DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

WORKING WITH THE TWITTER-API

FERTILITY, MORTALITY AND MIGRATION FROM TWEETS

TOM THEILE

EDSD NOVEMBER 2022 BARCELONA

LAB OF DIGITAL AND
COMPUTATIONAL DEMOGRAPHY

TWITTER DATA AND MIGRATION RESEARCH

- Information from Twitter data that can be used in migration research:
- ▶ Geo-tagged tweets
- ▶ Self-declared location information from profile
- ▶ Social networks: followers and friends
- ▶ Name
- ▶ Language
- Anything else?

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Potential problems:

- Tourists
- Travel blogs
- Bots
- …

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Definition of a migrant

- Who is an Immigrant?:
- “A person who moves to a country other than that of his or her usual residence for a period of at least a year.”
- - The period of stay determines whether the immigrant is a long-term immigrant or a short-term immigrant. Any period under twelve months is considered a short-term migrant, whereas any period over twelve months is considered a long-term migrant[1].
- [1]: UNDESA, Recommendations on Statistics of International Migration, Revision 1 (1998) para. 36

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Definitions of migrants on Twitter in the literature

▶ “A Twitter user has the nationality that others believe you have.” (Huang et al., 2014)

▶ “Any individual leaving Venezuela during the time window of observation.”(Mazzoli et al., 2020)

▶ “Anyone who tweeted exclusively from Venezuela in the time period between Feb. 1 and April 30 2017.” (Hausmann et al., 2018)

▶ “Migrants are users that are identified as people who moved to a different country for at least one of the 4-month periods.” (Zagheni et al., 2014)

▶ “A migrant is a person that has the residence different from the nationality.” (Kim et al., 2020)

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Identifying migrants on Twitter (Kim et al., 2020)

“A migrant is a person that has the residence different from the nationality.”

▶ Country of residence: “the country with the longest length of stay”

▶ Country of nationality: “the ensemble of features that make a person feel like they belong to a certain country”

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Identifying migrants (Mazzoli et al., 2020)

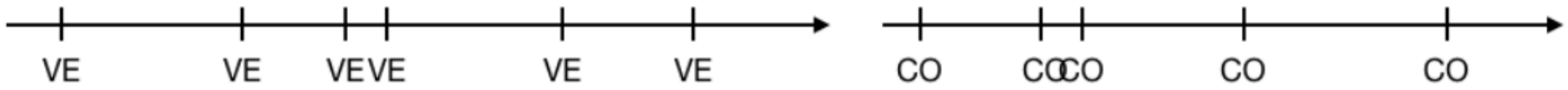


Figure: Mock example of geo-tagged tweet timeline

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Identifying migrants on Twitter (Kim et al., 2020) - Example

Twitter page: ▶ Homepage: ▶ Location:Germany ▶ Time zone:CET ▶ Interface language: English

▶ Tweets languages: (1) English: 97% (2) Italian: 2% (3) Korean: 1%

▶ Follower locations: (1) US: 151 (2) KR: 72 (3) DE: 3

▶ Following locations: (1) US: 80 (2) KR: 40 (3) IT: 30

▶ Tweets from (1) IT: 30 (2) KR:29

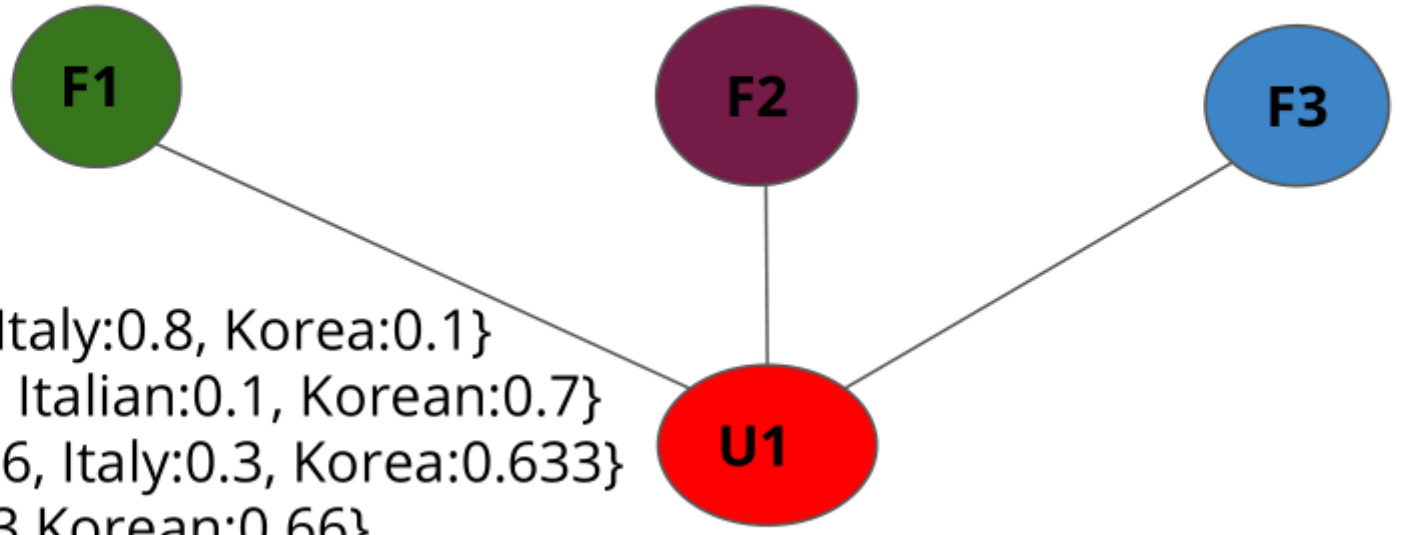
Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

loc^{F1}={France: 0.2, Italy:0.8}
$lang^{F1}$={Italian:1}

loc^{F2}={Italy:0.1,Korea:0.9}
$lang^{F2}$={Korean:1}

loc^{F3}={Korea:1}
$lang^{F3}$={Korean:1}



loc^{U1}={France: 0.1, Italy:0.8, Korea:0.1}
$lang^{U1}$={French:0.2, Italian:0.1, Korean:0.7}
$floc^{U1}$={France:0.066, Italy:0.3, Korea:0.633}
$flang^{U1}$={Italian:0.33,Korean:0.66}

Adapted from: Jisu Kim 11/2022

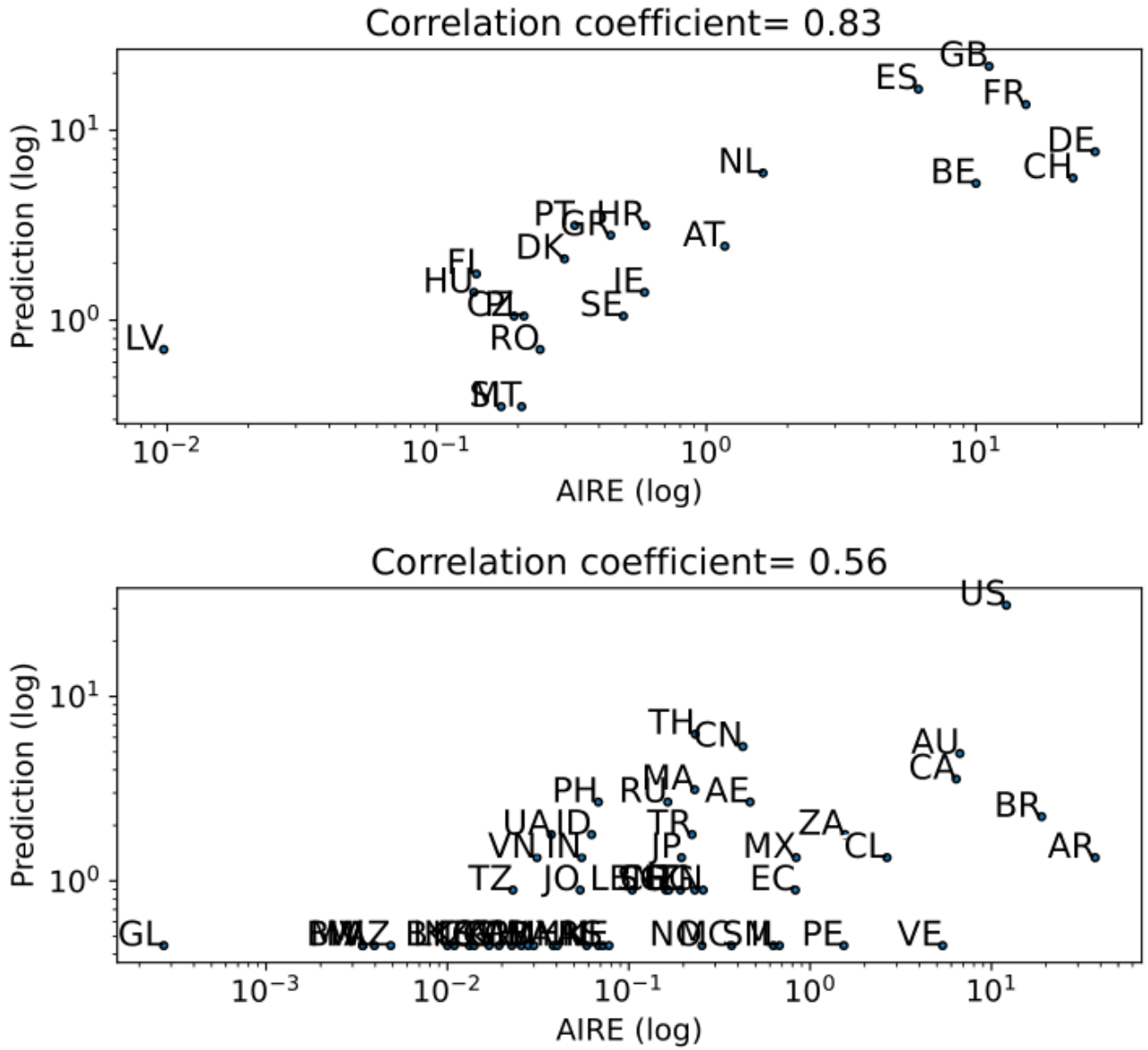
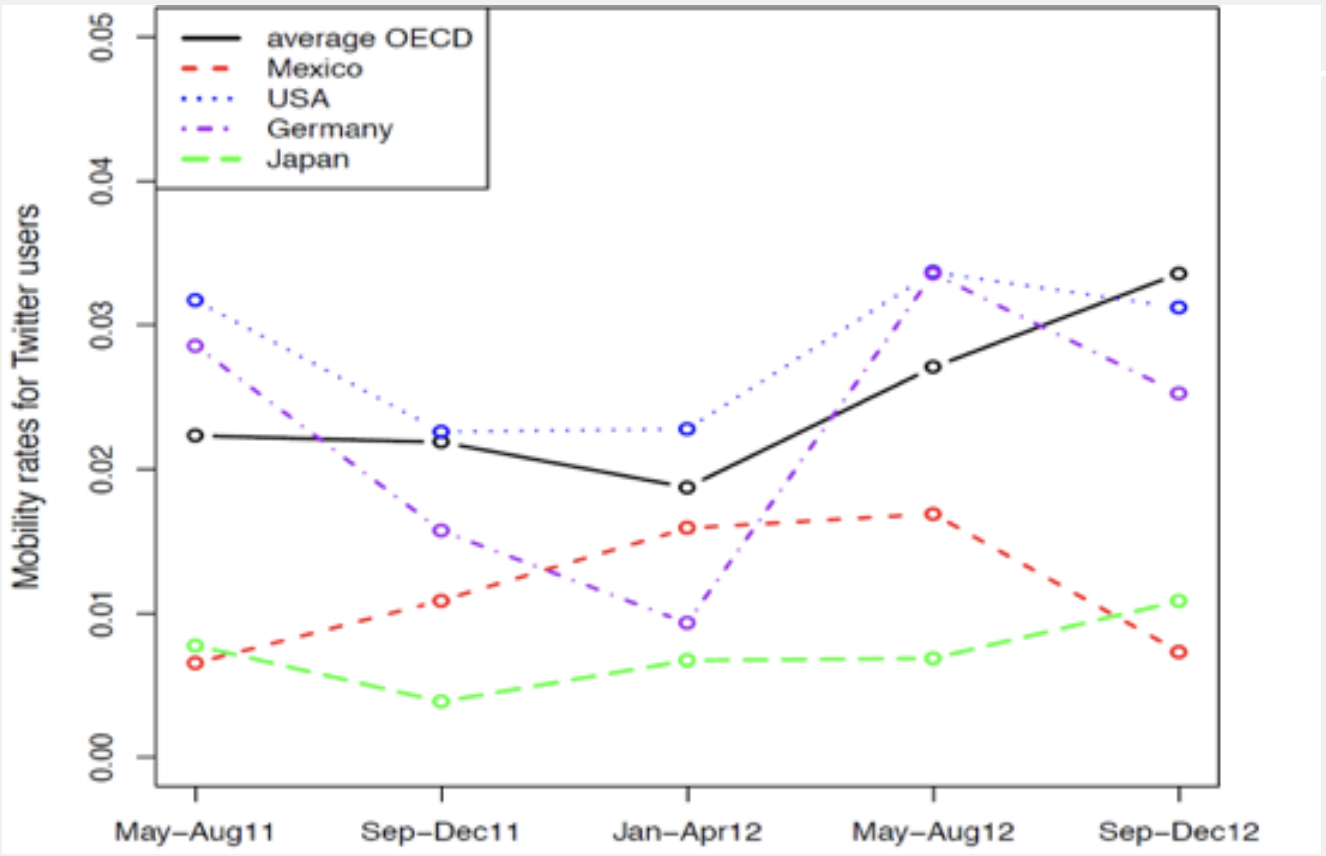


Figure: Correlation between predicted data and official statistics Kim et al., 2020

Adapted from: Usui Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

Inferring international and internal migration patterns from Twitter data (Zagheni et al., 2014)

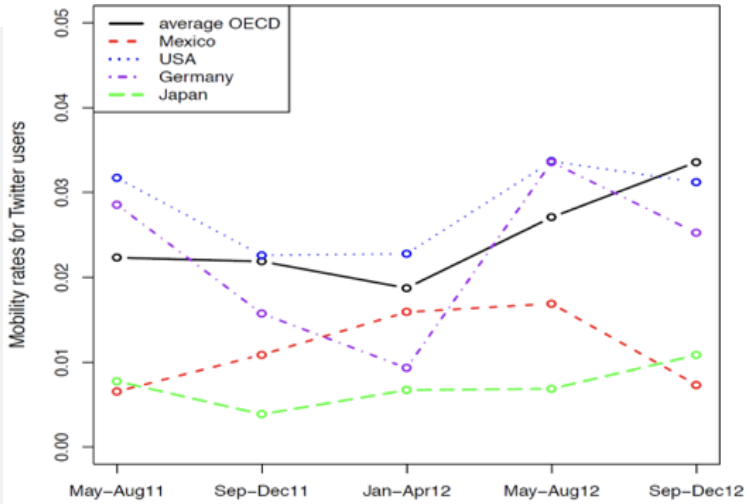


Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH
DIFFERENCE-IN-DIFFERENCES

Out-migration rates clearly an overestimate

- ▶ Non-representative user set
- ▶ Selection bias is changing over time
- ▶ Focus on between-country differences



$$\hat{\delta}_c^t = (m_c^t - m_{oecd}^t) - (m_c^{t-\Delta_t} - m_{oecd}^{t-\Delta_t})$$

→ Diff-in-diff estimator to evaluate relative changes in trends

Adapted from: Jisu Kim 11/2022

TWITTER DATA AND MIGRATION RESEARCH

TWITTER DATA AND MIGRATION RESEARCH

SENTIMENT ANALYSIS (ARCILA-CALDERÓN ET AL., 2021)

Hate speech towards migrants/refugees? What are the main topics behind the messages in Twitter in Spain with hate speech against refugees?

Topic	Most Common Words and Their Frequency within the Topic (the Underlined Words are the Most Determinant for the Labelling of the Topic)
Pull effect and consequences	"immigrants" (0.008) + "spain" (0.007) + "boat" (0.005) + "effect" (0.005) + "pull" (0.004) + "goes" (0.004) + "valencia" (0.004) + "consequences" (0.004) + "europe" (0.004) + "concentration" (0.004)
Pull effect and not welcoming "illegals"	"immigrants" (0.013) + "people" (0.008) + "spain" (0.008) + "pull" (0.007) + "effect" (0.007) + "illegal" (0.007) + "protectyourborders" (0.007) + "spaniards" (0.006) + "harbor" (0.006) + "aquariusnotwelcome" (0.006)
Not welcoming and terrorism	"go" (0.008) + "spain" (0.008) + "people" (0.006) + "aquariusnotwelcome" (0.005) + "country" (0.005) + "boko" (0.005) + "haram" (0.005) + "immigrants" (0.005) + "boat" (0.004) + "have" (0.004)
Smugglers and NGOs	"immigrants" (0.023) + "spain" (0.010) + "spaniards" + "come" (0.008) + "mafias" (0.008) + "people" (0.007) + "go" (0.006) + "illegal" (0.006) + "ngos" (0.006) + "boat" (0.006)
Money and jobs	"refugees" (0.011) + "immigrants" (0.009) + "pay" (0.007) + "spain" (0.007) + "countries" (0.005) + "boat" (0.005) + "spaniards" (0.005) + "people" (0.005) + "solution" (0.005) + "work" (0.005)
Entrance to Europe	"spain" (0.018) + "immigrants" (0.009) + "europe" (0.006) + "boat" (0.006) + "valencia" (0.005) + "spaniards" (0.005) + "government" (0.005) + "immigration" (0.005) + "north" (0.004) + "millions" (0.005)

QUANTIFYING MISSING MINORS IN GUATEMALA

We downloaded all tweets with images from an offcial GUA-Government acccount
All the important information was stored in the images. Because of that, we use OCR to read the Text in the images
Google-maps Geocoding API to get GPS-Coordinates from the Last-Seen-Place-string
Diego asked the goverment agency for all data about missing persons. It took time but we got it!
Comparison: Governemnt data <-> Twitter data!
The Twitter data has more information attached to every individual, individual based
But also limited to minors

- Source:

MISSING MINORS IN GUATEMALA

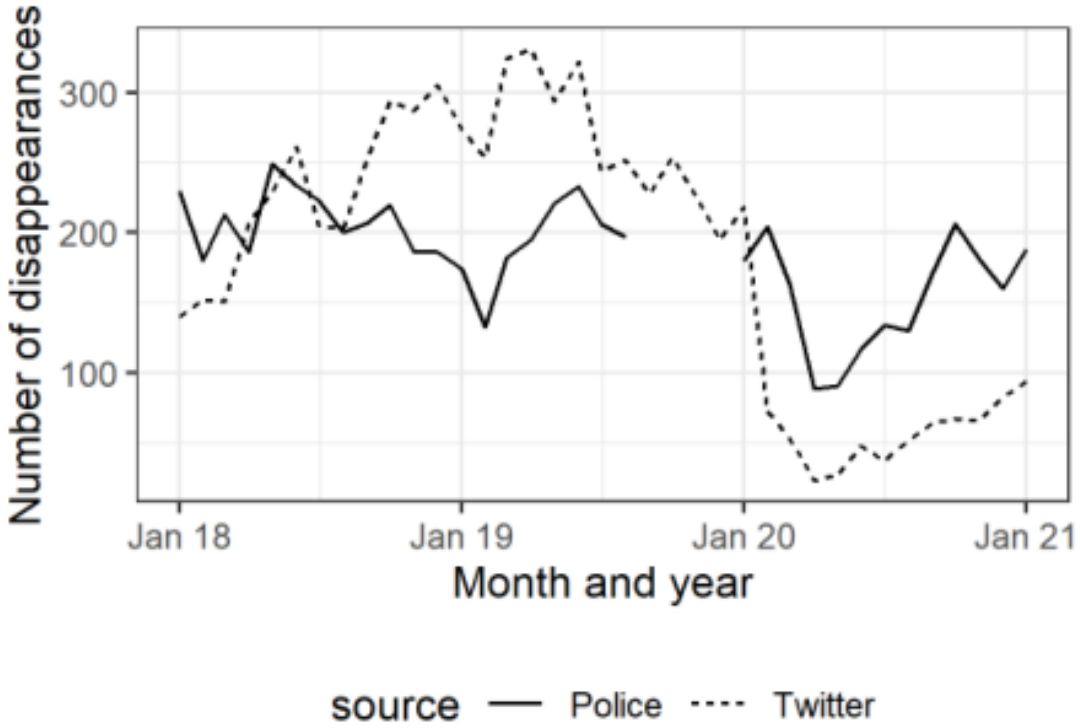


Figure 2: Number of missing children by month: a comparative overview of data from the Guatemalan National Police and *Alerta Alba-Keneth* Twitter data. Note that data for May-December 2019 were not provided by the National Police.

QUANTIFYING MISSING MINORS

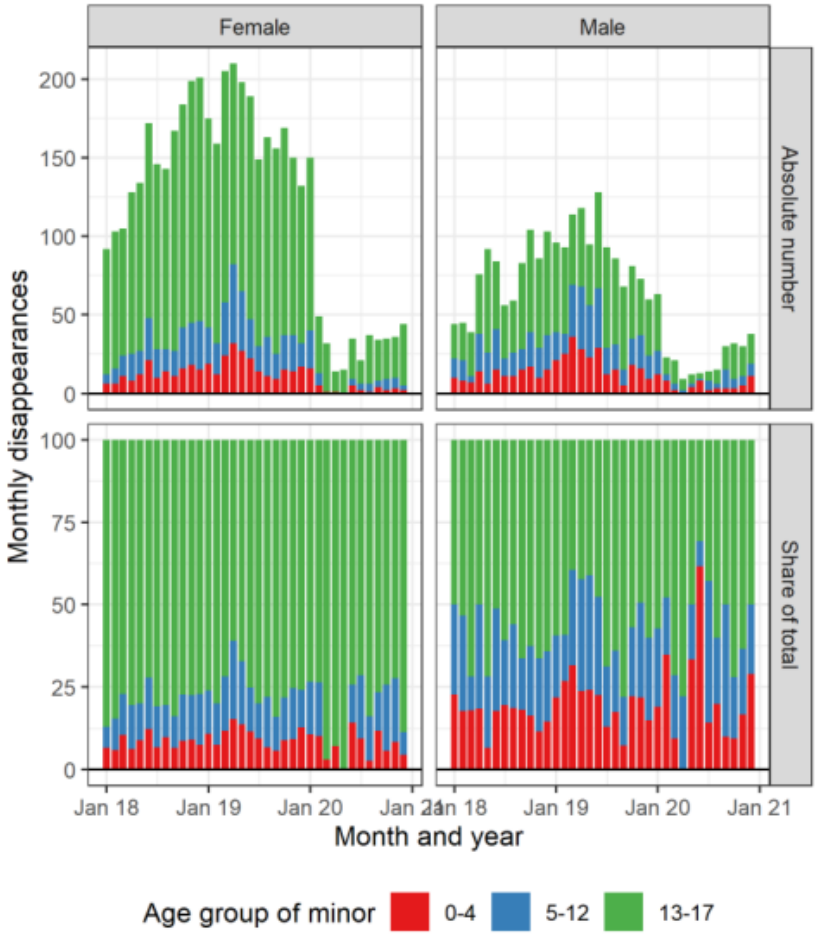


Figure 3: Age and sex distribution of the number of missing children by month of reported disappearance (2018 - 2020) according to the *Alerta Alba-Keneth* Twitter data.

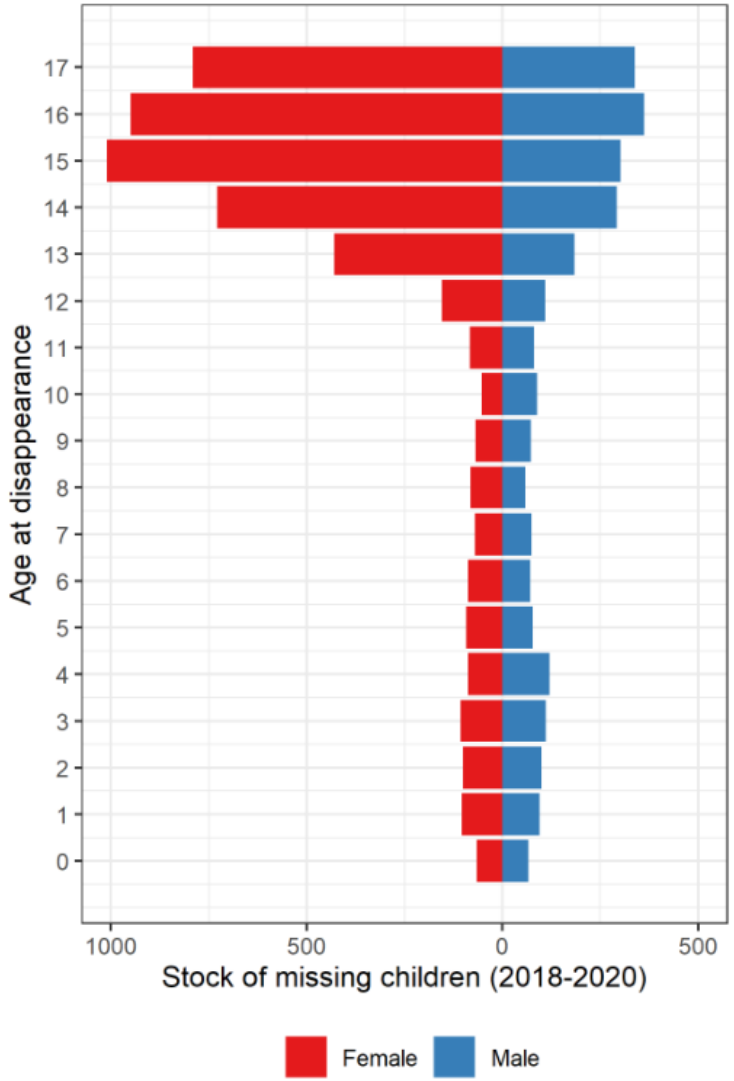


Figure 4: Stock of missing children: cumulative number of disappearances (2018 - 2020) according to the *Alerta Alba-Keneth* Twitter data.

QUANTIFYING MISSING MINORS IN GUATEMALA

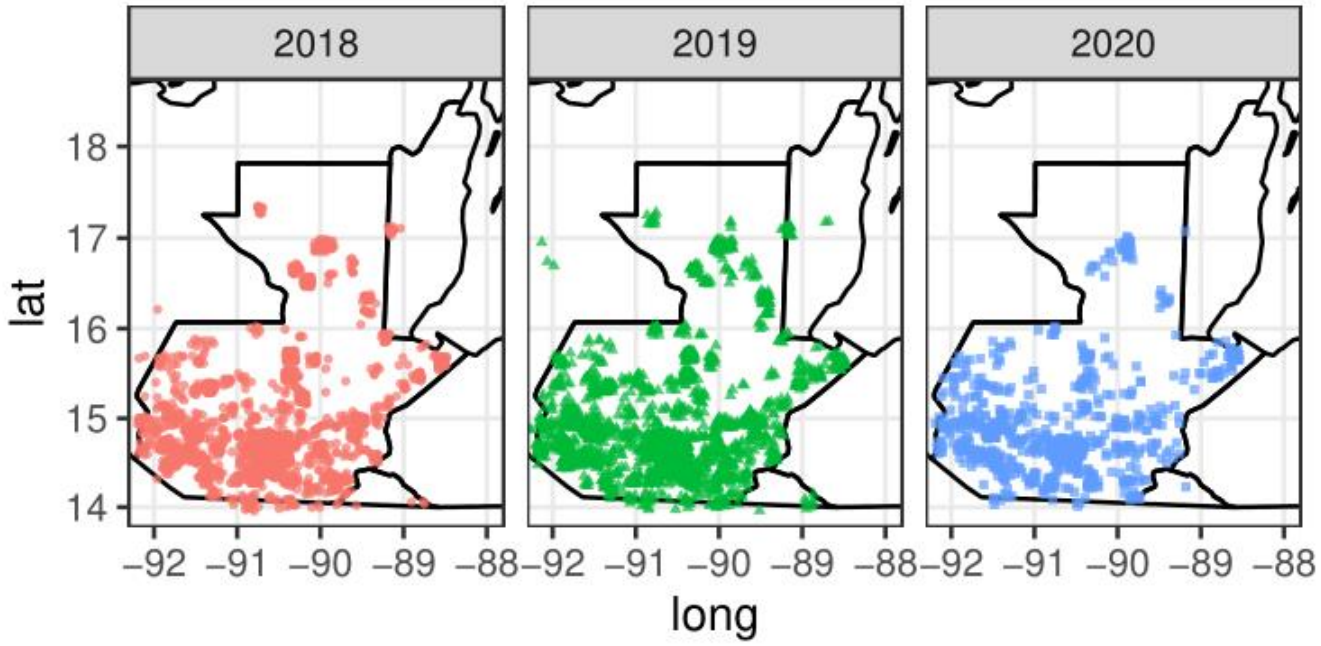


Figure 5: Geographic distribution of the number of reported missing children (2018 - 2020) according to the *Alerta Alba-Keneth* Twitter data. Location data were extracted from the "Place of disappearance" field of the tweet.

THE TWITTER API

https://developer.twitter.com/en

Different access levels (Commercial, free, academic access)

large documentation (but not always easy to find the right information): https://developer.twitter.com/en/docs/twitter-api

Access to tweets, tweet-counts, users, timelines, trends and more

Also possible: creating tweets, sending messages, following people

„Over 11% of Twitter's revenue in FY 2021, or $571.8 million, was from data licensing and other sources.“ (https://www.investopedia.com/ask/answers/120114/how-does-twitter-twtr-make-money.asp)

THE TWITTER API

https://developer.twitter.com/en/products/twitter-api

V1.1 and V2 (since November 2021)

Several levels of access limitations

V2 Access Levels

Essential

With Essential access, you can now get access to Twitter API v2 quickly and for free!

- Retrieve 500,000 Tweets per month
- 1 Project per account
- 1 App environment per Project
- Limited access to standard v1.1 (**only media endpoints**)
- No access to premium v1.1, or enterprise

Elevated

With Elevated access, you can get free, additional access to endpoints and data, as well as additional App environments.

- Retrieve 2 million Tweets per month
- 1 Project per account
- 3 App environments per Project
- Access to standard v1.1, premium v1.1, and enterprise

Academic Research

If you qualify for our Academic Research access level, you can get access to even more data and advanced search endpoints.

- Retrieve 10 million Tweets per month
- Access to full-archive search and full-archive Tweet counts
- Access to advanced search operators

Enterprise: Gnip 2.0

Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data.

Learn more >

Premium v1.1

The premium v1.1 endpoints offer scalable access to Twitter data for those looking to grow, experiment, and innovate by using historical search and subscribing to user activities.

Learn more >

Standard v1.1

The standard v1.1 endpoints were launched in 2012 and enables you to post, interact, and retrieve data for resources such as Tweets, Users, Direct Messages, Lists, Trends, Media, and Places.

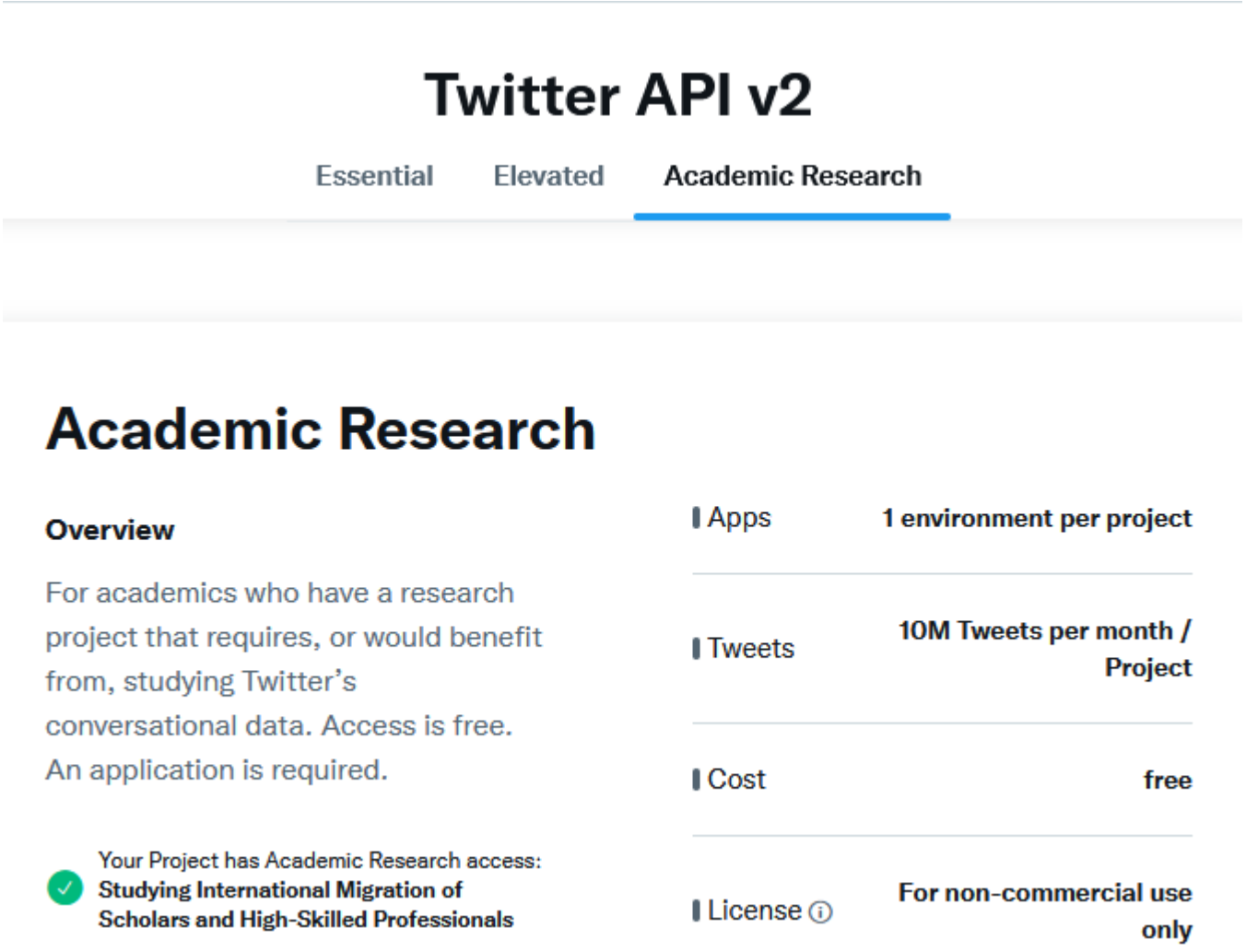
Learn more >

THE TWITTER API

https://developer.twitter.com/en/portal/products/academic-research

10 Million Tweets per month

Free for researchers



THE TWITTER API

https://developer.twitter.com/en/portal/products/academic-research

10 Million Tweets per month

Free for researchers

Academic Research features

Tweets

Tweets lookup

Retrieve detailed, up-to-date information about a specific Tweet, or Tweets.

Read the docs

Endpoints	Rate limit	Tweet cap	Special attributes
GET /2/tweets/:id	300 requests / 15 mins PER APP 900 requests / 15 mins PER USER	no	-
GET /2/tweets	900 requests / 15 mins PER USER 300 requests / 15 mins PER APP	no	-

THE TWITTER API

https://developer.twitter.com/en/portal/products/academic-research

10 Million Tweets per month

Free for researchers

Full-archive-search!

(The normal API is limited to the last 7 days)

Search Tweets

Search for historical Tweets, using advanced filtering tools to refine your search criteria.

Read the docs

Endpoints	Rate limit	Tweet cap	Special attributes
RECENT SEARCH			
GET /2/tweets/search/recent	180 requests / 15 mins **PER USER** 450 requests / 15 mins **PER APP**	yes	• 10 default results per response • 100 results per response • enhanced operators • 1024 query length
FULL-ARCHIVE SEARCH			
GET /2/tweets/search/all	300 requests / 15 mins **PER APP** 1 requests / second **PER USER** 1 requests / second **PER APP**	yes	• 500 results per response • 10 default results per response • enhanced operators • 1024 query length

THE TWITTER API

https://developer.twitter.com/en/portal/products/academic-research

10 Million Tweets per month

Free for researchers

Full-archive-search!

(The normal API is limited to the last 7 days)

Tweet counts

Understand and retrieve the volume of Tweet data for a search query, using the same advanced filtering tools available on the Search Tweets endpoints.

Read the docs

Endpoints	Rate limit ⓘ	Tweet cap ⓘ	Special attributes
RECENT TWEET COUNTS			
GET /2/tweets/counts/recent	300 requests / 15 mins **PER APP**	no	• enhanced operators ⓘ • 1024 query length ⓘ
FULL-ARCHIVE TWEET COUNTS			
GET /2/tweets/counts/all	300 requests / 15 mins **PER APP**	no	• enhanced operators ⓘ • 1024 query length ⓘ

TWITTER QUERIES

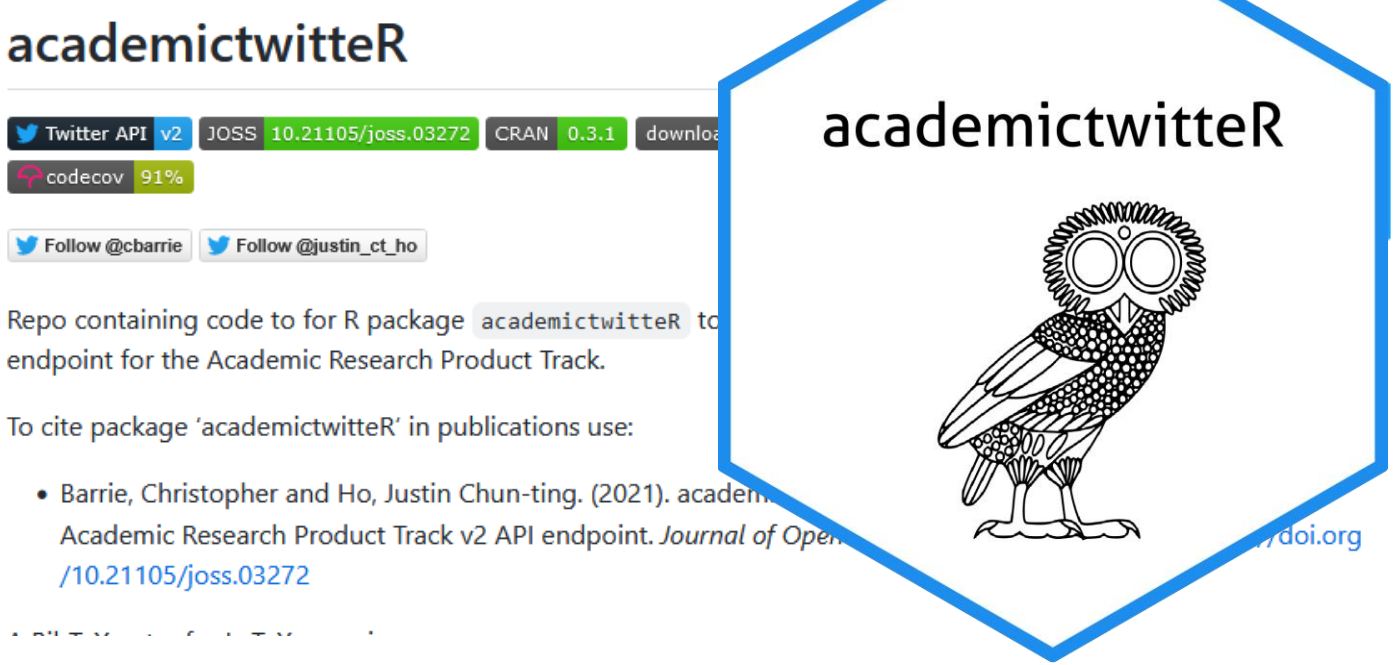
- https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query
- “computational demography" has:mentions (has:media OR has:links) -facebook

THE TWITTER API – ACCESS WITH ACADEMICTWITTER

https://github.com/cjbarrie/academictwitteR

Academictwittr is an easier to use alternative R-package with functions to query the twitter-API.

- Advantages: easier to use
- Disadvantage: you have to trust the package.
- The Twitter API might change and the package then stops working. Your own code is probably easier to adapt to the new changes



REFERENCES

Arcila-Calderón, Carlos et al. (2021). “Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius”. In: Sustainability 13.5, p. 2728.

Hausmann, Ricardo et al. (2018). Measuring venezuelan emigration with twitter. Tech. rep. Kiel Working Paper.

Huang, Wenyi et al. (2014). “Inferring nationalities of twitter users and studying inter-national linking”. In: Proceedings of the 25th ACM conference on Hypertext and social media, pp. 237–242.

Kim, Jisu et al. (2020). “Digital footprints of international migration on twitter”. In: International Symposium on Intelligent Data Analysis. Springer, pp. 274–286.

Kim, Jisu et al. (2021). “Home and destination attachment: study of cultural integration on Twitter”. In: arXiv preprint arXiv:2102.11398.

Lamanna, Fabio et al. (2018). “Immigrant community integration in world cities”. In: PloS one 13.3, e0191612

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Research Software Engineer

theile@demogr.mpg.de

THANK YOU!