MIRREM

Measuring Irregular Migration

www.irregularmigration.eu

Estimating Irregular Migrant Stocks Using Social Media Data and Machine Learning

MIrreM Briefing Paper

AUTHORS:

Alejandra Rodríguez-Sánchez

Jasper Tjaden

Funded by the European Union

Co-funded by:

UK Research and Innovation

Toronto Metropolitan University — Canada Excellence Research Chair in Migration & Integration

Deliverable Information:

Project Acronym:	Measuring irregular migration and related policies (MIrreM)
Project No.	101061314
WP	WP6 - Methods Innovation Lab
Deliverable Type:	Briefing Paper
Deliverable Name	D6.2 PS3 - Estimating irregular migrant stocks using social media data and machine learning
Version:	1
Date:	07/04/2025
Responsible Partner:	University of Potsdam (UP)
Contributing Partners:	
Authors	Alejandra Rodríguez-Sánchez (UP) and Jasper Tjaden (UP)
Reviewers:	Denis Kierans (UOXF) and Jill Ahrens (UWK)
Dissemination Level:	Public

Revision History:

Version	Date	Author	Organisation	Description
1	16/12/2024	Alejandra Rodríguez-Sánchez Jasper Tjaden	University of Potsdam	Draft version
1	18/12/2024	Jill Ahrens	UWK	Review and formatting check
1	14/01/2024	Albert Kraler	UWK	Review
1	22/01/2025	Alejandra Rodríguez-Sánchez Jasper Tjaden	University of Potsdam	Revised Version
1	15/03/2025	Denis Kierans	UOXF	Review
1	02/04/2025	Adriana Harm	UWK	Formatting check
1	07/04/2025	Alejandra Rodríguez-Sánchez	University of Potsdam	Published version 1

Executive Summary

Irregular migration is a sensitive policy issue around the world, often exacerbated by the lack of data. In data scarce contexts, triangulation relying on multiple sources and data points is the best strategy to arrive at evidence-based decisions. We want to contribute to this process by introducing a novel method for estimating the population size of migrants lacking legal residence, combining official statistics on international migrant stock, global social media data (Facebook), and predictive analytics through machine learning algorithms. We attain estimates of the size of the irregular migrant population for the 40 largest destination countries and, for some cases, the composition of the irregular migrant population by world regions of origin and destination. The estimates we obtain are consistent with alternative benchmarks in countries with available estimates. Inconsistencies can be explained by data limitations. We discuss potential limitations of the approach and broader implications for policy and migration research.

Table of contents

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

PIRM	Predictive Indirect Residual Method
ML	Machine Learning
UN	United Nations
Xgboost	Extreme Gradient Boosting

THE MIRREM PROJECT

MIrreM examines estimates and statistical indicators on the irregular migrant population in Europe as well as related policies, including the regularisation of migrants in irregular situations.

MIrreM analyses policies defining migrant irregularity, stakeholders' data needs and usage, and assesses existing estimates and statistical indicators on irregular migration in the countries under study and at the EU level. Using several coordinated pilots, the project develops new and innovative methods for measuring irregular migration and explores if and how these instruments can be applied in other socio-economic or institutional contexts. Based on a broad mapping of regularisation practices in the EU as well as detailed case studies, MIrreM will develop 'regularisation scenarios' to better understand conditions under which regularisation should be considered as a policy option. Together with expert groups that will be set up on irregular migration data and regularisation, respectively, the project will synthesise findings into a Handbook on data on irregular migration and a Handbook on pathways out of irregularity. The project's research covers 20 countries, including 12 EU countries and the United Kingdom.

TO CITE:

Rodríguez-Sánchez, A. & Tjaden, J. (2024). Estimating irregular migrant stock using social media data and machine learning. *MIrreM Briefing Paper*. Krems: University for Continuing Education Krems (Danube University Krems). DOI: 10.5281/zenodo.14808984

KEYWORDS

Social media; Machine Learning; Undocumented migration; Triangulation; Innovative Methods.

ACKNOWLEDGEMENTS

FUNDING ACKNOWLEDGEMENT

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

In addition, MIrreM benefit from funding provided by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee. The Canadian research component of this project is undertaken, in part, thanks to funding from the Canada Excellence Research Chairs Program of the Government of Canada.

1. INTRODUCTION

Previous research has shown that social media data on migrant populations – Facebook stocks in particular - strongly correlates with counts of migrants from official sources (Spyratos et al., 2018, 2019; Zagheni, Weber, & Gummadi, 2017). This research has not considered the presence of irregular migrants in social media data, nor how this relates to estimates of official migrant stocks. Given that irregular migrants may also be members of Facebook, the point of departure for our study is whether it could be possible to develop a method to estimate the total size of the irregular migrant population in a country based on social media data.

We build on this intuition and present an approach to estimate the size of the irregular migrant population for all countries with valid Facebook and UN data.

Given the demand for information on irregular migration on the one side, and the difficulty of arriving at reliable estimates, triangulation and validation of multiple data sources represents the best possible approach to narrow in on the most accurate estimate and to stimulate an evidence-based discussion. Triangulation helps identify the advantages and disadvantages of various pieces of information (Denzin, 2012), highlighting their disadvantages and merits. Our main contribution is to provide further data capturing irregular migration and enhance the viability of triangulation in this research space. We provide alternative estimates for many countries and propose a transparent, novel approach.

2. CONCEPTS & DEFINITIONS

Undocumented migrant, irregular migrant, unauthorized migrant, sans papiers, illegalized migrant, etc., are used interchangeably in the literature, although they do have different connotations. These terms – and related conceptualisations - have been heavily debated in the literature in terms of their adequacy to capture such a complex and dynamic phenomenon (Ambrosini & Hajer, 2023; Boudou, 2023). The most employed, yet inexhaustive, definition of irregular migrant considers an irregular migrant someone who does not have a valid visa, residence permit, or other documentation that allows them to stay and work in a country legally (International Organization for Migration, 2019)

The meaning of irregularity may vary by country of destination. For example, factors such as the specifics of the country of residence's migration legal system, the duration of stay of the migrant, main activity of the migrants, and previous legal statuses all interact with one another and play a crucial role in determining whether a migrant has fallen into irregularity (De Genova, 2002; Kraler & Reichel, 2011; Triandafyllidou & Bartolini, 2020).

Moreover, irregularity has a temporal dimension as well, as individuals fluidly move in and out of it over time, as shown in cases of migrants with precarious legal status trajectories (Goldring, 2022).

However, in practice, the definition of irregular migration is confined by the available data used to estimate this population. In our approach, we define the stock of irregular migrants to be the number of individuals who have previously lived in another country (hence, migrated) but who are not counted, or do not appear registered, in official statistics of migrant stocks in the destination country. This definition is coarser than other existing conceptualisations of migrant irregularity that encompass further dimensions, and which require more detailed and individual-level data (see Kraler & Ahrens 2023's taxonomy). The number of people "who have previously lived" in another country is user information provided by Facebook to advertisers on its platform. This classification used to be known as expats, in Facebook's marketing API. Facebook's definition does not match an official definition. For example, it does not account for the stay period. Although there is not a one-to-one correspondence, this information has been used to proxy migrant groups and is limited to the country of previous residence and country of current residence (Zagheni Weber, & Gummadi, 2017). However, it does not contain information on country of birth, nationality, year of immigration, mode of immigration or any other information related to legal status.

3. METHODS AND DATA

3.1 METHODS

Our approach starts with the idea that the Facebook stocks can be inflated to approximate the total migrant population in each country, that is, regular and irregular migrants. The method creates the hypothetical scenario in which all international migrants in the country would be on Facebook. If that were true, one could subtract the official count of regular migrants to get an estimate of the irregular migrant population. Here is this idea expressed as a formula:

$$Irregular\ stocks_{iz} = \frac{Facebook\ stocks_{iz}}{Penetration\ rate_{iz}} - Official\ Stocks_{iz}$$

Where the number of irregular migrants (from country i in destination z) is equal to number Facebook members from country i in destination country z, divided by the migrant group specific penetration rate, minus the "documented" or official number of migrants from country i in destination country z.

The migrant group specific penetration rate is the share of all migrant individuals in the total population of migrants that are present on Facebook for each country of origin. If the penetration rate were 1, this would mean that all migrants from a given country are members of Facebook and, therefore, that the Facebook stock of migrants would be a perfect measure of the total migrant stock including irregular migrants. If the penetration rate were 0.5, this would mean that the Facebook stock would need to be doubled to approximate the number of all individuals in the real population.

The problem is that the real penetration rate is unknown. The penetration is rate is defined as:

$$Penetration\ rate_{iz} = \frac{Facebook_{iz}}{Total\ Migrants_{iz}}$$

The whole migrant population (total migrants) includes irregular migrants and therefore the "true" penetration rate becomes inherently unknown as the number of irregular migrants is unknown.

We propose a novel method to approximate this penetration rate using a machine learning model. Employing a country-dyadic data set, we start by predicting the official "regular" count of migrants measured by the UN stock data based on a large range of predictors for total migration in a country (see Table 2 in Annex). We use a relative measure of the error in

the model, i.e. the ratio between UN stocks and the predictions from a machine learning model (i.e., extreme gradient boosting, XGBOOST, see Annex for further details of this model), as a statistical signal for the size of the undocumented population in a country as this quantity varies systematically between country-dyads. We can show empirically and mathematically that the error of the prediction model is systematically related to the size of irregular migrant stocks, and we use this signal to compute the migrant specific Facebook penetration rates. The technical details of this approach go beyond the scope of this Briefing Paper. For further details, please see Rodriguez-Sanchez, Tjaden & Weber (2025).

The final step of our approach consists of aggregating all the country-to-country estimates by destination country to arrive at our final estimate for the total undocumented stock by country of destination. In this last step, we consider which country-to-country corridors likely produce irregular migration and which do not (e.g. free movement zones between high-income countries such as within the EU).

The proposed method is a pilot test. While we find that it produces plausible estimates, further refinement and validation is needed. For more information on other innovative methods to estimate irregular migration, please see Rodriguez-Sanchez & Tjaden 2024.

3.2 DATA

Our approach employs two main data sources: 1) the International bilateral migration stock from the UN; and 2) the number of Facebook users classified as having 'lived abroad'.

The UN stocks are provided at 5-year intervals and is the most comprehensive data available on international migration. Various types of data go into the estimation of the UN bilateral stocks, depending on each country's national statistical system. Census data is often used to obtain these estimates, but in cases where there is no recent census, other sources are used (United Nations, 2020). In some countries, population wide register systems are used, whereas in others, surveys and projections have been employed to obtain bilateral stock estimates. Other forms of mobility, such as tourism, are excluded from these bilateral stocks. International migrants in the UN bilateral stocks are the foreign-born, though in some cases citizenship is considered in the definition (United Nations, 2020). The main drawback of the UN bilateral stocks is that they combine multiple kinds of data and is of limited quality (Willekens et al., 2016). Data quality of the UN bilateral stocks differs greatly from country to country, but the data is by far the largest international comparison on migration stocks available to date.

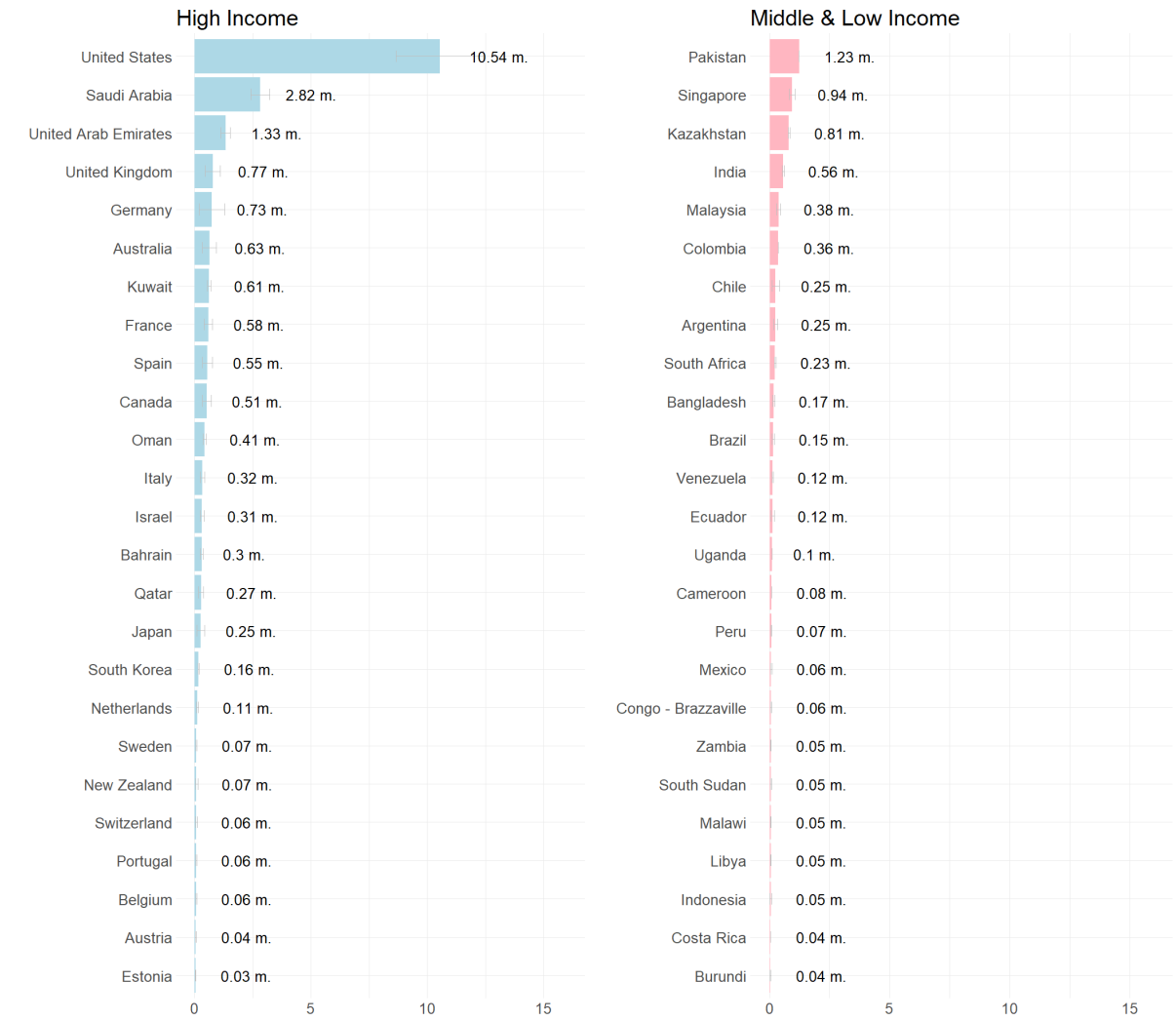
The Facebook stock, in turn, is the number of Facebook users who have previously lived in another country, formerly known as 'expats.' The algorithms used by Facebook to assign users to this category are not publicly available. However, the consensus in the literature is that the platform uses a combination of the history of the IP log-in location, the network of friends, the language used in the interfaces, and self-reported information in Facebook profiles (Spyratos et al., 2018). This data was passively collected employing Facebook's API for the whole of the year 2020 (Zagheni et al., 2017), at least twice per month. We averaged all the 2020 total values of expats for each country of origin available to obtain a single value

for the year. This data is only available for a selection of countries - where Facebook is available (i.e., 231 countries), and only distinguishes some groups of migrants from a reduced list of likely countries of origin (i.e., 90 countries). Some countries of residence potentially hosting important populations of undocumented migrants are not captured by this data source (e.g., Russia), and important countries of origin are also not reported (e.g., Syrians). Moreover, the Facebook stocks cannot be collected retrospectively and therefore access to the data relies on previous projects that carry systematic collection (Zagheni et al., 2017). Comparability of Facebook stocks over time is a major drawback of this data. For example, changes in Facebook's algorithms, lack of clarity of the algorithms used to classify users as having lived abroad, presence of fake accounts, and the lack of estimates for various countries of origin (less than a hundred are included as countries of origin) can all affect estimates of Facebook stocks.

4. RESULTS

Figure 1 shows the estimates we obtain for a selection of High-Income countries and Middle- & Low-Income countries. According to our results, among High-Income countries, the US is by far the country with the largest population of undocumented migrants. The US is followed by Saudia Arabia, United Arab Emirates, two of the most important countries in the Cooperation Council for the Arab States of the Gulf. Among the top ten high-income countries, there are only four European countries, namely the UK, Germany, France and Spain, all with estimates below 1 million, broadly in agreement with other estimates (Connor & Passel, 2019).

The recent overview of estimates of undocumented migration in twelve European countries (Kierans et al., 2024) - Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Poland, Spain, and the UK - puts the stock between 2.6 million to 3.3 million estimated undocumented migrants. According to our approach, these same countries host a population of 3.24 million undocumented migrants.



Note: Own elaboration. Selection of countries with the highest undocumented migrant stocks among selection of countries.

Figure 1 Estimates of undocumented migrants in a selection of countries in millions for the year 2020.

Among the countries in the Middle- and Low-Income group, the stocks of undocumented migrants are on average substantially smaller. The top five countries being in Asia - Pakistan, Singapore, Kazakhstan, India, and Malaysia, followed by important populations of undocumented migrants in Colombia, Chile, and Argentina.

Table 1 Spearman's rank correlation coefficient between estimates and other benchmarks.

Source	*Correlation*	*P.value*
Kierans et al. (2024)	0.82	6.81e-03
PEW (2019)	0.86	1.42e-03
Reuveny (2016)	0.80	1.20e-07
Other studies	0.85	3.44e-04

*Note: Own computations. Other studies are marked with * in reference list.*

When comparing these estimates with other compilations, as shown in Table 1, we also find that our results, in terms of the ranking of countries, largely agree with previous compilations or recently published single-country studies.

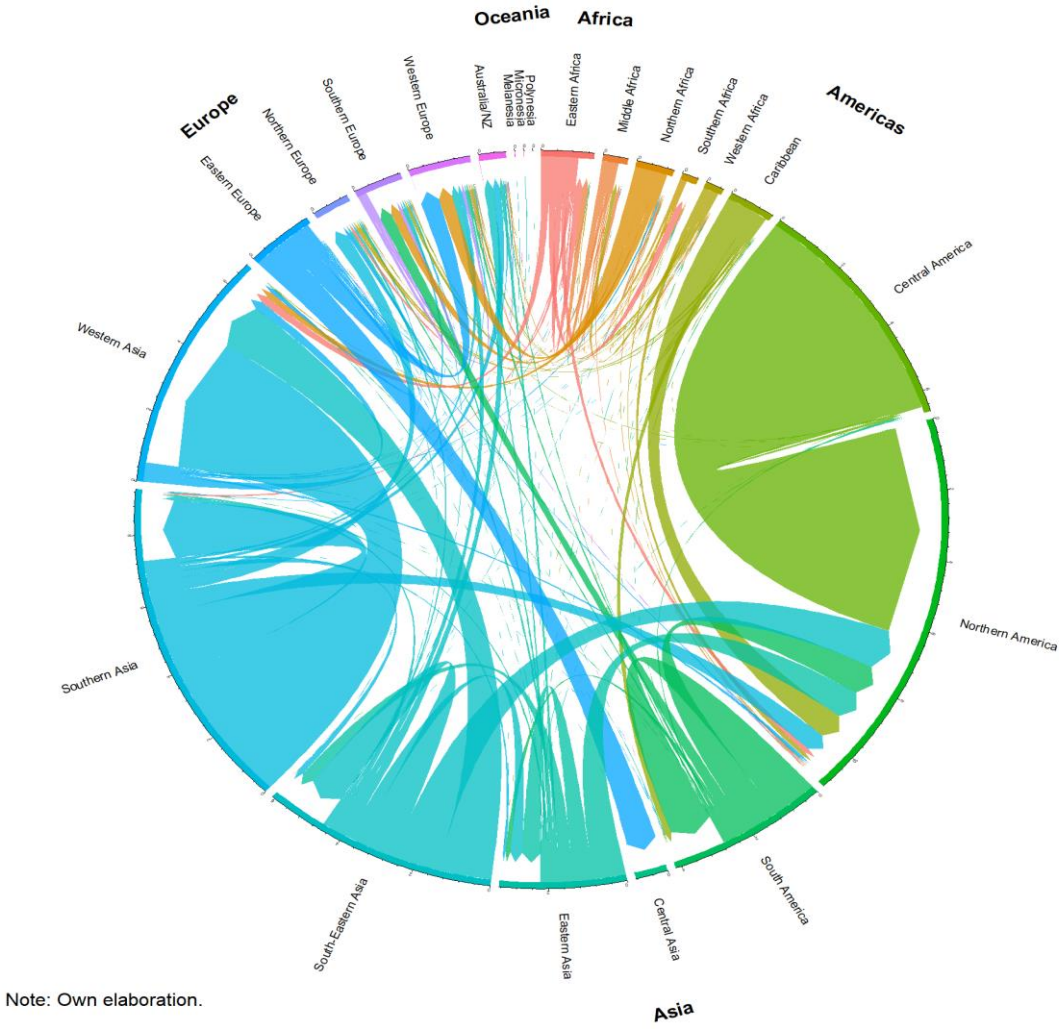


Figure 2. The composition of the undocumented migrant populations in a circular plot by world regions and continents in the world.

The composition of the undocumented migrant population by region of origin is another important result of our approach and helps us understand the composition of these migration flows. These are shown in Figure 2. Analysing the arrows going towards a specific region, for example towards North America, we see that Central Americans, Caribbeans, South-East Asians, South Americans and Central Asians are the five largest groups – in agreement with previous research (Baker & Warren, 2024). For Western Europe, we see that Eastern Europe, Northern Africa, Southern Asia, Western Asia and Eastern Asians appear as the top five. Moreover, other less studied world regions, such as Southern Asian, South America, and Eastern Asia, have large self-loops, and even one of the largest corridors would be from

Southern Asian towards Western Asia, as highlighted in migration towards the Gulf countries. These results hint at important populations of irregular migrants in South-to-South flows (e.g., the Venezuelan diaspora) that require further study.

5. DISCUSSION

We have presented a new methodology to measure the stock of irregular migrants employing a combination of traditional and innovative data sources (i.e., UN and Facebook's stock of migrants), and an ML model. Building on the idea that Facebook members include irregular migrants who are not captured in official statistics, we develop a method to estimate a penetration rate which expands the Facebook migrant stock to the full population.

ADVANTAGE

The estimates of the stock of irregular migrants resulting from this method are comparable to other estimates in many countries using traditional approaches. To the best of our knowledge, this method is the first to provide estimates of the irregular migrant population across as many countries, employing a single method. Hence, our results have the potential of generating new perspectives on the phenomena of irregular migration. However, careful attention to the data landscape in each country is needed, to put our estimates in relation to existing estimates and to assess the data quality in each case (e.g., UN DESA estimates vary strongly in quality)

Another advantage of the approach relates to the traceability of its assumptions and data sources. Our approach can be replicated with other data sources capturing migrant stocks (e.g., as shown in another study Kim et al., 2020). Our method could be made more reliable if the alternative data source used to measure the stock of migrants had a similar definition to that of the official stocks. Using Facebook, one could, for example, refine the approach by obtaining data on the population categorized as "having lived abroad" and residing in the country for at least 12 months. Currently, such fine grained categorization is not available through Facebook, but it could be possible in the future using Facebook, or by employing other innovative data sources, such as mobile phone data or Whatsapp.

RELIABILITY

The reliability of this approach is constrained by the main underlying data sources, namely the measurement of international migrant stocks by the UN and Facebook data. Both are subject to changes over time. Individual countries may change how the estimation of the migrant stock in their country is done. Facebook may change its working definition of users who lived abroad. At the same time, Facebook membership may change over time and differently across countries. These changes undermine the reliability of our method over time. Notwithstanding, it is possible that this approach can be replicated in the future, once more data becomes available, allowing for a direct comparison of different estimates over time or even across different platform with large, global user bases.

SCALABILITY

In terms of scalability of our approach, we have shown that it can provide estimates for many countries and is only limited by the availability of Facebook stocks. In fact, the estimates exploit variation across country-to-country information on migrant stock and Facebook membership. The approach does not work for a small number of countries. As such, scalability is not only possible, but also a key requirement of the method.

ESTIMATION ASSUMPTIONS

The main assumptions in our approach relate to data interpretation and the ML model. We are assuming that the UN bilateral stocks and Facebook stocks are a fraction of the total migrant stocks (including regular and irregular migrants). We also assume that irregular migrants use Facebook and that UN international migrant stocks are a good measure of the regular migrant population in a country.

ETHICAL CONSIDERATIONS

Finally, research in the "irregular migration" space is sensitive given its political salience, especially when uncertainty in the estimates is high. Any estimate of irregular migration has a dual use problem. Some actors may take estimates out of context and lobby for border control and deportations (Haas, 2024). Our estimates could be mistaken as target numbers of irregular migrants that ought to be identified and deported, which is an interpretation we refrain from. Others, in turn, may take estimates to lobby for regularization and access to social services (Abarca & Coutin, 2018; Obertino-Norwood & García, 2023) employing numbers to establish a form of legal deservingness in the eyes of the state (i.e., "bureaucratic visibility").

We acknowledge that attempts at counting the number of irregular migrants goes against the interest of at least some of them to remain hidden and uncounted. Our approach does not carry the risk of leading to the identification of a single individual, given that we only use country-level aggregate data. A separate ethical issue in relation to the use of social media data relates to obtaining informed consent from Facebook users for using data for the purpose of estimating irregular migration. Users consent to provide the information and allow third parties to access this information in the aggregate. However, at the time of registration, users are likely unaware of the implications of potential uses of their data. In other contexts, research using Facebook data has previously led to changes in the user data which Facebook makes available for research given ethical considerations. A related problem is that Facebook may decide to change is classification system or take away access to this information at any time, hampering attempts for further research and the use of our method. As such, we hope our empirical contribution also stimulate discussions among civil society organizations, law makers, and digital platforms on which information should be accessible to government and researchers and which items are deemed too sensitive.

REFERENCES

Abarca, G. A., & Coutin, S. B. (2018). Sovereign intimacies: The lives of documents within US state-noncitizen relationships. *American Ethnologist*, 45(1), 7–19. https://doi.org/10.1111/amet.12595

Ambrosini, M., & Hajer, M. H. (2023). Irregular migration: IMISCOE short reader. Springer Research Series. https://library.oapen.org/handle/20.500.12657/63574

Baker, B., & Warren, R. (2024). Estimates of the unauthorized immigrant population residing in the United States: January 2018–January 2022. Office of Immigration Statistics, U.S. Department of Homeland Security. Retrieved from Office of Immigration Statistics, U.S. Department of Homeland Security website: https://ohss.dhs.gov/sites/default/files/2024-06/2024_0418_ohss_estimates-of-the-unauthorized-immigrant-population-residing-in-the-united-states-january-2018%25E2%2580%2593january-2022.pdf

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259–280. https://doi.org/*10.1257/jep.32.3.259*

Bonneau, L. (2023). Migration status: A key structural social determinant of health inequalities for undocumented migrants, Working Paper. Brussels: PICUM. https://picum.org/wp-content/uploads/2023/12/Migration-status_A-key-structural-social-determinant-of-health-inequalities-for-undocumented-migrants_EN.pdf

Boudou, B. (2023). Migration and the critique of "state thought": Abdelmalek Sayad as a political theorist. *European Journal of Political Theory*, 22(3), 399–424. https://doi.org/10.1177/14748851211041906

Capps, R., Gelatt, J., Soto, A. G. R., & Van Hook, J. (2020). Unauthorized immigrants in the United States: Stable numbers, changing origins. *Migration Policy Institute*. https://coilink.org/20.500.12592/t4b8nps

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Connor, P., & Passel, J. S. (2019). Europe's unauthorized immigrant population peaks in 2016, then levels off. Retrieved from https://www.pewresearch.org/global/2019/11/13/europes-unauthorized-immigrant-population-peaks-in-2016-then-levels-off/

Conte, M., P. Cotterlaz and T. Mayer (2022), "The CEPII Gravity database". CEPII Working Paper N°2022-05, July 2022. https://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=8

Czaika, M., Haas, H. de, & Villares-Varela, M. (2017). The global evolution of travel visa regimes: An analysis based on the DEMIG VISA database. Oxford University Research Archive. https://ora.ox.ac.uk/objects/uuid:3be426d7-7cce-4d74-af46-259726b16e42

De Genova, N. P. (2002). Migrant "illegality" and deportability in everyday life. *Annual Review of Anthropology*, 31(1), 419–447. https://doi.org/10.1146/annurev.anthro.31.040402.085432

Denzin, N. K. (2012). Triangulation 2.0. Journal of Mixed Methods Research, 6(2), 80–88. https://doi.org/10.1177/1558689812437186

Deutschmann, E., Gabrielli, L., & Recchi, E. (2023). Roads, rails, and checkpoints: Assessing the permeability of nation-state borders worldwide. *World Development*, 164, 106175. https://doi.org/10.1016/j.worlddev.2022.106175

Haas, H. de. (2024). Changing the migration narrative: On the power of discourse, propaganda and truth distortion (Working Paper No. 3). PACES Project. https://www.austriaca.at/0xc1aa5572_0x003fc5c0#page=104

International Organization for Migration. (2019). Glossary on migration. International Organization for Migration. https://publications.iom.int/system/files/pdf/iml_34_glossary.pdf

Gálvez Iniesta, I. (2020). The size, socio-economic composition and fiscal implications of the irregular immigration in Spain. https://e-archivo.uc3m.es/rest/api/core/bitstreams/1f19ed21-ac46-4007-9637-bd60e73bbc19/content

Goldring, L. (2022). Precarious legal status trajectories as method, and the work of legal status. *Citizenship Studies*, 26(4-5), 460–470. https://doi.org/10.1080/13621025.2022.2091228

Jandl, M. (2011). Methods, approaches and data sources for estimating stocks of irregular migrants. *International Migration*, 49(5), 53–77. https://doi.org/10.1111/j.1468-2435.2011.00701.x

Jolly, A., Thomas, S., & Stanyer, J. (2020). London's children and young people who are not british citizens: A profile. *Greater London Authority*. https://trustforlondon.fra1.cdn.digitaloceanspaces.com/media/documents/Londons_Children_and_Young_People_Who_Are_Not_British_Citizens_report.pdf

Kierans, D., Vargas-Silva, C., Ahmad-Yar, A. W., Bircan, T., Cacciapaglia, M., Carvalho, J., ... Sohst, R. R. (2024). *MIrreM public database on irregular migration stock estimates (version 2).* Krems: University for Continuing Education Krems (Danube University Krems). https://zenodo.org/records/13856861

Kim, J., Sîrbu, A., Giannotti, F., & Gabrielli, L. (2020, April). Digital footprints of international migration on twitter. In *International Symposium on Intelligent Data Analysis* (pp. 274-286). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-44584-3_22

KNOMAD/World Bank. (2021). Bilateral Remittance Matrix 2021. Retrieved from https://www.knomad.org/data/remittances

Koser, K. (2010). Dimensions and dynamics of irregular migration. *Population, Space and Place*, 16(3), 181–193. https://doi.org/10.1002/psp.587

Kraler, Albert, & Reichel, D. (2011). Measuring irregular migration and population flows–what available data can tell. *International Migration*, 49(5), 97–128. https://doi.org/10.1111/j.1468-2435.2011.00699.x

Kraler, A., & Ahrens, J. (2023). Conceptualising migrant irregularity for measurement purposes (Version 3), *MIrreM Working Paper No. 2*. Krems: University for Continuing Education Krems (Danube University Krems). https://zenodo.org/records/7868237

LeVoy, M., & Geddie, E. (2009). Irregular migration: Challenges, limits and remedies. *Refugee Survey Quarterly,* 28(4), 87–113. https://doi.org/10.1093/rsq/hdq010

Obertino-Norwood, H., & García, A. S. (2023). Hard to count? The 2020 census "citizenship question" and bureaucratic visibility among undocumented Latin Americans in Chicago. *Social Service Review*, 97(3), 540–568. https://www.journals.uchicago.edu/doi/abs/10.1086/725212

O'Hare, W. P. (2019b). *Differential Undercounts in the US Census: Who is Missed?* Springer Nature. http://library.oapen.org/handle/20.500.12657/23087

Recchi, E., Deutschmann, E., Gabrielli, L., & Kholmatova, N. (2021). The global visa cost divide: How and why the price for travel permits varies worldwide. *Political Geography*, 86, 102350. https://doi.org/10.1016/j.polgeo.2021.102350

Recchi, E., Deutschmann, E., & Vespe, M. (2019). Estimating transnational human mobility on a global scale. EUI RSCAS Working Papers – MPC Series, (2019/30). http://dx.doi.org/10.2139/ssrn.3384000

Recchi, E., Deutschmann, E., & Vespe, M. (2020). Global transnational mobility dataset. https://cadmus.eui.eu/handle/1814/67634

Reuveny, R (2016). Illegal immigrants in high income countries and politically autonomous units: recent estimated stocks by country and unit. *Journal of Globalization Studies*, 7(1), 30-46. https://www.sociostudies.org/journal/files/jogs/2016_1/030-046.pdf

Rodriguez-Sanchez, A., & Tjaden, J. (2023). Estimating irregular migration – a review of traditional and innovative methods (version 2). *MIrreM Working Paper No. 4*. https://doi.org/10.5281/zenodo.8380854

Rodriguez-Sanchez, A., Weber, I. & Tjaden, J. (2023). Leveraging social media and machine learning to estimate uncounted migrant stocks (forthcoming).

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). *Migration Data Using Social Media: A European Perspective*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/964282

Spyratos, S., Vespe, M., Natale, F., Weber, Z., Zagheni, E., & Rango, M. (2019). Quantifying international human mobility patterns using Facebook network data. PloS One, 14(10), e0224134. https://doi.org/10.1371/journal.pone.0224134

Triandafyllidou, A., & Bartolini, L. (2020). Understanding irregularity. In S. Spencer & A. Triandafyllidou (Eds.), *Migrants with irregular status in Europe. Evolving conceptual and policy challenges* (pp. 11–31). Springer. https://link.springer.com/chapter/10.1007/978-3-030-34324-8_2

UNESCO. (2024). Global flow of tertiary-level students. UNESCO Institute of Statistics. Retrieved from https://uis.unesco.org/en/uis-student-flow

United Nations. (2020). International migrant stock 2020: documentation. United Nations; Retrieved from https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2021/Jan/undesa_pd_2020_international_migrant_stock_documentation.pdf

UNHCR. (2024). Measuring forced displacement and statelessness. UNHCR Refugee Population Statistics Database. Retrieved from https://www.unhcr.org/refugee-statistics/methodology/

Van Hook, J., Morse, A., Capps, R., & Gelatt, J. (2021). Uncertainty about the size of the unauthorized foreign-born population in the United States. Demography, 58(6), 2315–2336. https://doi.org/10.1215/00703370-9491801

Warren, R. (2021). In 2019, the US undocumented population continued a decade-long decline and the foreign-born population neared zero growth. *Journal on Migration and Human Security*, 9(1), 31–43. https://doi.org/10.1177/2331502421993746

Willekens, F., Massey, D., Raymer, J., & Beauchemin, C. (2016). International migration under the microscope. *Science*, 352(6288), 897–899. https://www.science.org/doi/10.1126/science.aaf6545

World Bank. (2024). World development indicators. Retrieved from https://datatopics.worldbank.org/world-development-indicators/

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. Population and Development Review, 721–734. ttps://doi.org/10.1111/padr.12102

ANNEX 1

The MIrreM Methods Lab conducted a review of 21 traditional and innovative methodological approaches for estimating irregular migrant stocks and flows. Each approach was assessed based on its core concept, data sources, definition and coverage of irregular migration, estimation assumptions, reliability, scalability, general assumptions, and ethical considerations.

Building on this review, we developed six innovative approaches that have the potential to advance research on irregular migration.

As part of the broader MIrreM project, the WP6 Methods Innovation Lab carried out the following six Pilot Studies (PS). Please find the MIrreM Briefing Papers about the other Pilot Studies linked below:

MIrreM Briefing Papers	Authors	DOI
PS1 - Exploring the use of aggregate air passenger data for estimating overstayer inflows	Luca Bernasconi Ettore Recchi	https://doi.org/10.5281/zenodo.14809013
PS2 - Measuring the participation of irregular migrants in the informal economy	Aslı Salihoğlu Carlos Vargas-Silva	https://doi.org/10.5281/zenodo.14809000
PS3 - Estimating irregular migrant stocks using social media data and machine learning	Alejandra Rodríguez-Sánchez Jasper Tjaden	https://doi.org/10.5281/zenodo.14808984
PS4 - Irregular migration: What can mortality reveal?	Johan Surkyn Tuba Bircan	https://doi.org/10.5281/zenodo.14808979
PS5 - Estimating irregular migration in the UK using a health care reform	Alejandra Rodríguez-Sánchez Jasper Tjaden	https://doi.org/10.5281/zenodo.14808948
PS6 - Measuring irregular migration stocks through social media surveys	Jasper Tjaden Alejandra Rodríguez-Sánchez	https://doi.org/10.5281/zenodo.14801999

ANNEX 2 – List of covariates used in the estimation of the migrant specific penetration rate

Table 3 provides a brief description of the various data sources, the type (in terms of variables being dyadic and or symmetric), the years for which the values were taken, as well as the number of countries and country-dyads in the case of dyadic data. For countries in which information from more than one year was available, we used the most recent estimates.

Table 3 Explanation of Data Sources Utilized for Predictive Features of Real Migrant Stocks

Type	Variables	Description of the data	Data type	Years	Number of countries	Number of Country-dyads
Target	Migration stocks (United Nations)	The international migrant stock by destination and origin for the mid-point (1 July) of each year.	dyadic; not symmetrical	2020	231	11951
Predictive Features	Population size	United Nations Department of Economic and Social Affairs estimates of the population by gender and age groups.	non-dyadic	2020	235	-
	Remittances	Bilateral Remittance Estimates for 2021 using Migrant Stocks, Host Country Incomes, and Origin Country Incomes (millions of US$).	dyadic; not symmetrical	2021	211	10204
	Facebook users	Audience estimates by country or residence or location	non-dyadic	2023	184	-
	Facebook users classified as Lived abroad	Audience estimates of expats by country of origin and destination	non-dyadic	2021	90	-
	Facebook Social Connectedness Index	The Social Connectedness Index (SCI) is a measure of how well or strongly connected two different regions are. The value of the SCI measures the relative probability that two individuals from two locations are Facebook friends with each other.	dyadic; symmetrical	2023	184	33672
	Internet access rates	Individuals using the Internet as a percentage of the total population based on data from World Bank, the Iternational Telecommunication Union (ITU), and the World Telecommunication/ICT Indicators Database.	non-dyadic	2020	215	-
	UNHCR Refugees and asylum seekers	This it the total number of refugees, asylum seekers and internally displaced persons residing in each country.	dyadic; not symmetrical	2020	171	5291
	International students	UNESCO's Global Flow of Tertiary-Level Students. This is the nuber of inbound internationally mobile students by country of origin.	dyadic; not symmetrical	2015 - 2020	180	33686
	Airport flow mobility	Sabre data on cross-border air passenger mobility (country to country). Sabre is a private company that gathers its data from various airline companies.	dyadic; not symmetrical	2019	237	38209
	Global Transnational Mobility Dataset	Estimates of country-to-country cross-border human mobility ("trips") on the basis of global statistics on tourism and air passenger traffic. See Rechi et al. 2019 for details.	dyadic; not symmetrical	2012 - 2016	196	38416
	Permeability of nation-state borders	The index of border permeability for land boders based on data from OpenStreetMap and the World Food Programme, detecting cross-border transport infrastructure and checkpoints. See for details.	dyadic; symmetrical	2022	118	308
	Costs of Visas	The Global Visa Cost Dataset was built through manual and an automatized web-scraping-based data collection on a	dyadic; not symmetrical	2019	198	38574

	country-to-country basis. The authors compiled visa costs for tourism-, study-, business-, work-, family reunification-, transit-, and other motives-related visas. All costs are provided in US dollars (USD). See for details.				
DEMIG VISA database	The DEMIG VISA database tracks the different VISA requirements for nationalities and countries of destination. The categories are: "0" = Visa/Exit permit NOT needed; "1" = Visa/Exit permit needed; and "2" = Individuals are not allowed to travel to this country ("blacklisted"). This information was originally obtained from the IATA Travel Information Manual. See for details.	dyadic; not symmetrical	1973 - 2013	227	45320
Socioeconomic indicators (World Bank)	A collection of multiple socioeconomic indicators on health, employment, poverty, inequality, agriculture, education, economy, etc. from the World Bank	non-dyadic	2020	266	-
CEPII's Cultural and historical proximity	Cultural indicators include common official and other languages, common colonizer, common religion and a previous colonial dependency, or if the ever were colonized by a similar political unity. See for details.	dyadic; symmetrical	2021	235	55225
CEPII's Geographical distance and contiguity	We include here various bilateral distance metrics between countries and an indicator whether these countries share a land border. See for details.	dyadic; symmetrical	2021	235	55225
CEPII's Trade flows	Trade flows data from three sources: the CEPII's BACI database, the UNSD's Comtrade data and the IMF's DOTS data. See for details.	dyadic; not symmetrical	1962 - 2020	235	55225
CEPII's Trade facilitation or international agreements	Binary indicators of whether country is a member of the GATT, the WTO, the EU. See for details.	dyadic; not symmetrical	1990 - 2020	235	55225
CEPII's Macroeconomic indicators	We employ GDP per capita in 2011 PPP US dollars. See for details.	dyadic; not symmetrical	1990 - 2017	235	55225

Note: own compilation of variables.

ANNEX 3 – Further methodological details

The Machine Learning model we chose for this exercise is an extreme gradient boosting model: Xgboost (Chen & Guestrin, 2016), a highly flexible, nonparametric model that can capture complex relationships between multiple predictors and the target variable. We estimate this model employing 10-fold cross-validation to tune the most important hyperparameters of this model on the train data (i.e., number of trees, tree depth, minimum number of observations in final leaf, and stopping iterations), and fixed the learning rate at a value of 0.01 and evaluate on the test data. Predictions are then obtained for the whole sample. Other ML models were tested, but this one provided the smallest error.

ABOUT THE AUTHORS

Alejandra Rodríguez-Sánchez is currently a postdoc at the University of Potsdam, Germany. Jasper Tjaden is Professor of Applied Social Research and Public Policy at the Economic and Social Science Department of the University of Potsdam, Germany.

COPYRIGHT CLAUSE

THE MIRREM CONSORTIUM

University for Continuing Education Krems (Coordinator)

European University Institute

University of Osnabrück

University of Maastricht

University of Turku

Complutense University Madrid

Hellenic Foundation for European and Foreign Policy (ELIAMEP)

University of Milan

University of Potsdam

Platform for International Cooperation on Undocumented Migration (PICUM)

International Centre for Migration Policy Development (ICMPD)

Migration Policy Institute Europe (MPI-E)

University of Warsaw

Vrije Universiteit Brussel

Instituto Universitário de Lisboa

Associated Partners:

Toronto Metropolitan University

University of Leicester

University of Oxford