the migration flows to each county as inbound and outbound migration flows, Spearman’s rank-order correlation test suggests that the Twitter migration flows are highly correlated with the IRS estimates. For example, for county outbound

\[ \text{Figure 1. Log-log plot of the probability density function for the number of tweets in each Twitter user’s most tweeted county in a year from 2013 to 2015.} \]
migration flows between 2013–2014 and 2014–2015, the coefficient of determination ($R^2$) value from the linear regression between the two estimates when using each threshold value is shown in Table 1 (The same table for county inbound migration flows is shown in Supplement Materials, S2).

As the value of the minimum number of tweets required for assigning a Twitter user’s county-of-residence increases, the $R^2$ value increases (Table 1). At the same time, the number of common pairwise county-to-county migration flows between Twitter estimates (denoted as Twitter OD pairs) and the IRS estimates (denoted as N) decrease. As it is shown in Figure 1, it also results in a decrease in the number of Twitter users qualified for the study (see Supplement Materials, S2). Considering that there are 46,128 and 35,770 county-to-county pairs in the 2013–2014 and 2014–2015 IRS datasets, we chose 30 as the threshold value as the $R^2$ approaches the highest while at the same level the number of Twitter OD pairs is comparable to the IRS estimates. In other words, we have determined that a Twitter user’s county-of-residence can only be assigned if his/her most tweeted county has at least 30 tweets.

### 3.4. Migration pattern comparison

To evaluate the degree to which the migration events estimated by using geo-located Twitter data reflect actual human migration activity, we examine the correlation of the county-to-county migration flows (both inbound and outbound flows) between the IRS and Twitter estimates. Notably, many OD county pairs do not have migration flows. Given 3,141 U.S. counties in this study, there should be 9,862,740 possible OD county pairs, but only several hundred thousand pairs have migration flows from both data sources. There are also instances where OD county pairs with migrants exist in the IRS estimates but do not exist in the Twitter estimates and vice versa. In this study, all county-to-county migration flows are kept as long as they exist in either of the data sources.

To identify the counties that Twitter-estimated migration flows under- or over-represent the IRS estimates, we aggregated the flows to each county as inbound and outbound migration flows. Further, we evaluated such representativeness using a normalized difference representation index ($R$), which is defined as follows (Eq. 1):

\[
R = \frac{(X - X_{min})}{(X_{max} - X_{min})}
\]
\[ R_j = \left( \frac{T_j}{T_{all}} - \frac{I_j}{I_{all}} \right) / \left( \frac{T_j}{T_{all}} + \frac{I_j}{I_{all}} \right) \]

where \( T_j \) denotes the number of (outbound or inbound) Twitter-estimated migration flows in county \( j \), \( I_j \) denotes the number of (outbound or inbound) IRS migration flows in county \( j \), \( T_{all} \) denotes the number of total Twitter migration flows, and \( I_{all} \) denotes the total IRS migration flows of the U.S. A value of \( R_j \) equal to 0 indicates that the percentage of Twitter migration from/to the county \( j \) is proportional to the national average of the IRS estimates. In particular, the representation index is effective in illustrating the availability of migration flows (i.e. a value of \( R_j \) equal to 1 or \(-1\) suggests the absence of migration flows from IRS or Twitter estimates). A value of \( R_j \) which approaches 1 or \(-1\) indicates an over-representation or an under-representation of Twitter-estimated migration from/to the county \( j \).

The representation index provides an overall view of the spatial distribution of Twitter-estimated migration flows representing the IRS estimates. However, the spatial patterns are evaluated by fixing on migration flows aggregated to (i.e. origin or destination) counties. The aggregation completely misses information regarding the origin counties of the inbound flows or the destination counties of the outbound flows. In other words, it is impossible to learn how different counties contribute to the representation index.

### 3.5. Multidimensional spatial scan statistics

When a migration flow is directionally defined by its origin and destination (OD), comparing spatial patterns of two sets of pairwise OD flows is challenging (Gao et al. 2018). It is intuitive to compare spatial patterns of two OD flows as aggregated to/from certain areal units, such as aggregating migration flows to counties as mentioned in the previous section. Yet, due to the process of aggregation, it is impossible to trace the origins of inbound flows or the destinations of outbound flows, which may have contributed to the differences. To address this issue, Gao et al. (2018) proposed a multidimensional spatial point data model to integrate the origin and destination of a migration flow into a single analysis unit, where each migration flow is modeled as a spatial point in a 4D OD space. Specifically, a migration flow \( M_i = \langle S_{oi}, S_{di} \rangle \) represents a migrant that moves from an origin \( S_{oi} = \langle x_{oi}, y_{oi} \rangle \) to a destination \( S_{di} = \langle x_{di}, y_{di} \rangle \), where \( x \) and \( y \) are the coordinates of the locations. In this case, both \( S_{oi} \) and \( S_{di} \) are 2D geographic points. To represent \( M_i \) as an integrated analytical unit, \( M_i \) is modeled as a spatial point \( P_{Mi} = \langle x_{oi}, y_{oi}, x_{di}, y_{di} \rangle \) in a 4D OD space, which is the Cartesian product of the origins’ 2D geographic space and the destinations’ 2D geographic space \( S_{oi} \times S_{di} \). Therefore, the concepts and methods for 2D spatial point pattern analysis can be employed to compare migration flows.

To evaluate the migration flows at a fine spatial scale, this study utilizes the annual IRS county-to-county migration flow estimates as the ground truth (i.e. an approximation of real migration flows). The null hypothesis is that the estimated county-to-county migration flows of Twitter users are a fully accurate representation of the real migration flows (i.e. the IRS estimates). To detect the differences between the two migration flow distributions against this null hypothesis, we must find pairs of OD
regions where there is a statistically significant high/low existence of one type of migration flows over the other. When both migration flows are modeled as 4D points using a multidimensional spatial scan statistics approach, those regions correspond to the high/low clusters in the 4D OD space (Gao, Li, et al. 2018).

A multidimensional spatial scan statistics approach involves the following steps. First, a large collection of 4D scanning windows is generated in the study area. Second, the number of 4D points representing the migration flows in each scanning window is calculated. Third, the maximum likelihood of each scanning window to be a cluster (high or low), \( L_W \), and the maximum likelihood under the null hypothesis with no clusters, \( L_{W0} \), are calculated according to the data model. The scan statistics \( \lambda \) of a scanning window is calculated as the likelihood ratio (Eq. 2). Fourth, the primary cluster is detected as the scanning window with the highest likelihood ratio. The secondary clusters are detected as the scanning windows with the highest likelihood ratio that do not intersect existing clusters. Finally, a Monte Carlo simulation is performed to obtain the statistical significance of each detected cluster.

\[
\lambda = \frac{L_{W0}}{L_W}
\]  

**3.6. Poisson process model**

The Bernoulli model developed by Gao et al. (2018) only handles events that are in either one of two states and assumes two competing 4D point sets are generated from two independent univariate point processes with the same spatial distributions (Warden 2008). Therefore, the Bernoulli model is mostly suited for case-control studies where two OD flows are drawn from the same dataset. It is ill-suited for our study as the migration flows are generated from two separate data sources.

In the existing literature, gravity models are popular frameworks for modeling various forms of flows, such as trade, cargo shipping, and migration between city pairs and regions (Anderson 2011). While a more recent study suggests that the radiation model outperforms a general gravity model in predicting migration volumes (Simini et al. 2012), gravity models allow explicit consideration of the spatial distributions of the movement flows as an explanatory variable (O’Kelly 2015, Rodrigue et al. 2016). A recent study suggests that the spatial pattern of Twitter user movements across the geographic space can be modeled by a gravity model (Yin et al. 2017). Many methods were proposed to fit gravity models in applications seeking to understand the spatial patterns of movement flows, such as the Log-normal and the Poisson model. The OD migration flow volumes between a pair of regions are commonly assumed to follow a Poisson distribution (Fischer and Wang 2011). In terms of empirical migration data, the literature suggests that the Poisson model is preferred over the Log-normal model (Flowerdew and Aitkin 1982, Wesolowski et al. 2013). Therefore, a multidimensional spatial scan statistic approach based on a Poisson process model is developed in this paper to detect the pairs of origin and destination regions where Twitter-estimated migration flows over- or under-represent the IRS counterparts.

Assuming the migration flows between a pair of regions follow a Poisson distribution, the probability to have \( k \) migrants from region \( A \) to region \( B \) under such a Poisson model is shown in Eq. 3, where \( n_{ab} \) is the migrant counts from region \( A \) to
region $B$ and $\lambda_{ab}$ is the expectation of $n_{ab}$. Since migration flows are estimated at the county level we used the centroids of U.S. counties as the coordinates of the flows. A migration flow from the origin county’s centroid to the destination county’s centroid, represented as a 4D point, can be modeled as $P_{Mi} = \langle x_{oi}, y_{oi}, x_{di}, y_{di} \rangle$, where $V_i$ is the migration flow counts from the origin county to the destination county. From an OD space perspective, $V_i$ represents the number of points within the Cartesian product of the spatial extents of the two counties. It is worth noting that a region in this context is an aggregated form of counties, so a region may contain one or multiple counties. Therefore, the migrant counts from five counties in Florida to six counties in California are also expected to follow a Poisson distribution. 

$$P(n_{ab} = k) = \frac{(\lambda_{ab})^k e^{-\lambda_{ab}}}{k!}$$  \hspace{1cm} (3)

Given the null hypothesis that the Twitter-estimated migration flows are fully representative of real migration flows, $\lambda_{ab}$ should be proportional to the IRS estimates from region $A$ to region $B$. Those regions where Twitter over-represents or under-represents real migration flows can be detected by identifying local high or low point clusters in the 4D Poisson point process. Given the study area with $M$ real migrants in the IRS estimates and $N$ migrants in the Twitter estimates, and a 4D scanning window $W$ has $M_W$ real migrants and $N_W$ migrated Twitter users, the likelihood ratio $\lambda$ for a Poisson model is shown in Eq. 4, where $E[N_W] = N_c M_W / M$ is the expected number of points in window $W$ following the null hypothesis (Kulldorff 1997). It works for both high and low clusters.

$$\lambda = \left( \frac{N_W}{E[N_W]} \right)^{N_W} \left( \frac{N - N_W}{N - E[N_W]} \right)^{N - N_W}$$ \hspace{1cm} (4)

4. Results

4.1. Estimated pairwise county-to-county migration flows

Taking the pairwise county-to-county migration flow matrices generated from Twitter and IRS data resulted in 75,540 and 54,159 pairwise county-to-county migration flows for the 2013–2014 and 2014–2015 datasets. The number of flows is denoted as 0 if the migration flows do not exist in one dataset. Tables 2 and 3 show the top 10 pairwise county-to-county migration flows estimated from IRS (field ‘IRS flow’) and Twitter in 2013–2014 (field ‘Twitter 13–14’) and 2014–2015 (field ‘Twitter 14–15’). The top 10 Twitter-estimated migration flows for 2013–2014 and 2014–2015 are quite consistent: 8 OD county pairs are shared, yet the orders of these pairs are slightly different. The consistency over time between Twitter and IRS estimates is relatively low, with only three OD county pairs shared by the two tables. The first noticeable inconsistency is from New York County, NY to Los Angeles County, CA (highlighted in color blue). The Twitter-estimated migration flows in both directions are among the top 10 in both 2013–2014 and 2014–2015. However, their ranks in the IRS estimates are significantly lower. Second, three OD county pairs, Harris County, TX to Fort Bend County, TX, Kings County, NY to Queens County, NY, and Cook County, IL to DuPage County, IL
Table 2. Top 10 migration flow estimates from the IRS data between 2013 and 2014.

| IRS rank (Top 10) | Twitter rank | From County | To County | IRS flow | Twitter 13–14 flow | Twitter rank (Top 10) | IRS rank | From County | To County | IRS Flow | Twitter 13–14 flow |
|-------------------|--------------|-------------|-----------|----------|---------------------|-----------------------|-----------|-------------|-----------|----------|---------------------|
| 1                 | 4            | Los Angeles, CA | Orange, CA | 19238    | 241                 | 1                     | 5          | New York, NY | Kings, NY | 13159    | 500                 |
| 2                 | 3            | Orange, CA     | Los Angeles, CA | 16554    | 271                 | 2                     | 24         | Kings, NY   | New York, NY | 8488     | 301                 |
| 3                 | 14           | Miami-Dade, FL | Broward, FL | 15819    | 134                 | 3                     | 2          | Orange, CA  | Los Angeles, CA | 16554   | 271                 |
| 4                 | 10           | Los Angeles, CA | San Bernardino, CA | 15750    | 154                 | 4                     | 1          | Los Angeles, CA | Orange, CA | 19238    | 241                 |
| 5                 | 1            | New York, NY   | Kings, NY   | 13159    | 500                 | 5                     | 235        | New York, NY | Los Angeles, CA | 2392    | 192                 |
| 6                 | 45           | Harris, TX     | Fort Bend, TX | 12948    | 80                  | 6                     | 40         | New York, NY | Queens, NY   | 7195     | 191                 |
| 7                 | 92           | Kings, NY      | Queens, NY  | 11611    | 56                  | 7                     | 68         | Queens, NY  | New York, NY | 5373     | 173                 |
| 8                 | 8            | San Bernardino, CA | Los Angeles, CA | 11245    | 170                 | 8                     | 8          | San Bernardino, CA | Los Angeles, CA | 11245   | 170                 |
| 9                 | 53           | Cook, IL       | DuPage, IL  | 11178    | 74                  | 9                     | 361        | Los Angeles, CA | New York, NY | 1790     | 162                 |
| 10                | 12           | Broward, FL    | Miami-Dade, FL | 10744    | 142                 | 10                    | 4          | Los Angeles, CA | San Bernardino, CA | 15750   | 154                 |
| IRS rank (Top 10) | Twitter rank | From County       | To County       | IRS flow | Twitter rank 14–15 flow | Twitter rank (Top 10) | IRS rank | From County       | To Country  |
|------------------|--------------|-------------------|-----------------|----------|-------------------------|-----------------------|----------|-------------------|------------|
| 1                | 4            | Los Angeles, CA   | Orange, CA      | 15621    | 220                     | 1                     | 20       | Kings, NY         | New York, NY |
| 2                | 3            | Orange, CA        | Los Angeles, CA | 13709    | 233                     | 2                     | 5        | New York, NY      | Kings, NY   |
| 3                | 15           | Los Angeles, CA   | San Bernardino, CA | 13581   | 97                      | 3                     | 2        | Orange, CA        | Los Angeles, CA |
| 4                | 16           | Miami-Dade, FL    | Broward, FL     | 12655    | 93                      | 4                     | 1        | Los Angeles, CA   | Orange, CA   |
| 5                | 2            | New York, NY      | Kings, NY       | 10955    | 322                     | 5                     | 56       | Queens, NY        | New York, NY |
| 6                | 61           | Kings, NY         | Queens, NY      | 10576    | 47                      | 6                     | 233      | New York, NY      | Los Angeles, CA |
| 7                | 45           | Harris, TX        | Fort Bend, TX   | 10309    | 54                      | 7                     | 41       | New York, NY      | Queens, NY   |
| 8                | 13           | San Bernardino, CA | Los Angeles, CA | 9425     | 99                      | 8                     | 331      | Los Angeles, CA   | New York, NY |
| 9                | 9            | Broward, FL       | Miami-Dade, FL  | 9024     | 123                     | 9                     | 9        | Broward, FL       | Miami-Dade, FL |
| 10               | 23           | Dallas, TX        | Tarrant, TX     | 8800     | 75                      | 10                    | 30       | DeKalb, GA        | Fulton, GA   |
are ranked in the top 10 in the IRS estimates but are ranked significantly lower in the Twitter flows.

4.2. Outbound and inbound migration flows by county

The inbound and outbound migration flows aggregated by county are statistically correlated between the Twitter and IRS estimates for both the 2013–2014 and 2014–2015 estimates. By performing a linear regression of the outbound and inbound migration flow counts, the coefficient of determination ($R^2$) value is 0.81 (inbound) and 0.91 (outbound) for the 2013–2014 estimates, and $R^2$ value is 0.79 (inbound) and 0.88 (outbound) for the 2014–2015 estimates. The coefficient values suggest that the migration flow estimates using Twitter data can provide a reasonably good prediction of the IRS estimates at the county level.

It should be noted that while neither Twitter nor IRS migration estimates cover the whole population, the Twitter-estimated migration flows may raise more concerns due to the skewed demographic composition of the Twitter user population (Luo et al. 2016, Yin et al. 2018). To potentially mitigate this challenge, counties, where the Twitter data tended to over-represent or under-represent the IRS estimates, were identified by using the representation index ($R_j$) ($0 < R_j \leq 1$ indicates an over-representation in county $j$ and $-1 \leq R_j < 0$ indicates an under-representation in county $j$). Figures 2 and 3 depict the counties where Twitter under-represents (low value, blue counties) and over-represents (high value, red counties), as opposed to the IRS outbound and inbound migration flows between the 2013–2014 and 2014–2015 estimates.

Both over- and under-representation appear to be spatially clustered. The spatial clusters are illustrated by running the Getis-Ord $G^*$ statistic in GeoDa (Supplement Materials, S4). In terms of outbound migration flows, there are a total of 1,639 out of 2,976 counties with over-representation and 1,337 counties with under-representations. Except for a few counties with large populations, such as Los Angeles County, CA, and New York County, NY, the over-representation is likely concentrating on less populated counties, such as ones with large rural populations. Note that the availability of locations of geo-located tweets is highly related to human populations/activities. There are fewer tweets in less visited/populated regions and may be no tweets at all in certain areas, such as mountains/forests, lakes, and deserts (Supplement Materials, S1). Counties with over-representation as shown in Figure 2 are very similar to those with significant population losses as identified from the 2000 to 2010 Census. They are located in the Great Plains, Corn Belt, Mississippi Delta, parts of the northern Appalachians, and the industrial and mining belts of New York and Pennsylvania (Johnson 2012). The figure used by Johnson (2012) showing the nonmetropolitan population change in the U.S. from 2000 to 2010 is shown in Supplement Materials, S3. Because these areas are mostly less populated, it is likely that the associated migration flows (from/to) are underestimated in the IRS data. IRS county-to-county migration data include only larger flows due to due to privacy concerns (DeWaard et al. 2018). However, it appears that these migration flows are captured in Twitter data, which is why it is shown up in red in the visualization. There are also some exceptions, especially in the Mountain regions, such as the states of Wyoming,
Colorado, and Arizona, possibly as some counties in those regions are popular tourist destinations. One possible explanation is that Twitter users may actively post geolocated tweets when they are visiting those places and mislead our proposed heuristics to treat those counties as one of their counties of residence. This explanation is evident in the observation of over-representations in counties with well-known tourist attractions, such as Coconino County, AZ (Grand Canyon), Teton County, WY (Yellowstone and Grand Teton), and Monroe County, FL (Florida Keys and Everglades).

In terms of inbound migration flows, there are a total of 1,345 out of 2,931 counties with over-representation and 1,586 counties with under-representations. The overall spatial distribution of the representativeness in the inbound migration flows is quite similar to the outbound migration flows. There are several cases where the outbound migration flows in a particular county are under-represented by Twitter estimates, but its inbound migration flows are over-represented, such as Cook County, IL, and Miami-Dade County, FL. However, the margin of the representativeness around value 0 is quite narrow.

The results of the representation index for both inbound and outbound migration flows from the 2014 to 2015 estimates are very similar to the ones from the 2013 to 2014 cases illustrated in Figures 2 and 3 (see Supplement Materials, S4). While the maps showing the spatial patterns of the representation index are informative, the analysis is exploratory and hence the findings are limited. Because the aggregated migration flows to each county are treated separately as inbound and outbound migration flows, it is impossible to know the origin and destination county pairs of the Twitter migration flows that are under- or over-represented in their IRS migration counterparts.

Figure 2. The representation index between the 2013 and 2014 Twitter and IRS outbound migrants by county.
4.3. Spatial patterns of the Twitter migration flow representativeness

To better visualize the spatial clusters representing the origin and destination regions where Twitter-estimated migration flows under-represent or over-represent their IRS counterparts, the top 5 high and low clusters from the 2013 to 2014 estimates in each scenario are shown in Figures 4 and 5. The rank of clusters is determined by the likelihood ratio $\lambda$ and the statistical significance (i.e. $p$-value) is estimated by performing a Monte Carlo simulation. The results are generated with an increment size of scan windows of 25 km and the maximum cluster radius is 1000 km as suggested by Gao et al. (2018). As using 25 km as the increment size of the scan window may miss some neighboring counties in the calculation, the same method with a 50 km increment size of scan window was tested (see Supplement Materials, S5). In both figures, the origin extent (solid line) and the destination extent (dashed line) for each cluster are displayed in the same color. The arrows represent the direction of migration flows between the centers of clusters. The summary statistics of these clusters are shown in Supplement Materials, S5. The clusters were determined without allowing intersecting origin and destination spatial extent. Note that the source code shared with this article has the option to allow detected clusters to be overlapped. Because our study is carried out at the national scale, clusters with overlapping origin and destination extent will confuse the direction of migration flows between nearby counties with a large population leading to cluttered visualization and providing little information about inter-regional migration patterns (see Supplement Materials, S5). The following analysis will focus on the common characteristics of the clusters rather than explaining the details regarding what each cluster represents.
Regarding the spatial pattern of Twitter-estimated migration flows under-representing the IRS estimates, the top five low clusters all exist in regions with dense populations. The most likely cluster (cluster 1) originates from the East and Midwest parts to the West Coast parts of the U.S. (mostly California). The IRS migration flows in this cluster (55,370) are approximately half of the expected value (102,219) if Twitter and IRS migration flows have the same spatial distribution in these regions. The second cluster (cluster 2) is the opposite, originating from the West Coast and going to the Northeast region (36,752), where IRS migration flows are also approximately half of the expected value (71,173). The Twitter-estimated migration flows under-represent the IRS estimates from the South to the Northeast region (cluster 3), the South region to the Northwest region (cluster 4), and the Northeast to the Southeast region (cluster 5). The results provide different insights as to when using the representation index Twitter-estimated migration flows are likely to under-represent the IRS estimates between populated regions. The identified clusters can also inform where the under-represented migration flows are from (i.e. origin regions) and to (i.e. destination regions). The observation stays consistent when the increment size of the scan window is increased up to 50 km (Supplement Materials, S5) and is subject to a slight change in the extent, center, and rank of the clusters. Indeed, the top 5 low clusters from the 2014 to 2015 estimates are similar to the ones from the 2013 to 2014 estimates (Supplement Materials, S5).

The top five high clusters correspond to the origin and destination regions of Twitter-estimated migration flows over-represented by the IRS estimates. Compared to the low clusters (i.e. under-representation) the sizes of the spatial extent of the high clusters (i.e. over-representation) are much smaller. It is both interesting and odd to see

Figure 4. Centers and spatial extents of top five low clusters (i.e. under-representation) with a maximum radius of 1000 km without intersecting OD flows.
three of the destination regions are centered in the Capital Region of Texas, which is still the case when the increment size of the scan window is set to 50 km (Supplement Materials, S5). The frequent appearances of the clusters between Florida and Texas from both 2013–2014 and 2014–2015 (Supplement Material, S5) seem to contradict the observation that Twitter migration flows tend to under-represent the IRS estimate in populated regions. However, it may be because the number of migrants who have filed tax returns in those regions is more than the real number of actual migrants. Observation of the Capital Region of Texas as the most frequent origin and destination region, where Twitter-estimated migration flows over-represent the IRS estimates, may suggest that the IRS data itself does not capture enough migration flows either from or to this region. In contrast, the origin and destination extent of clusters 2 and 3 are reversed between the region around Boulder County, CO, and the region around the Denver Metropolitan area. This makes sense given that the Twitter-estimated migration flows tend to over-represent the IRS estimates from one metropolitan region to its adjacent region with tourist attractions. Indeed cluster 2 shows up as cluster 3 observed in the 2014–2015 estimates. It is also evident from cluster 2 in the 2014–2015 estimates, which are adjacent regions around the Seattle Metropolitan area (Supplement Materials, S5).

5. Discussions and conclusions

In this study, we used geo-located Twitter data and evaluated spatial patterns for representativeness of measuring human migration. Using geo-located tweets continuously
collected from 2013 to 2015 in the U.S., we assigned each Twitter user with a county-of-residence, then extracted Twitter users who migrated per changes in county-of-residence every two years, and generated Twitter-estimated county-to-county migration flows. We performed a regression analysis of the Twitter-estimated county-to-county migration flows with the ones extracted from the IRS migration data. The initial results suggest that geo-located Twitter data could be a sound statistical proxy for measuring human migration. It is worth noting that under the assumption of normal distribution for migration flow values, the Pearson’s $r$ value was high from the regression analysis between the pairwise county-to-county migration flows generated from the Twitter and IRS datasets. However, they failed Spearman’s rank-order correlation test which states that the two sets of migration flows are not correlated. After aggregating the pairwise county-to-county migration flows to each county as inbound and outbound migration flows, Spearman’s rank-order correlation test does suggest a strong correlation between the two sets of migration flows. Specifically, the coefficient of determination ($R^2$) value is 0.81 (inbound) and 0.91 (outbound) for the 2013–2014 estimates, whereas the $R^2$ value is 0.79 (inbound) and 0.88 (outbound) for the 2014–2015 estimates. The coefficient values suggest that the Twitter-estimated migration flows do provide a fairly good prediction of the IRS estimates at the county level. To evaluate the spatial variation of the inbound and outbound migration flows estimates from the two data sources, we used a representation index to depict the counties where Twitter user migration flows tend to over-represent or under-represent the IRS estimates. The results showed that the over-representations of the inbound and outbound Twitter-estimated migration flows are likely concentrated in less populated counties, such as ones with large rural populations. Over-representations are commonly seen in counties with tourist attractions.

We applied a Poisson process model-based multidimensional spatial scan statistic approach to quantitatively assess the spatial patterns of Twitter-estimated county-to-county migration flows representing the IRS estimates. This approach evaluates the spatial patterns of Twitter-estimated migration flows by identifying statically significant high (over-represent) and low (under-represent) spatial clusters. The results not only showed the center and spatial extent of those identified regions but also informed where the migration flows are generated from and to. The overall trend suggested that Twitter-estimated migration flows are likely to under-represent the IRS estimates in regions with dense populations. However, the frequent appearance of Texas as the origin and destination region with over-representation suggests that there is a migrant population that is captured by Twitter but may not be reflected in the IRS data. Over-representation is also likely to be observed from a metropolitan region to its adjacent region with tourist attractions. The developed spatial scan statistic approach can be useful for related research when using geo-located Twitter data as a proxy for migration studies, as data calibration is necessary within the detected clusters.

### 5.1. Limitations

There are still some limitations in the current study that should be carefully attended to in future research. The heuristics in determining the county-of-residence can be
improved. For example, a user lives in county A but sends more tweets at work in county B. The current approach is not able to recognize county A as the user’s county-of-residence, which will add uncertainty to the migration flows estimated between neighboring/adjacent counties. Although the 30-day time constraint for determining residency and the usage of a minimum of 30 tweets in determining county-of-residence does provide a more stable Twitter user population, there is also a risk of excluding many migrants who simply do not frequently send Twitter messages. Users who have migrated to more than one county are also not considered. The time interval in defining a migration event can be problematic. Although there is no official requirement for a specific period, the county-of-residence of a user is determined by a year’s observation. There are cases a migrant will not be recognized, such as the user who moved later in the new year but did not record many tweets in the new place of residence. Last but not least, Twitter has changed its geocoding policy so precise location tagging is turned off by default, which will significantly reduce the number of geo-located tweets and covers even fewer groups of Twitter users (Hu and Wang 2020). Considering this change, we envision that methods for inferring Twitter users’ home locations based on tweets content, such as developing a machine learning-based classifier (Mahmud et al. 2014) and performing sentiment analysis (Mostafa et al., 2022), will be utilized more by the research community and will be improved over time. Even if Twitter data became no longer accessible, our approach can still be used to evaluate many other types of movement flows. Our proposed multidimensional spatial scan statistic approach remains effective in detecting the pairs of origin and destination regions where the movement flows generated from those data sources over-/under-represent the ground truth.

Acknowledgments
The authors thank the editor and the anonymous reviewers for their constructive comments on earlier versions of the article. Appreciation is extended to Dr. Jennifer Van Hook at The Pennsylvania State University for her many helpful suggestions.

Author contributions
All authors conceived and designed the study and revised the manuscript. Junjun Yin and Yizhao Gao processed the data and coded methods. Junjun Yin performed experiments and wrote the original draft. Guangqing Chi contributed to the refinement of the proposed concepts, method and manuscript writing, and the discussion of the findings.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Data and codes availability statement
The data and source code that support the findings of this study are available in [figshare] at [https://doi.org/10.6084/m9.figshare.14759976]. The geo-located Twitter data are not publicly available according to the Twitter data sharing policy as precise geo-locations could
compromise the privacy of Twitter users. However, a sample collection of 2000 raw geo-located
tweets are included in the shared file.

**Funding**

This research was supported in part by the National Science Foundation [Awards #1541136, 
#1823633, and #1927827]; the Eunice Kennedy Shriver National Institute of Child Health and 
Human Development [Award #P2C HD041025]; the USDA National Institute of Food and 
Agriculture and Multistate Research Project #PEN04623 [Accession #1013257]; and the Social 
Science Research Institute, Population Research Institute, and Institute for Computational and 
Data Sciences of the Pennsylvania State University.

**Notes on contributors**

**Junjun Yin** is an Assistant Research Professor in the Computational and Spatial Analysis Core at 
The Pennsylvania State University, University Park, PA 16802. E-mail: jyin@psu.edu. His research 
interests include computational geography approaches and geospatial big data to model 
human-urban environment interactions about urban environmental sustainability, resilience, 
and mobility.

**Yizhao Gao** received his Ph.D. at the Department of Geography and Geographic Information 
Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801. E-mail: ygao29@illinois. 
edu. His research interest includes CyberGIS, movement analysis, and spatiotemporal analysis.

**Guangqing Chi** is a Professor of Rural Sociology, Demography, and Public Health Sciences in the 
Department of Agricultural Economics, Sociology, and Education and Director of the 
Computational and Spatial Analysis Core at The Pennsylvania State University, University Park, 
PA 16802. E-mail: gchi@psu.edu. His research interests focus on socio-environmental systems, 
aiming to understand the interactions between human populations and built and natural envi-
ronments and to identify important assets to help vulnerable populations adapt and become 
resilient to environmental changes by developing and implementing spatial and big data ana-
lytic methods.

**ORCID**

Junjun Yin [http://orcid.org/0000-0002-4196-2439](http://orcid.org/0000-0002-4196-2439)

Guangqing Chi [http://orcid.org/0000-0003-0888-7964](http://orcid.org/0000-0003-0888-7964)

**Reference**

Anderson, J.E., 2011. The Gravity Model. *Annual Review of Economics*, 3 (1), 133–160.

Belyi, A., *et al.*, 2017. Global multi-layer network of human mobility. *International Journal of 
Geographical Information Science*, 31 (7), 1381–1402.

Blumenstock, J.E., 2012. Inferring patterns of internal migration from mobile phone call records: 
evidence from Rwanda. *Information Technology for Development*, 18 (2), 107–125.

Deville, P., *et al.*, 2014. Dynamic population mapping using mobile phone data. *Proceedings of 
the National Academy of Sciences of the United States of America*, 111 (45), 15888–15893.

DeWaard, J., Johnson, J., and Whitaker, S., 2018. Internal migration in the United States: a com-
parative assessment of the utility of the consumer credit panel. *SSRN Electronic Journal*, 2018, 
192.

Duggan, M., and Smith, A., 2014. Social media update 2013. Pew research center. [https://www. 
pewresearch.org/internet/2013/12/30/social-media-update-2013/](https://www.pewresearch.org/internet/2013/12/30/social-media-update-2013/) [Accessed 18 December 2019].
Fiorio, L., et al., 2017. Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In Proceedings of the 2017 ACM on Web Science Conference – WebSci ’17, 103–110. New York, New York, USA: ACM.
Fischer, M.M., and Wang, J., 2011. Spatial data analysis: models, methods and techniques. Berlin, Heidelberg: Springer Science & Business Media.
Flowerdew, R., and Aitkin, M., 1982. A method of fitting the gravity model based on the poisson distribution. Journal of Regional Science, 22 (2), 191–202.
Gao, Y., 2018. Data-intensive spatial pattern discovery based on generalized spatial point representations. https://www.ideals.illinois.edu/handle/2142/101698 [Accessed 23 June 2021].
Gao, Y., Li, T., et al., 2018. A multidimensional spatial scan statistics approach to movement pattern comparison. International Journal of Geographical Information Science, 32 (7), 1304–1325.
Gayo-Avello, D., Metaxas, P.T., and Mustafaraj, E., 2011. Limits of electoral predictions using twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Limits, 193–200. Barcelona.
González, M.C., Hidalgo, C.A., and Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature, 453 (7196), 779–782.
Hawelka, B., et al., 2014. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41 (3), 260–271.
Hu, Y., and Wang, R.Q., 2020. Understanding the removal of precise geotagging in tweets. Nature Human Behaviour, 4 (12), 1219–1221.
Huang, X., et al., 2020. Twitter reveals human mobility dynamics during the COVID-19 pandemic. PLoS One, 15 (11), e0241957.
Hübl, F., et al., 2017. Analyzing refugee migration patterns using geo-tagged Tweets. ISPRS International Journal of Geo-Information, 6 (10), 302.
Jensen, E.B., 2013. A review of methods for estimating emigration. US Census Bureau, Population Division.
Jiang, Y., Li, Z., and Ye, X., 2018. Understanding demographic and socioeconomic biases of geo-tagged Twitter users at the county level. Cartography and Geographic Information Science 46 (3):228–242.
Johnson, K., 2012. Rural demographic change in the new century slower growth, increased diversity. The Carsey School of Public Policy at the Scholars’ Repository, 2012. 159.
Jones, N.M., et al., 2016. Tweeting negative emotion: an investigation of Twitter data in the aftermath of violence on college campuses. Psychological Methods, 21 (4), 526–541.
Jurdak, R., et al., 2015. Understanding human mobility from Twitter. PLoS One, 10 (7), e0131469.
Kaplan, G., and Schulhofer-Wohl, S., 2012. Interstate migration has fallen less than you think: consequences of hot deck imputation in the current population survey. Demography, 49 (3), 1061–1074.
Kim, A.E., et al., 2013. Methodological considerations in analyzing Twitter data. Journal of the National Cancer Institute, 2013 (47), 140–146.
Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics, 26 (6), 1481–1496.
Leetaru, K., et al., 2013. Mapping the global Twitter heartbeat: the geography of Twitter. First Monday, 18 (5), 5–6.
Lin, A.Y., Cranshaw, J., and Counts, S., 2019. Forecasting U.S. domestic migration using internet search queries. In The Web Conference 2019 – Proceedings of the World Wide Web Conference, WWW 2019, 1061–1072. Association for Computing Machinery, Inc.
Lin, J., and Cromley, R.G., 2018. Inferring the home locations of Twitter users based on the spatiotemporal clustering of Twitter data. Transactions in GIS, 22 (1), 82–97.
Longley, P.A., and Adnan, M., 2016. Geo-temporal Twitter demographics. International Journal of Geographical Information Science, 30 (2), 369–389.
Longley, P.A., Adnan, M., and Lansley, G., 2015. The geotemporal demographics of Twitter usage. Environment and Planning A, 47 (2), 465–484.
Luo, F., et al., 2016. Explore spatiotemporal and demographic characteristics of human mobility via Twitter: a case study of Chicago. Applied Geography, 70, 11–25.
Mahmud, J., Nichols, J., and Drews, C., 2014. Home location identification of Twitter users. *ACM Transactions on Intelligent Systems and Technology*, 5 (3), 1–21.

McNeill, G., Bright, J., and Hale, S.A., 2017. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science*, 6 (1), 24.

Mislove, A., et al., 2011. Understanding the demographics of Twitter users. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 554–557. Barcelona.

Mostafa, A., et al., 2022. Pre-HLSA: predicting home location for Twitter users based on sentiment analysis. *Ain Shams Engineering Journal*, 13 (1), 101501.

Nguyen, Q.C., et al., 2016. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, 73, 77–88.

O’Kelly, M.E., 2015. Spatial Interaction. In *International Encyclopedia of the Social & Behavioral Sciences*. Amsterdam: Elsevier, 152–156.

Rodrige, J.-P., Comtois, C., and Slack, B., 2016. *The geography of transport systems*. London: Routledge.

Simini, F., et al., 2012. A universal model for mobility and migration patterns. *Nature*, 484 (7392), 96–100.

Smith, M.A., and Leigh, B., 1997. Virtual subjects: using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, 29 (4), 496–505.

Soliman, A., et al., 2015. Where Chicagoans tweet the most. In *Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, 55–58. New York, NY, USA: ACM.

Warden, C.R., 2008. Comparison of poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. *International Journal of Health Geographics*, 7 (1), 51.

Weber, I., et al., 2014. Using geolocated Twitter data to study recent patterns of international and internal migration in OECD Countries. In *European Population Conference 2014*. Budapest: Hungary, 1–9.

Wesolowski, A., et al., 2013. The use of census migration data to approximate human movement patterns across temporal scales. *PLoS One*, 8 (1), e52971.

Willekens, F., et al., 2016. International MIGRATION. International migration under the microscope. *Science*, 352 (6288), 897–899.

Wilson, R., et al., 2016. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal Earthquake. *PLoS Currents*, 8, 1–25.

Xin, Y., and MacEachren, A.M., 2020. Characterizing traveling fans: a workflow for event-oriented travel pattern analysis using Twitter data. *International Journal of Geographical Information Science*, 34 (12), 2497–2516.

Yildiz, D., et al., 2017. Using Twitter data for demographic research. *Demographic Research*, 37, 1477–1514.

Yin, J., et al., 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*, 31 (7), 1293–1313.

Yin, J., and Chi, G., 2021. Characterizing people’s daily activity patterns in the urban environment: a mobility network approach with geographic context-aware Twitter data. *Annals of the American Association of Geographers*, 111 (7), 1–21.

Yin, J., Chi, G., and Van Hook, J., 2018. Evaluating the representativeness in the geographic distribution of twitter user population. In: R. S. Purves and C. B. Jones, eds. *Proceedings of the 12th Workshop on Geographic Information Retrieval*. New York, NY: ACM, 1–2.

Zagheni, E., and Weber, I., 2012. You are where you e-mail. In *Proceedings of the 3rd Annual ACM Web Science Conference on – WebSci ’12*, 348–351. New York, NY: ACM Press.

Zagheni, E., et al., 2014. Inferring international and internal migration patterns from Twitter data. In: *Proceedings of the 23rd International Conference on World Wide Web – WWW ’14 Companion*. New York, NY, 439–444.