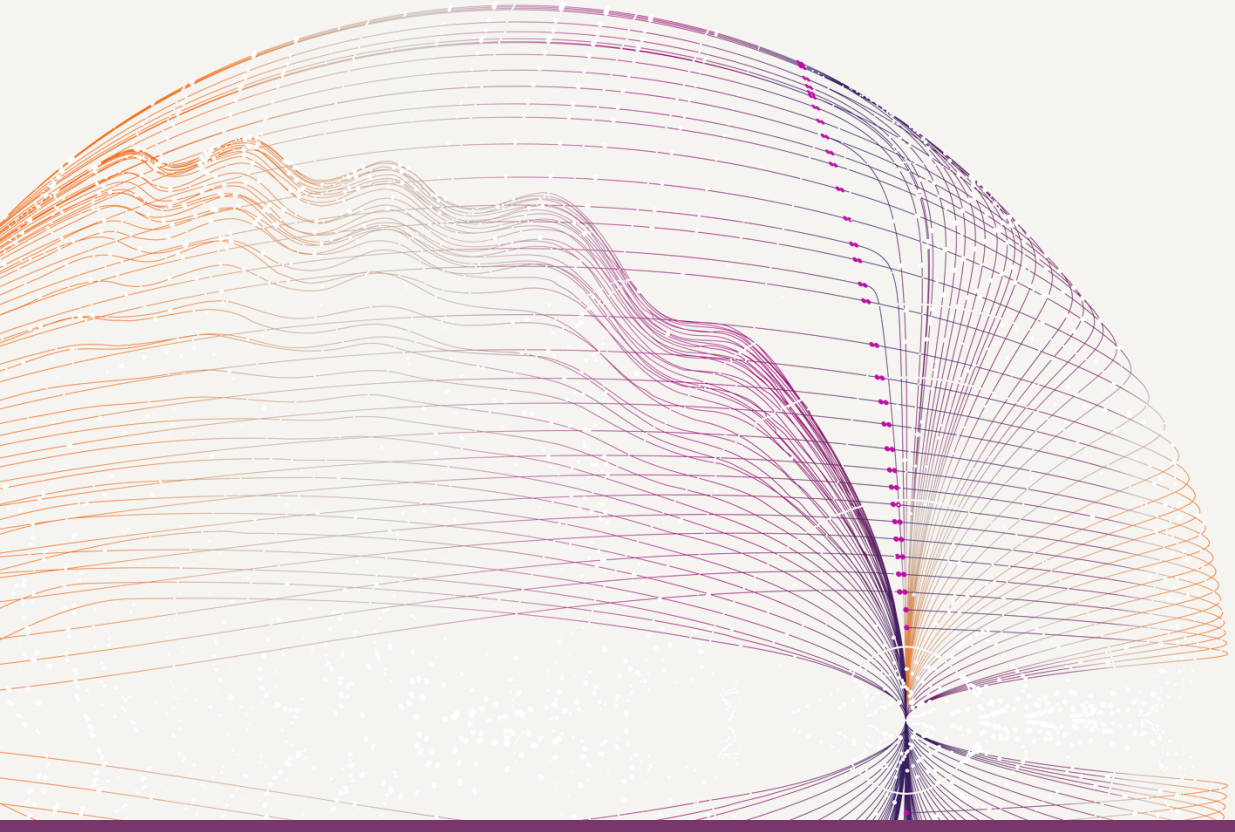
NORC Global

FINAL REPORT
JULY 2024

DEMOCRATIC BACKSLIDING AND MIGRATION INTENTIONS IN LATIN AMERICA AND THE CARIBBEAN

USING TWITTER DATA TO STUDY MIGRATION



NORC Global is a Division of NORC at the University of Chicago

Table of Contents

List of Figures

List of Tables

Acronyms

ACS	American Communities Survey
API	Application Programming Interface
CBP	Customs and Border Protection
LAC	Latin America and The Caribbean
NORC	NORC at the University of Chicago
RQ	Research Question

Executive Summary

Research Overview

This study examines potential uses of Twitter (now X) data for examining migration in the Latin America and Caribbean (LAC) region. The goal of the overall project is to examine whether democratic erosion contributes to increased migration to the U.S. and other countries in the LAC region, after accounting for other push and pull factors that also influence migration trends. The overall project investigated this question using multiple data sources, including data on encounters at the U.S. Southern border from U.S. Customs and Border Protection (CBP), opinion poll data, and internet search data from Google Trends. The results are presented in three companion reports, available here.

This report summarizes the findings related to whether and how social media data from Twitter can be used to complement other data sources to study factors that affect migration trends in LAC. Because of the well-known limitations of existing data sources on population movements, migration researchers have begun to use social media data as an alternative in recent years.[1] We focus on Twitter as our social media platform of interest due to the open availability of its data. Unlike Facebook, Instagram, and other platforms, Twitter makes data from users' posts available to researchers and scholars.[2] **The goal of the analysis is to examine whether this data can be used for studying how domestic factors in countries of origin affect migration trends when administrative data is unavailable or of low quality.**

Methodology

The analysis in this report draws on a dataset of geolocated tweets from January 2015 to December 2022 for target countries in LAC. The research team developed a set of algorithms for identifying residents of countries of interest and tracking cross-border movement over time. Tweets contain varying geospatial specificity: all tweets have information on the country from which they are posted and most also contain more precise locations (regions, neighborhoods, point of interest and GPS coordinates). Analysis is conducted at different levels, including country, administrative department (for origin countries in LAC) and state levels (for destinations in the United States). The team developed methods for distinguishing between visitors and emigrants, offering insights into migration patterns observed through Twitter data.

[1] Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet, and J. J. Ramasco. "Migrant mobility flows characterized with digital data." *PLOS ONE* 15, no. 3 (2020): e0230264.
[2] Twitter allows researchers to obtain access to all posts for use in research and analysis through its Application Programming Interface (API). Previously, access was free to the public. Since 2023, researchers must pay for access. At present, a Basic account (which allows access to 10,000 posts per month) is $100/month; a Pro account (which allows access to 1 million posts per month) is $5,000/month. Higher levels of access are available at higher rates. See https://developer.x.com/en/docs/twitter-api.

Main Findings

We report four main findings:

- Twitter data is not well suited for estimating monthly or annual migration flows to the U.S. from countries in LAC. Because there are relatively small numbers of Twitter users in many LAC countries, "upscaling" the Twitter-based migration estimates to generate population estimates can be unreliable.[3] In addition, researchers cannot distinguish between irregular and regular migrants using Twitter data. Therefore, it is not recommended as a substitute for administrative data collected by CBP or regional governments.

- Twitter data can be useful for tracking population movements during major crises when large numbers of people depart specific countries, as illustrated by prior research on the Venezuelan exodus in 2018.[4]

- Twitter data can be used to examine migrant transport routes to the U.S. and settlement locations after crossing the border. The report provides an illustration by tracking migration routes through Mexico and U.S. state-level settlement patterns for migrants from five LAC countries. This type of analysis could provide insights into precise locations in which migrants need most support.

- Twitter data can be used to examine where migrants come from in their home countries. While data on sub-national origins for migrants encountered at the U.S. Southwest border is also collected by the U.S. CBP, that data is not publicly available and has only been collected in recent years. Twitter-based estimates could serve as a valuable complementary data source. While the data is not well suited for estimating the number of migrants from different sub-national areas, it could be used to track changes in the proportion of migrants coming from specific areas, relative to the total number of migrants from a specific country, over time. These data could be used to study how the root causes of migration – including violence, economic shocks, and natural disasters – affect migration dynamics, leading to higher rates of displacement in some areas within affected countries. To do so, researchers could examine how the proportion of migrants

[3] The challenges relate to the small size of the population of Twitter users is more acute for smaller countries, including much of Central America, than for larger countries like Mexico. Nonetheless, our investigation found that even for larger countries included in this analysis (including Mexico) the Twitter-based estimates were unreliable.
[4] Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet, and J. J. Ramasco. "Migrant mobility flows characterized with digital data." *PLOS ONE* 15, no. 3 (2020): e0230264. K. C. Roy, M. Cebrian, and S. Hasan, "Quantifying human mobility resilience to extreme events using geo-located social media data," EPJ Data Science, vol. 8, no. 1, pp. 1–15, 2019. Hawelka, Bartosz, et al. "Geo-located Twitter as proxy for global mobility patterns." *Cartography and geographic information science* 41.3 (2014): 260-271. Huang, Xiao, et al. "Twitter reveals human mobility dynamics during the COVID-19 pandemic." *PloS one* 15.11 (2020): e0241957. Wang, Yan, and John E. Taylor. "Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake." *Natural hazards* 92 (2018): 907-925.

from particular sub-national areas changes over time in response to changes in the push factors that drive out-migration across those areas, even if the Twitter data cannot be used to estimate the number of migrants. At the same time, considerable technical expertise is required to access and process Twitter data, limiting its potential value for most potential uses.

Goals and Research Questions

NORC at the University of Chicago (NORC) seeks to understand whether and how democratic backsliding affects migration in the LAC region, with a particular interest in two priority countries, Nicaragua and El Salvador, that have experienced sustained crises of democratic governance in parallel with rising out-migration to the U.S. and other countries in the region. NORC conducted research to shed light on whether democratic erosion contributes to increased migration to the U.S. and other countries in the LAC region, after accounting for other push and pull factors that also influence migration trends. Additionally, the research team examined potential uses of alternative data sources from social media for studying country-to-country migration.

To expand the evidence base for investing in interventions designed to build and maintain democracy in the region, the research team conducted complementary analysis using multiple data sources and analytic techniques. The first report examines *migration intentions* using opinion poll data throughout the LAC region and in two priority countries: Nicaragua and El Salvador. The second report examines migration trends to the U.S. with data from U.S. CBP and Google Trends. This report summarizes findings related to using Twitter data for studying migration trends.

The overall activity is guided by three specific Research Questions (RQs) listed below. This report focuses on RQ3.

RQ1: Does democratic backsliding in the LAC region increase migration to the US and/or other countries in the region?

RQ2: What are the challenges related to using CBP data from the United States and other receiving countries for studying the connection between democratic backsliding in LAC countries and out-migration? What best practices should be adopted for using this data? What alternative sources should be used as complements to the CBP data?

RQ3: Can alternative methods and data sources be used to estimate country-to-country migration trends and intentions to migrate for countries in the LAC region? What methods should be adopted for using these alternative data sources?

1. Prior Literature

The challenges associated with obtaining high-quality data on cross-border population movement, particularly in the Global South, have led scholars to explore alternative data sources that leverage recent increases in mobile phone penetration and social media usage. The combination of these factors provides scholars with new data sources regarding users and their behavior, which traditionally could only be obtained through more costly methods. Data from social media platforms has been used to study global patterns of human mobility, estimate cross-border movements, and examine immigrant integration.[5]

Twitter is especially useful because the platform has historically provided open access to tweets and has employed transparent data policies – unlike most other social media platforms.[6] However, Twitter has several challenges that limit its relevance for particular applications. First, the number of users in countries in the Global South is often small, making it difficult to generate precise estimates of population movement. For example, as of January 2021 there were about 160,000 users in Nicaragua and 500,000 in El Salvador.[7] In addition, national samples on social media platforms are known to be more educated, urban, and affluent than the overall population – creating challenges for using this data to generate population estimates.[8] Moreover, it is difficult to distinguish migrants from other types of travelers, and it is not possible to differentiate irregular migrants (who enter foreign countries without necessary authorizations or stay beyond permitted periods) from regular migrants who enter through official channels.[9] Finally, in 2023 Twitter moved to a subscription-based model, which imposes constraints on researchers who have traditionally relied on free access through Twitter's application programming interface (API).

Particularly relevant to this study is prior work by Mazzoli et al. (2020), which used Twitter data to study the migration patterns of Venezuelan emigrants in 2018.[10] The study developed a methodology for identifying Venezuelan residents and tracking their movement across national borders. Critically, the authors were able to validate the methodology against official data from multiple sources, including the U.N, High Commissioner for Refugees, the International

[5] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, and J. J. Ramasco, “Cross-checking different sources of mobility information,” PloS one, vol. 9, no. 8, p. e105184, 2014. M. Lenormand, A. Tugores, P. Colet, and J. J. Ramasco, “Tweets on the road,” PloS one, vol. 9, no. 8, p. e105407, 2014. K. C. Roy, M. Cebrian, and S. Hasan, “Quantifying human mobility resilience to extreme events using geo-located social media data,” EPJ Data Science, vol. 8, no. 1, pp. 1–15, 2019.
Hawelka, Bartosz, et al. "Geo-located Twitter as proxy for global mobility patterns." *Cartography and geographic information science* 41.3 (2014): 260-271. Huang, Xiao, et al. "Twitter reveals human mobility dynamics during the COVID-19 pandemic." *PloS one* 15.11 (2020): e0241957. Wang, Yan, and John E. Taylor. "Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake." *Natural hazards* 92 (2018): 907-925.
[6] Starting in February 2023, Twitter began restricting free public access to data from its platform and now makes information available only for specific research projects and through subscription-based plans.
[7] https://www.statista.com/statistics/977791/number-twitter-users-in-latin-american-countries/
[8] Mosley, Mohsen, and David G. Rand. “Who is on Twitter (X)? Identifying demographic of Twitter Users.” Working paper, February 2024. Rosenzweig, Leah, Parrish Bergquist, Katherine Hoffmann Pham, Francesco Rampazzo, and Matto Mildenberger. "Survey sampling in the Global South using Facebook advertisements." (2020).
[9] Armstrong, Caitrin, et al. "Challenges when identifying migration from geo-located Twitter data." *EPJ Data Science* 10.1 (2021): 1.
[10] Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet, and J. J. Ramasco. “Migrant mobility flows characterized with digital data.” *PLOS ONE* 15, no. 3 (2020): e0230264.

Organization for Migration, and the Brazilian government, demonstrating the potential of using Twitter to create accurate estimates of migration flows in the LAC region. The present study, under the direction of one of the lead authors of this prior research, draws on methods developed in that case and explores their application to a wider set of countries and contexts in LAC.[11]

2. Data And Methods

Twitter is one of the most widely used social media platforms globally, boasting a total of 1.3 billion existing accounts, with an estimated 500 million active users. For this project, the research team used Twitter's API to obtain tweets from January 2015 to December 2022 for nine LAC countries: Belize, El Salvador, Guatemala, Honduras, Panama, Costa Rica, Venezuela, Mexico, and Nicaragua. The geolocation data included with tweets allows for their use in spatial analysis. While a relatively small number of tweets in the sample (6 percent of the total dataset) included precise latitude and longitude coordinates, most provided a less precise bounding box showing the user's approximate location at the time the tweet was sent.

Using the approach developed in Mazzoli et al. (2020)[12] as a starting point, the research team developed procedures for identifying residents in target LAC countries and tracking their movement across national borders. The method included the following steps:

1. **Obtain sample.** Start with a large database of geolocated tweets in target countries during the period under study.
2. **Remove bots.** A filter to remove bots was implemented, excluding users who on average tweet more than three times per hour throughout their tweeting lifespan.
3. **Remove users with anomalous activity.** The team excluded tweets generated by accounts tweeting from vastly different locations in a very short period. To do so, the team calculated users' scroll speed based on consecutive tweets and their respective coordinates. Users with scroll speeds exceeding the typical airplane speed of 750 km/h were excluded from the database (plane-speed rule). These mainly correspond to business accounts used by multiple people to tweet from distant locations in short time spans.
4. **Remove tweets with large bounding boxes (tweets with imprecise spatial location information).** For analysis focused on sub-regions within countries, we applied an area filter to retain only tweets whose bounding boxes have an area of less than 40 square kilometers. Larger areas do not provide precise information about the location where the user was when the tweet was sent.

[11] It is important to note that that number of Twitter users is substantially larger in Venezuela than in other countries of interest for this project. Data from Statista indicates that as of January 2021, there were 1.4 million users in Venezuela (about 5% of the population), compared to about 500,000 in El Salvador (about 8 percent of the population) and 160,000 in Nicaragua (2 percent of the population) at the same time.
[12] Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet, and J. J. Ramasco. "Migrant mobility flows characterized with digital data." *PLOS ONE* 15, no. 3 (2020): e0230264.

5. **Identify residents of target countries.** A key task was to identify users residing in each country of origin included in the study. To do so, we checked the area where users had the highest volume of activity. For each user, we identified the most common location for tweets posted every month. This approach allowed us to create the users' history as a sequence of administrative location flags, with one flag for each month. If, during a given month, no tweets were posted or there was a draw between two or more locations, we labeled the month as "undetermined." We then applied the following criterion to define country of residence: We defined a user as a resident of country X (e.g., El Salvador) if the user's location was in that country for three consecutive months. If more than one country met this criterion, the last country with three consecutive months was set as the user's residence country. Additionally, we used an alternative, less restrictive criterion to define residence countries. The alternative approach counted users with two months (out of three) in a specific country as a resident of that country.
6. **Identify migrants to the U.S.** Once we had a pool of users identified as residents for each target country, we quantified the number who subsequently traveled to the U.S. We distinguish between two types of visitors: short-term visitors and long-term migrants. We define a user as a migrant if the user's location (as defined above) was in the U.S. for three consecutive months.
7. **Upscale Twitter data to generate population estimates.** The final step is to generate annual estimates of the overall population of long-term migrants from each country of origin to the U.S. For this, we multiply the number of observed migrants in the data by the inverse of the Twitter population share in each country of origin. For example, if the sample of unique Twitter users for a particular country is 10 percent, we multiply the raw number of migrants per year by 10 to generate population estimates. We use constant upscaling factors for each country for years between 2015 and 2022. This was done because of the low volume of Twitter data available for some countries and years, which results in a high volatility in upscaling factor values, especially for the recent years, as Twitter activity has decreased significantly.

Table 1 provides details on the estimates generated using the above methods. Column 2 shows the total number of unique twitter users who tweeted once or more in each country between 2015 and 2022. Column 3 shows the number of unique twitter users identified as residents using the three-month rule described above. Column 4 shows the number of those residents who were excluded from the sample due to violating the plane-speed rule. Column 5 shows the final number of unique Twitter users identified as residents of each country during the period under study. Column 6 shows the total population in 2015, taken from the World Bank's World Development Indicators. Finally, column 7 shows the ratio of Twitter users to the overall population.

As noted above, a key limitation is that the sample of Twitter users identified as residents is relatively small for several LAC countries. Thus, for El Salvador and Nicaragua, the number of residents identified over the entire eight-year period of study (column 5) is just 33,008 and 22,357 respectively. These represent small shares of the overall populations for each country – just 0.5 percent and 0.3 percent respectively. As a result, the data for these countries require large upscaling factors, resulting in unstable population estimates.

Table 1: Data Summary

(1) Country	(2) Twitter users observed in each country	(3) Residents (initial estimate)	(4) Users removed due to plane speed filter	(5) Residents (final estimate)	(6) 2015 population	(7) Ratio: residents / population
El Salvador	11,4866	36,822	3,814	33,008	6,231,066	0.0053
Nicaragua	74,835	26,462	4,105	22,357	6,298,598	0.0035
Venezuela	739,828	321,567	34,434	287,133	30,529,716	0.0094
Costa Rica	194,326	55,216	9,304	45,912	4,895,242	0.0094
Guatemala	160,607	59,123	6,693	52,430	15,567,419	0.0034
Honduras	89,134	29,572	3,844	25,728	9,294,505	0.0028
Panama	174,843	57,544	10,427	47,117	3,957,099	0.0119
Mexico	2,993,756	1,181,035	144,267	1,036,768	120,149,897	0.0086

3. Findings

Finding 1: Twitter data is not well suited for estimating routine migration flows from LAC to the U.S.

While Twitter data has many documented uses for studying human mobility, it is not well-suited for generating estimate of routine migration from LAC countries to the U.S. for three reasons:

- First, the number of active Twitter users in most LAC countries is small, as noted above.[13] To develop reasonably precise estimates, researchers need a sufficient volume of activity in target countries. The methods described above generate estimates with a certain level of statistical resolution, which depends on the upscaling factors that adjust the Twitter estimates to reflect the size of the national population. Higher upscaling factors result in higher uncertainty.

- The methods do not adequately distinguish between short-term visitors and long-term migrants. To deal with this challenge, the research team implemented common-sense coding rules to remove individuals who were present for short periods in either the origin or destination country. However, they cannot identify other types of travelers who should be excluded from the estimates, including U.S. expats living abroad who return to the U.S and foreign nationals who come to the U.S. as students or for longer visits to

[13] In most LAC countries, particularly those in Central America that have been major sources of migrants to the US in recent years, Twitter users make up less than 10 percent of the population. For example, data from Statista show that in 2021, Twitter users made up about five percent of the population in Venezuela, eight in El Salvador, and two in Nicaragua. For details, see https://www.statista.com/statistics/977791/number-twitter-users-in-latin-american-countries/

relatives based in the U.S. Moreover, the data cannot distinguish between irregular and regular forms of migration.

- Official data required to validate the estimates is not available. For mobility between countries in Central America and the United States, the data is not reliable due to the high volume of irregular immigration present in these regions.

To illustrate these limitations, Table 2 shows the estimated number of migrants to the U.S. for 2019.[14] Column 5 shows the number of migrants per country estimated based on the Twitter data. For comparison, column 6 shows the number of migrant encounters recorded by the U.S. Customs and Border Protection (CBP) for the same period. We do not expect a perfect match since the Twitter data includes all residents (as identified in the coding rules described above) from target countries who subsequently came to the U.S. (and stayed for at least three months), not just those who entered at the Southern border. However, the lack of association between the Twitter estimates and the CBP data is a cause for concern. For example, the Twitter-based approach estimates that 263,049 people migrated to the U.S. from Costa Rica, while the actual number of CBP encounters was just 73. The mismatch is likely due to the fact that the Twitter estimates include travelers who do not count as migrants according to standard definitions and the imprecision due to the very large upscaling factors used for the Twitter-based estimates.

Table 2: Estimated Travelers to the U.S. in 2019

(1) Country	(2) Number of Twitter residents	(3) Number of Twitter migrants to U.S.	(4) Upscaling factor	(5) Estimated number of migrants to the U.S. using Twitter data	(6) CBP encounters at U.S. Southern Border
El Salvador	9,940	374	631.79	236,289	83,030
Nicaragua	5,509	164	1209.66	198,343	11,516
Venezuela	50,834	995	569.89	567,044	2,875
Costa Rica	13,725	710	370.49	263,049	73
Guatemala	12,201	511	1360.54	695,238	224,945
Honduras	7,065	63	1409.62	88,806	227,808
Panama	13,661	672	309.86	208,226	36
Mexico	314,224	24,157	398.12	9,617,472	180,020

[14] We chose this year for comparison purposes because it precedes the onset of Covid-19 and the implementation of the Title-42 restrictions that led to increased rates of repeat crossing at the U.S. Southern border.

Finding 2: Twitter is useful for tracking migrant routes from LAC countries to the U.S. and settlement in the U.S.

Despite the limitations noted above, Twitter data can be useful for tracking migrant routes and settlement locations in recipient countries. By using geotagged tweets, researchers and policymakers can pinpoint specific locations along migrant routes, identify transit hubs, and track the movement of individuals and groups. Furthermore, monitoring tweets allows for the examination of border crossings and studying settlement after entry. Twitter can be useful here because for these applications the data does not need to be upscaled to generate population estimates; rather, the analysis is based on the raw number of Twitter users in each case.

As an illustration, Figure 1 plots the location of tweets by residents from the sample of LAC countries included in the analysis for this report[15] inside Mexico and compares the results to information on common migrant routes obtained from prior qualitative research.[16] The estimates in Figure 1 are based on the volume of tweets within grid cells measuring 0.5 degrees of longitude and latitude during the entire time period for the study, 2015-2022. The data is aggregated from all LAC origin countries included in the study, except Mexico.

Figure 1: **Heatmap of Tweets by Residents from Target LAC Countries in Mexico (Left) and Known Migrant Routes (Right)**

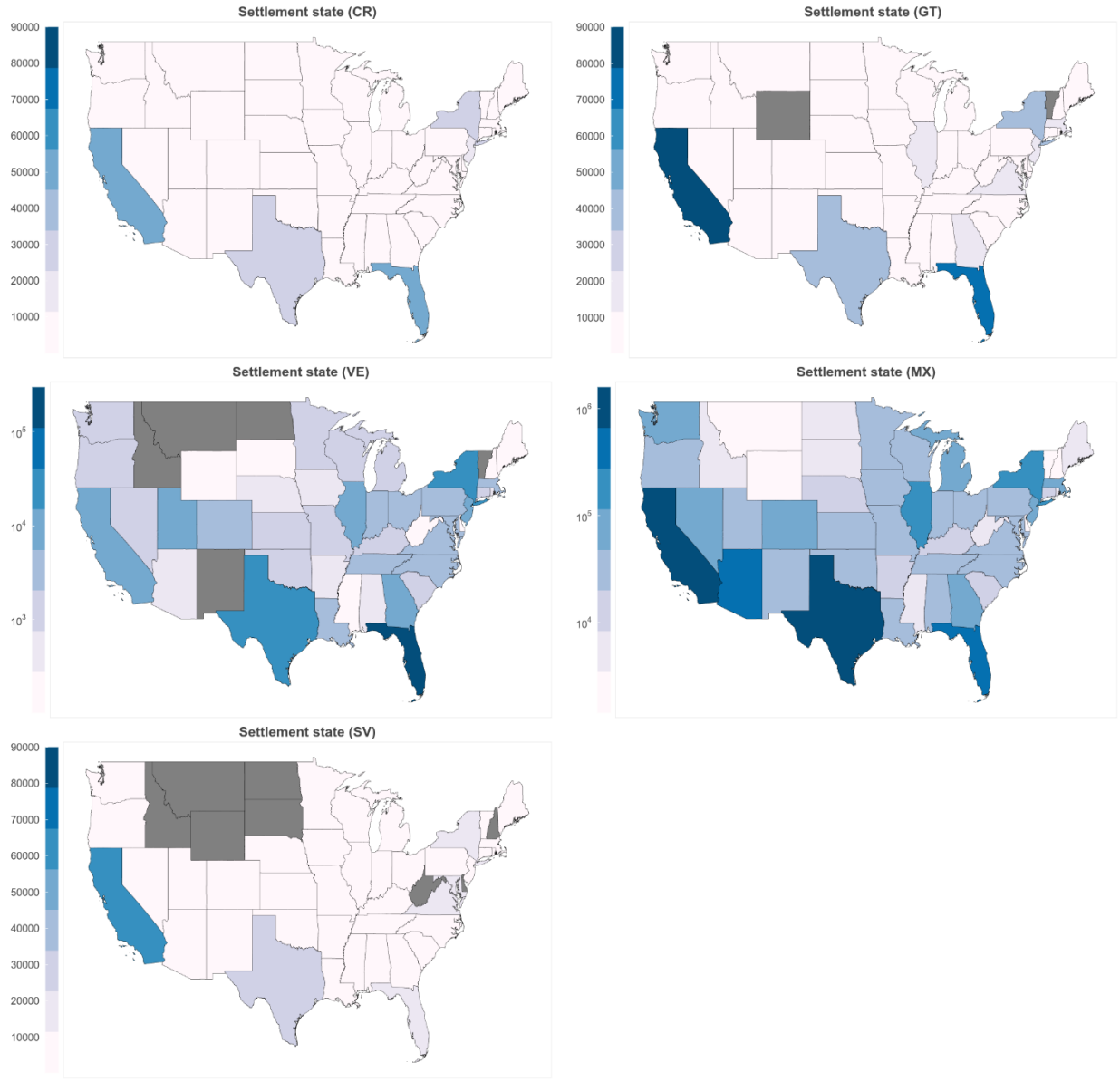
Notes: The heatmap scale (left figure) is presented logarithmically with a smoothing filter applied. Additionally, a minimum threshold of 1,000 detected persons per cell is applied for enhanced visualization and interpretation.

The Twitter results closely align with known information about migration routes through Mexico provided by other studies. The map on the right side shows major roads/paths and train routes used by irregular migrants through Mexico to the United States. Consistent with this information, we observe two major routes illustrated by the Twitter results in Figure 1: one on the East Coast approaching Texas and a second on the Center/West coast to Arizona and California. We

[15] This includes: El Salvador, Nicaragua, Venezuela, Costa Rica, Guatemala, Honduras, Panama, and Mexico (excluded from this analysis since we are tracking migrant routes through Mexico here).
[16] The map in Figure 1 (right side) is from: https://www.bbvaresearch.com/en/publicaciones/map-2020-of-migrant-houses-shelters-and-soup-kitchens-for-migrants-in-mexico

observe a bifurcation in Southern Mexico close to the border with Guatemala. Additionally, we detect peaks of activity at the borders where these routes converge, such as the Tamaulipas and Texas border in the East, as well as the Sonora and Baja California border with Arizona and California. Similarly, we observe that in the central region of the country, migrants also take routes leading to the border areas of Chihuahua, Coahuila, and Tamaulipas with Texas.

Figure 2: Settlement States by Country of Origin



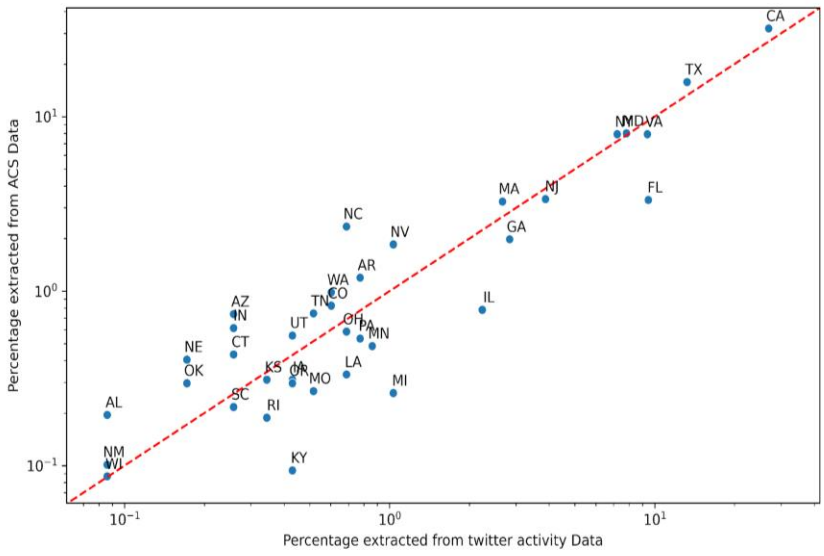
Notes: This figure plots settlement locations for migrants from Costa Rica (CR), Guatemala (GT), Venezuela (VE), Mexico (MX), and El Salvador (SV).

To illustrate how Twitter data can be used to study settlement patterns in the U.S., Figure 2 shows the distribution of settlement locations for migrants during the entire period of study (2015-2022) by state within the U.S. for individuals who were previously identified as residents in El Salvador (SV), Mexico (MX), Venezuela (VE), Guatemala (GT) and Costa Rica (CR). To detect settlement locations (states), the research team used the same procedure developed to code residence in countries of origin.

The results show important differences by country of origin. Specifically, during the period under study, California was the top destination for residents from El Salvador, Mexico, and Guatemala, while Florida was the top destination for migrants from Venezuela and Costa Rica. Texas remained an important destination for migrants from Mexico, and to a lesser extent from the other countries shown in Figure 2, with the exception of Venezuela.

Finally, to validate the state-level settlement estimates based on the Twitter data, we compare the estimates to data from the American Communities Survey (ACS), compiled by the Migration Policy Institute.[17] The results, shown in Figure 3, are plotted on a logarithmic scale for clearer visualization, particularly for states with low percentages. The results show a very close correlation between the Twitter estimates and the ACS data obtained from high-quality surveys conducted by the U.S. Census Bureau.

Figure 3: State-level Migration Estimates from El Salvador, 2015-2022



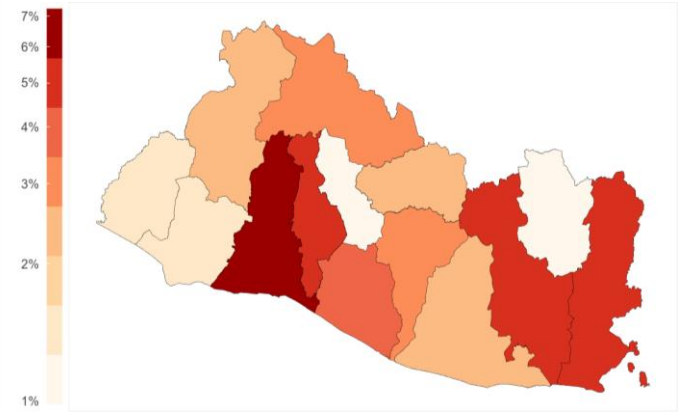
[17] Migration Policy Institute. https://www.migrationpolicy.org/programs/data-hub/charts/us-immigrant-population-state-and-county. Accessed: February 2024.

Finding 3: Twitter data is useful for examining where migrants come from within countries of origin

Finally, Twitter data can be used to study the sub-national origins of migrants to the U.S. Using the geographic information contained in tweets and applying the residency assignment rules mentioned in the methodology section, we can identify specific regions within countries of origin where users resided before emigrating. Figure 4 provides an illustration with data from El Salvador, showing the percentage of migrants by region for the period 2015-2022. The estimates are based on users who subsequently settled in the United States.

The ability to track sub-national location for migrants could be useful for studying how the root causes of migration – e.g., violence, economic shocks, natural disasters, and conflict – operate within countries. While the CBP dataset has historically contained information on prior residence within countries of origin for migrants encountered at the U.S. Southwest border, that data is not publicly available. For this reason, Twitter estimates could provide a complementary data source for certain types of analysis. For example, analysis could compare how the share of migrants coming from specific areas of a country changes in response to changes in sub-national crime rates, economic conditions, or natural disasters that disproportionately affect some areas. At the same time, analysts will need to be mindful of potential biases in the Twitter data, since usage rates are likely to be higher in urban areas in LAC countries.

Figure 4: Region of Origin for Migrants from El Salvador (percentage of total), 2015-2022



4. Conclusion

This project assessed the value of Twitter data for estimating migration from LAC countries. Despite the tremendous potential social media holds for studying human mobility, the report found that several inherent issues – particularly the small size of the population of Twitter users in many LAC countries – limit the value of Twitter data for quantifying migration flows from LAC. Other applications are more promising: for example, Twitter data can be used to track large-

scale outflows during major national crises, as in Venezuela's mass exodus in 2018.18 It can also be used to study migrant routes, settlement patterns in receiving countries, and the local origins of migrants within countries of origin. However, this analysis is limited by the inability to use the data to track flows over time (month and year) and to assess regular vs. irregular flows.

[18] Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet, and J. J. Ramasco. "Migrant mobility flows characterized with digital data." *PLOS ONE* 15, no. 3 (2020): e0230264.

References

Armstrong, Caitrin, et al. “Challenges when identifying migration from geo-located Twitter data.” *EPJ Data Science* 10.1 (2021): 1.

Hawelka, Bartosz, et al. “Geo-located Twitter as proxy for global mobility patterns.” *Cartography and geographic information science* 41.3 (2014): 260-271.

Huang, Xiao, et al. "Twitter reveals human mobility dynamics during the COVID-19 pandemic." *PloS one* 15.11 (2020): e0241957.

Huang, Xiao, et al. “Twitter reveals human mobility dynamics during the COVID-19 pandemic.” *PloS one* 15.11 (2020): e0241957.

Lenormand, M., M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, and J. J. Ramasco. “Cross-checking different sources of mobility information.” *PloS one*, vol. 9, no. 8, p. e105184, (2014).

Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet, and J. J. Ramasco. “Migrant mobility flows characterized with digital data.” *PLOS ONE* 15, no. 3 (2020): e0230264.

Migration Policy Institute. https://www.migrationpolicy.org/programs/data-hub/charts/us-immigrant-population-state-and-county. Accessed: February 2024

Mosley, Mohsen, and David G. Rand. “Who is on Twitter (X)? Identifying demographic of Twitter Users.” Working paper, February 2024.

National Security Council. *U.S. Strategy for Addressing the Root Causes of Migration in Central America*. July 2021.

Rosenzweig, Leah, Parrish Bergquist, Katherine Hoffmann Pham, Francesco Rampazzo, and Matto Mildenberger. “Survey sampling in the Global South using Facebook advertisements.” (2020).

Roy, K.C., M. Cebrian, and S. Hasan. “Quantifying human mobility resilience to extreme events using geo-located social media data.” *EPJ Data Science*, vol. 8, no. 1, pp. 1–15 (2019).

Wang, Yan, and John E. Taylor. "Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake." *Natural hazards* 92 (2018): 907-925.