Measuring Venezuela Emigration with Twitter

Measuring Venezuelan Emigration with Twitter

Ricardo Hausmann, Julian Hinz and
Muhammed A. Yildirim

CID Faculty Working Paper No. 342
May 2018

KIEL WORKING PAPER NO. 2106 | MAY 2018

ABSTRACT

MEASURING VENEZUELAN EMIGRATION WITH TWITTER

Ricardo Hausmann, Julian Hinz, and Muhammed A. Yildirim

Venezuela has seen an unprecedented exodus of people in recent months. In response to a dramatic economic downturn in which inflation is soaring, oil production tanking, and a humanitarian catastrophe unfolding, many Venezuelans are seeking refuge in neighboring countries. However, the lack of official numbers on emigration from the Venezuelan government, and receiving countries largely refusing to acknowledge a refugee status for affected people, it has been difficult to quantify the magnitude of this crisis. In this note we document how we use data from the social media service Twitter to measure the emigration of people from Venezuela. Using a simple statistical model that allows us to correct for a sampling bias in the data, we estimate that up to 2,9 million Venezuelans have left the country in the past year.

Keywords: migration, social media

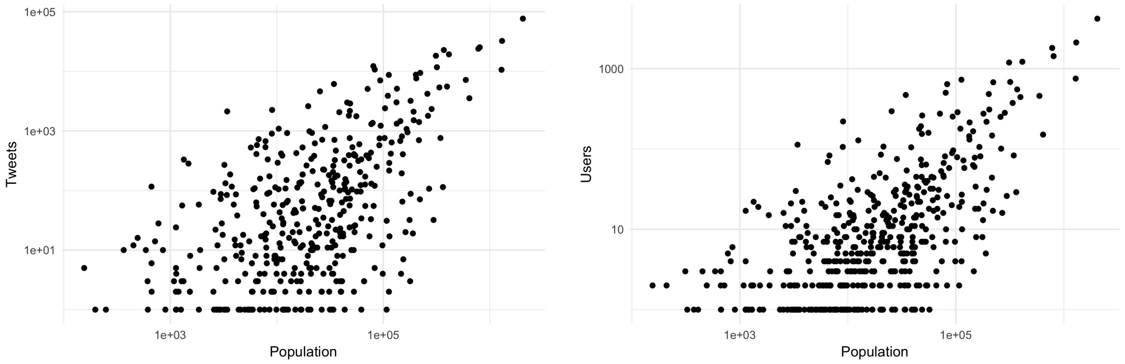
JEL classification: F22, C55

Ricardo Hausmann
Harvard Kennedy School and Center for International Development at Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
United States
Email:
ricardo_hausmann@hks.harvard.edu
www.hks.harvard.edu

Julian Hinz
Kiel Institute for the World Economy
Kiellinie 66
D-24105 Kiel,
Germany
Email:
Julian.hinz@ifw-kiel.de
www.ifw-kiel.de

Muhammed A. Yildirim
Koç Üniversitesi, Turkey
Rumelifeneri Yolu
34450 Sarıyer, İstanbul
Turkey
Email:
mayildirim@ku.edu.tr
www.case.ku.edu.tr

The responsibility for the contents of this publication rests with the author, not the Institute. Since working papers are of a preliminary nature, it may be useful to contact the author of a particular issue about results or caveats before referring to, or quoting, a paper. Any comments should be sent directly to the author.



(a) Population and number of tweets by location **(b)** Population and number of users by location

Venezuela has seen an unprecedented exodus of people in recent months. In response to a dramatic economic downturn in which inflation is soaring,[1] oil production tanking,[2] and a humanitarian catastrophe unfolding,[3] many Venezuelans are seeking refuge in neighboring countries. However, the lack of official numbers on emigration from the Venezuelan government, and receiving countries largely refusing to acknowledge a refugee status for affected people, it has been difficult to quantify the magnitude of this crisis. In this note we document how we use data from the social media service Twitter to measure the emigration of people from Venezuela.

This note is structured as follows. In section 1 we describe the data we use in more detail. In section 2 we outline our quantification strategy and describe the resulting migration figures. Section 3 concludes.

1 Data

We use data collected from the Twitter Streaming API.[4] The API provides a 1% random sample of all geolocalized tweets at any given moment. The geolocation is provided either by the user's device's GPS coordinates, or a self-assigned location.[5] Morstatter et al. (2013) finds that this random sample creates an acurate picture of the entire population of geolocated tweets. Hawelka et al. (2014) use a dataset similar to ours to show global mobility patterns and compare Twitter users' locations to tourism flows. Jurdak et al. (2015) employ a smaller-scale dataset to study city-to-city travel in Australia.

The question that is particularly relevant for this current endeavor is the representativeness of the observed Twitter data for the overall population of Venezuelans. Figures 1a and 1b show the number of tweets and Twitter users mapped against data from the "Gridded Population of the World" project (CIESIN, Columbia University, 2017) at a resolution of 15 arc minutes. The correlation of the local population with the number of tweets is 0.85, the correlation with Twitter users is 0.87.

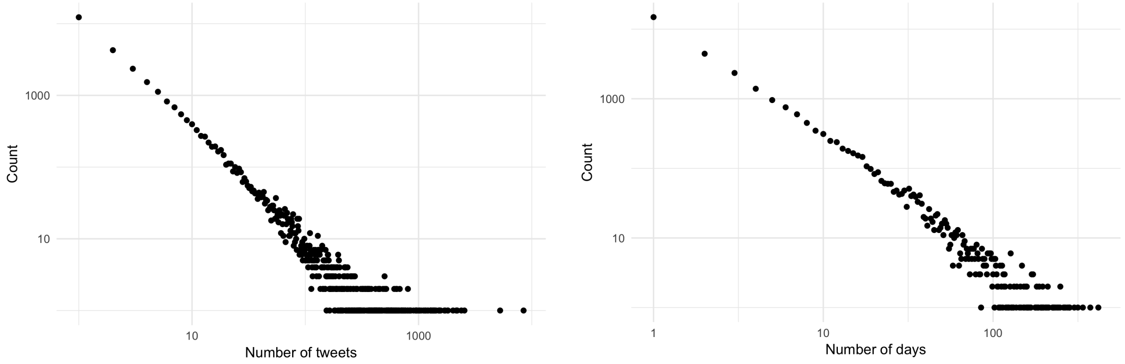
[1] See e.g. `http://www.finanzasdigital.com/2018/05/cendas-canasta-alimentaria-familiar-de-abril-de-2018-se-ubico-en-100-174-98098-bolivares-aumentando-bs-48-131-75770-925-con-respecto-a-marzo-de-2018/`.

[2] See e.g. `http://www.opec.org/opec_web/en/data_graphs/335.htm`.

[3] See e.g. `https://www.catholicnewsagency.com/news/caritas-venezuela-warns-that-280000-children-could-die-of-malnutrition-46072`.

[4] For a detaied description see `https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter`.

[5] As described in more detail below, we restrict the sample to only those tweets where we can be sure to observe a user's true location provided as reported by the GPS coordinates.



(a) Number of tweets per user in the dataset **(b)** Number of days a user is observed in the dataset

1.1 Use of social media in Venezuela

According to the “Digital in 2017 Global Overview report” of the social media agency “We Are Social” and social media marketing company “HootSuite” 44% of Venezuelans are using social media and 35% do so from a mobile device.[6] According to a 2016 report by “ Tendencias Digitales”, 56% of internet users in Venezuela use Twitter or comparable social media services, [7] which together mirror closely other sources that report a penetration rate of 26% of Twitter in Venezuela.[8]

1.2 Descriptive statistics

For the purpose of this project, we restrict the dataset to a sample of tweets from users that at some point since February 2017 tweeted from Venezuela. This yields a total of roughly 5.4 million tweets. To ensure that we strictly observe human users we follow Chu et al. (2012) and restrict the sample to those tweets sent out from Twitter clients for mobile phones, yielding about 490.000 tweets. These tweets were sent out from about 30.000 Twitter users.

2 Quantifying Venezuelan emigration

2.1 Ad-hoc emigration numbers and destinations

In order to measure the emigration from Venezuela we first need to determine who could have emigrated. We define as Venezuelan anyone who tweeted exclusively from Venezuela in the time period between February 1 and April 30, 2017. In the rest of the analysis, unless otherwise stated, we focus on these 9623 Twitter users.

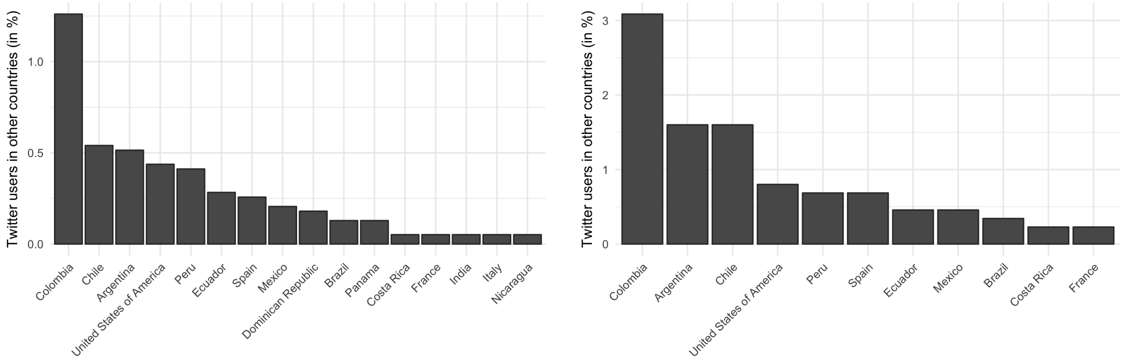
One ad-hoc way to measure the emigration is to look at the last recorded location of these people. In total, 3887 show up in the data at least once more after April 30, 2017. 94% of them have their last recorded location in Venezuela. Figure 3a shows the distribution of countries of the remainder. These numbers, in particular the one for Venezuela, may however contain people that stopped tweeting altogether.

Another way to measure the emigration is to look at the same time window from February 1 and April 30 in 2018. In total, 875 users from the first period also show up in this second period. Out

[6] See https://www.slideshare.net/wearesocialsg/digital-in-2017-south-america.

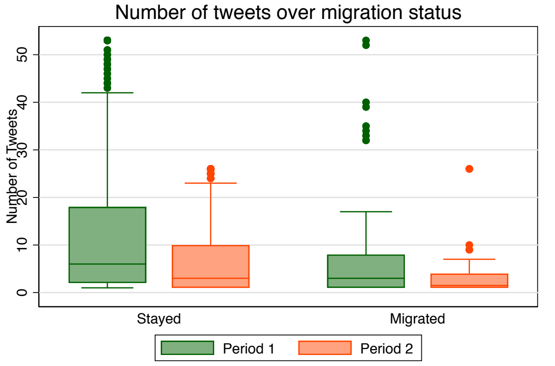
[7] See http://tendenciasdigitales.com/web/wp-content/uploads/2017/02/Reporte_Penetracion_vzla_2016.pdf.

[8] See https://www.statista.com/statistics/754520/venezuela-penetration-social-networks/.



(a) Distribution of countries of last recorded locations of users outside Venezuela

(b) Distribution of countries of users between February and April 2018 outside Venezuela



Note: Because of the heavy tail, the users who are at the top 90% of the tweet counts are top-coded t

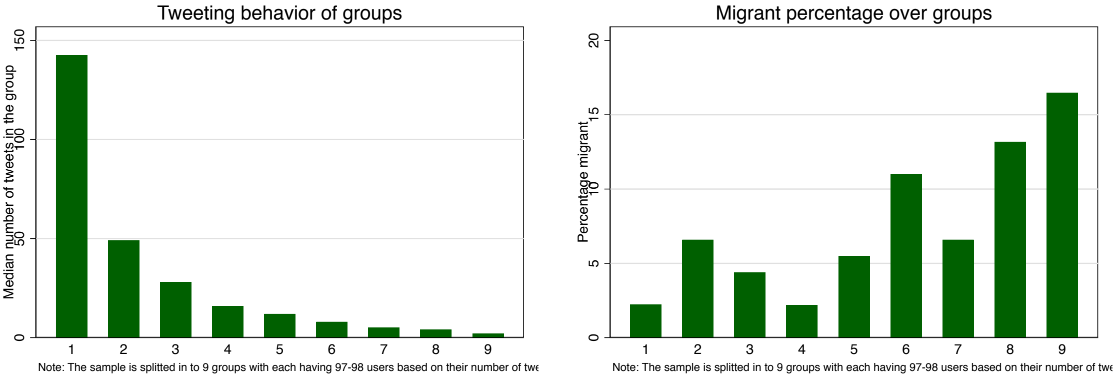
Figure 4: Tweets by migrants and non-migrants in two periods

of these people, 86% tweeted exclusively from Venezuela, 4% tweeted from Venezuela and at least one other country, while 10% tweet exclusively from outside of Venezuela. Figure 3b shows the distribution among these countries.

2.2 Accounting for heterogeneity of Tweet frequency

However, the above figures do not take into account the fact the heterogeneity of Twitter users with respect to the frequency of tweeting. To get a more precise estimate of the percentage of the emigrants leaving Venezuela, we narrow our sample even further to users who exclusively tweeted from Venezuela, which reduced the number of users to 818. Out of these users, 62 tweeted exclusively outside of Venezuela. We call these users “migrants”.

Figure 4 shows that the distribution of tweets shows a significantly different pattern between users who stayed in Venezuela and those who migrated from Venezuela. We rank the users based on their combined tweet counts in both periods and divided the users into 9 groups with 97 to 98 users in each group. Figure 5a shows the median of the total number of tweets for each user in each group. The distribution of the medians is consistent with the overall power-law like structure of the number of tweets per user displayed in 2a. Surprisingly, though, when we calculate the percentage of migrants from each group, the last two groups, which consists of users with the least number of tweets, have a higher ratio of individuals who become migrants.



(a) Median number of tweets from each group.

(b) Percentage of migrants from each group.

2.3 A simple statistical model

We now set up a simple statistical model that yields a *weight* attributed to each user that corrects for this sample bias. Suppose that the probability of an individual i tweeting exactly x tweets in a three-month period is given by,

$$p_{i,x} = Pr\{tw_i = x\}$$

where tw_i is the random variable denoting the tweets by individual i. We will assume that this probability distribution remains constant across all observed periods.

We are interested in estimating the fraction of people that moved out of Venezuela between two periods. We have Twitter sample data for two three-month periods: an initial one from February to May 2017, and a latter one from February to May 2018. The problem is that Twitter only provides a $s = 1\%$ sample of all Tweets, independent of the user, with $q = (1 - s) = 99\%$ of Tweets not being reported.

Let U^0 (U^1) denote the set of all users that are observed at least once in the initial (latter) sample. Let x denote the tweets sent in the initial period and y those sent in latter period. Hence the probability of observing an individual who tweeted x_i tweets during the initial period can be written as

$$Pr\{i \in U^0 | tw_i^0 = x\} = 1 - q^x.$$

The same individual will be in the second sample if out of her y_i tweets, at least one is in the 1% sample provided by Twitter, such that

$$Pr\{i \in U^1 | tw_i^1 = y\} = 1 - q^y$$

We can write the probability of an individual to be present in both samples as, assuming the independence between two samples, as

$$
\begin{aligned}
&Pr\{i \in U^0 \text{ and } i \in U^1\} \\
&\qquad = \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} Pr\{i \in U^0 | tw_i^0 = x\} Pr\{tw_i^0 = x\} Pr\{i \in U^1 | tw_i^1 = y\} Pr\{tw_i^1 = y\} \\
&\qquad = \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} p_{i,x}(1-q^x)p_{i,y}(1-q^y) \\
&\qquad = \sum_{x=0}^{\infty} p_{i,x}(1-q^x) \sum_{y=0}^{\infty} p_{i,y}(1-q^y) \\
&\qquad = (1-E_i[q^x])^2
\end{aligned}
$$

where E_i is the expected value operator taking the expectation over p_i. We can use the probability generating function defined as $G_i(q) \equiv E_i[q^x]$ to re-write the equation above as

$$Pr\{i \in U^0 \text{ and } i \in U^1\} = (1 - G_i(q))^2.$$

Furthermore, we can also model the individuals tweeting behavior as a Poisson process. Let us assume that each individual has a Poisson tweet rate in a three month period with λ_i. With the Poisson distribution, we can then easily write the probability generating function as

$$G_i(q) = e^{-\lambda_i(1-q)} = e^{-\lambda_i s}.$$

Hence, we can write the probability of being observed in both periods as

$$Pr\{i \in U^0 \text{ and } i \in U^1\} = (1 - e^{-\lambda_i s})^2 \tag{1}$$

with $s = 0.01$ in our case.

2.4 Outflow over time

We compute the probabilities expressed in equation (1) for each user. In table 1 we show the results of the estimation of the number of emigration and immigration for Venezuela and other countries. The unweighted measure denotes the simple shares of users outside of the respective country in the second period for the emigration number, in the first period for the immigration number. The *weighted* number reports the comparable figure with the probabilities of being observed in the data as the weights.

When looking at the emigration numbers, Venezuela does not seem extraordinary compared to other Latin American countries like Colombia, Argentina or Brazil. Germany even surpasses Venezuela's number. However, this picture changes dramatically when contrasting this number against implied immigration numbers. Here, other countries report numbers similar to their respective emigration numbers. This suggests that, owing to the limited time windows, our estimation may pick up significant yet regular short-term movements of people. Venezuela, however, reports a large difference between emigration and immigration numbers, even when expanding the time window to 6 months (column 6). This strongly suggests that other mechanisms than short-term travel is at play.

To put these numbers in perspective, we compute some back-of-the-envelope estimates of absolute migration numbers. Venezuela has a population of 30 million people, 26 % of which are reported to be active users of Twitter. Assuming that the distribution of geocoded Tweets is representative

		(1) Venezuela	(2) Colombia	(3) Argentina	(4) Brazil	(5) Germany	(6) Venezuela	(7) Colombia
Emigration	*unweighted*	6,76%	7,78%	7,62%	3,88%	11,59%	6,99%	6,06%
	weighted	9,59%	7,84%	7,92%	3,97%	13,18%	7,98%	6,10%
Immigration	*unweighted*	2,01%	5,21%	10,48%	3,59%	11,27%	1,77%	5,21%
	weighted	2,22%	5,48%	10,70%	3,67%	12,41%	1,70%	5,37%
Difference	*unweighted*	-4,75%	-2,57%	2,86%	-0,29%	-0,32%	-5,22%	-0,85%
	weighted	-7,37%	-2,36%	2,78%	-0,30%	-0,77%	-6,28%	-0,73%
Annualized weighted perc.		-9,7%	-3,1%	3,7%	-0,4%	-1%	-12,1%	-1,4%
Period 1		02–04/17	02–04/17	02–04/17	02–04/17	02–04/17	12/16–04/17	12/16–04/17
Period 2		02–04/18	02–04/18	02–04/18	02–04/18	02–04/18	12/17–04/18	12/17–04/18

Source: Authors' calculations.

Table 1: Computed emigration and immigration numbers

of all Tweets and that the rate of emigration has remained constant for the last 12 months, $1-(1-0.0737)^{12/9}=9.7\%$ of users have left the country over the past year.[9] Assuming this trend is representative for all of Venezuelans, which does not appear to be a strong assumption as laid out above, the number stands at roughly 2,9 million people. Assuming only Twitter users did so, this still constitutes at a minimum about 815 thousand people.

3 Conclusion

In this note we outline how data from the social media service Twitter can be used to shed light on the magnitude of the emigration of Venezuelans in response the the political and humanitarian crisis in the Latin American country. Using a simple statistical model that allows us to correct for a sampling bias in the data, we estimate that approximately 2,9 million Venezuelans have left the country in the past year. Even in the highly unlikely scenario that only Twitter users emigrated, the number of emigrants would still stand at about 815 thousand people, placing it among the highest estimates of the outflow of people from Venezuela.

[9] Note that the numbers for column (6) for longer time windows, but only a 6 months period to be counted as migrant, implies a similar, even slightly higher annual rate at $1-(1-0.0628)^{12/6}=12.1\%$.

References

Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing 9*(6), 811–824.

CIESIN, Columbia University (2017, 20180523). Gridded population of the world, version 4 (gpwv4): Population count, revision 10.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science 41*(3), 260–271. PMID: 27019645.

Jurdak, R., K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth (2015, 07). Understanding human mobility from twitter. *PLOS ONE 10*(7), 1–16.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley (2013). Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. *CoRR abs/1306.5204*.