Use of non-traditional data sources to nowcast migration trends through Artificial Intelligence technologies.

Diletta Goglia*[1], Laura Pollacci[1], Alina Sîrbu[1]

[1]Department of Computer Science, University of Pisa, Pisa, Italy
* Corresponding author: d.goglia@studenti.unipi.it

Keywords: International migration · migration flows · bilateral migration · Facebook social connectedness index · migration nowcasting · migration prediction · migration drivers

Extended Abstract

Background and motivation

In recent years the pursuit of original drivers and methods is becoming an increasing requirement for migration studies, considering the new technologies used to characterise and understand the human migration phenomenon. In addition to the traditional data typically used in migration studies (e.g., indicators related to the labour market or economic status, measures obtained from surveys and official statistics, either from national censuses or from the population registries), many researchers have proposed to employ non-traditional data sources to study migration trends (Bosco et al. (2022); Fiorio et al. (2017); Gendronneau et al. (2019); Kim, Sîrbu, Rossetti, Giannotti, and Rapoport (2021); Salah (2021); Spyratos et al. (2018); Sîrbu et al. (2021); Zagheni, Garimella, Weber, and State (2014); Zagheni, Polimis, Alexander, Weber, and Billari (2018); Zagheni, Weber, and Gummadi (n.d.)). These can consist in news data, satellite data, but also in digital traces of humans generated by using internet services, mobile phones, IoT devices, fidelity cards, online social networks and many others. This unconventional approach is intended to find an alternative methodology to answer open questions about the human migration framework (i.e., nowcasting flows and stocks, studying the integration of multiple sources and knowledge, and investigating migration drivers). The new data have the advantage of timeliness and large geographical coverage, but also disadvantages in terms of selection bias and amount of resources required to process (Pollacci, Milli, Bircan, and Rossetti (2022); Sîrbu et al. (2021)). Therefore, models extracted from these data need to be carefully validated, typically with traditional data sources. In this context of meaningful data combination, many types of data exist, still very scattered and heterogeneous, making integration far from straightforward.

Contributions

Our work focuses on the integrated use of heterogeneous traditional datasets and new data types. We present two different contributions: MIMI, a new multi-feature dataset (Goglia (2022); Goglia, Pollacci, and Sirbu (2022)), and a new regression analysis that could significantly contribute to the study of migration drivers and, in future work, to forecast emerging trends through the use of Artificial Intelligence technologies.

The MIMI dataset

The Multi-aspect Integrated Migration Indicators (MIMI) dataset is intended to be exploited in migration studies, and is a concrete example of integration of traditional and non-traditional data sources. It includes official data about bidirectional human migration (traditional country-to-country flow and stock data, retrieved from EUROSTAT and United Nations public datasets), multidisciplinary variables and original indicators, including economic, demographic, cultural and geographic indicators, together with the Facebook Social Connectedness Index (SCI)[1] Bailey, Cao, Kuchler, Stroebel, and Wong (2018); Meta (2021). The dataset was released under the Creative Commons Attribution 4.0 International

[1] bit.ly/Facebook_SCI

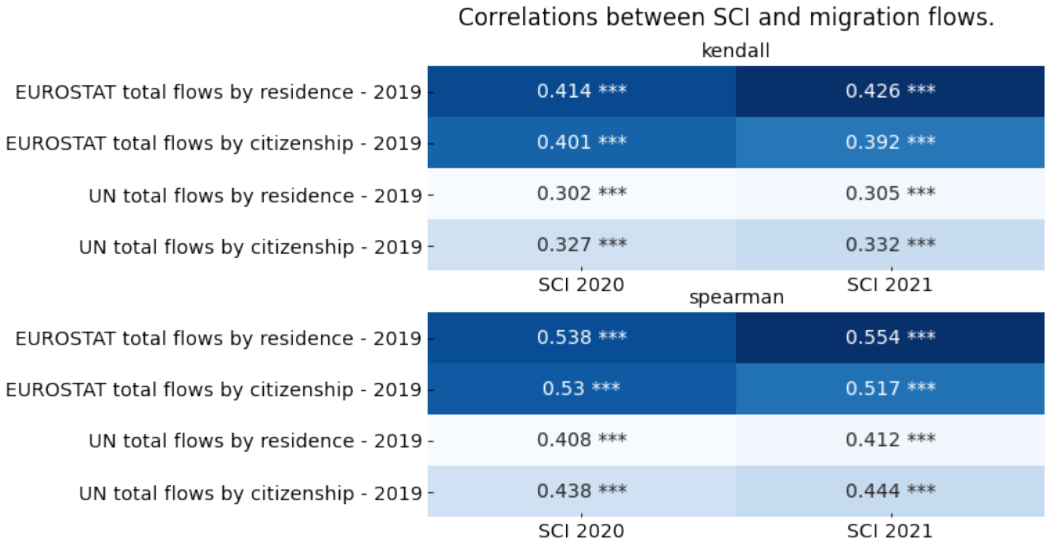


Figure 1: Kendall's tau-b and Spearman rank-order correlations between SCI and migration flows by citizenship and by residence for both UN and EUROSTAT sources. Measures of Facebook SCI are available in the MIMI dataset for both 2020 and 2021. They have been correlated with the most recent data available for migration flows, i.e. 2019.

Public License (CC BY 4.0[2]) and is publicly available on Zenodo [3]. It contains more than 28,000 records and 870 different variables and covers 255 different countries, identified by ISO-3166 standard notation.

The integration process uniformised the data coming from various sources, in an attempt to fill in gaps and missing data, standardise location and time dimensions, and ultimately facilitate use by the research community. Thanks to this variety of knowledge, experts from several research fields (demographers, sociologists, economists) could exploit MIMI to investigate the trends in the various indicators, and the relationship among them. Moreover, it could be possible to develop complex models based on this dataset to assess human migration by evaluating related interdisciplinary drivers, and to nowcast/predict traditional migration indicators through non-traditional variables, such as the strength of social connectivity. Here, the Facebook SCI could have an important role. It guarantees an anonymised collection of information on users and their friendships, measuring the relative probability that two individuals across two countries are friends on Facebook. Therefore it could be employed as a proxy of social connections across borders to be studied as a possible driver of migration.

As example of salient patterns observed among various indicators included in the MIMI dataset, Figure 1 shows correlations between migration flows and Facebook SCI using the Spearman and Kendall coefficients.[4]. We observe significant positive correlation between migration flows and SCI, indicating that social connectedness may be an important migration driver, and that SCI could be employed to estimate migration flows.

Migration drivers

Our second contribution is an analysis of the relation between indicators included in the MIMI dataset, through regression analysis. We present a new measure, the Bidirectional Migration Index (BMI)

[2] https://creativecommons.org/licenses/by-nc/4.0/

[3] 10.5281/zenodo.6493325

[4] Spearman rank-order formulation represents a non linear monotonicity between variables, while Kendall's tau-b formulation is used to check ranking correspondence. Spearman's correlations varies in the range [-1, +1], with 0 implying no correlation, while in Kendall's correlations values close to the positive or to the negative bound indicate, respectively, strong agreement or disagreement of rankings. P-values have been computed in order to confirm of refute the relevance of each correlation value: results are indicated in heatmaps with a number of asterisks proportional to the relevance obtained:

no asterisks	no relevance	p-value ≥ 0.05
*	moderate relevance	$0.01 \leq$ p-value < 0.05
**	high relevance	$0.001 \leq$ p-value < 0.01
***	very high relevance	p-value < 0.001

indicator, which takes into account both the inflows and outflows shared by two countries *i* and *j*, and which is defined for each year *t* as follows:

$$BMI(t) = \frac{Flow_{i \to j}(t) + Flow_{j \to i}(t)}{Pop_i(t) * Pop_j(t)} \tag{1}$$

We predict the values of the BMI starting from SCI and other indicators through an ordinary least squares statistical model (OLS) that performs a linear regression to estimate migration trends in order to understand which variables, among those included in the MIMI dataset, are related to the flows and therefore could be considered migration drivers. At this stage, our goal is to understand which factors are significant for the purpose of predicting migration. Specifically, the OLS model fits a subset of variables derived from the MIMI dataset and related to each country pair *i* and *j*, and evaluates their relevance for estimating the BMI. Besides Facebook SCI, we include:

- as numerical variables: the distance between the two countries, difference and mean GDP[5], area, number and percentage of Facebook users of the two countries, difference between each cultural indicator of the two countries (Hofstede, 1980; Kaasa, Vadi, & Varblane, 2016; Kaasa, Anneli, 2014).
- as binarized categorical variables: indicators that express whether the two countries share border, religion, language or continent.

The initial structure for the linear regression analysis consists in four different settings, analysing separately migration data by residence and by citizenship and then considering or not the Facebook SCI among the list of independent variables. The BMI indicator, representing the dependent variable, is then predicted for all the four settings following the backward elimination approach, which consists in the exclusion, at each step, of the variable considered to be the least relevant. At the end of all the necessary iterations, the final result for each one setting is a model with a reduced list of variables (i.e., those considered to be the most significant).

The results are summarised in Figure 2, where the initial and the final models are reported for each of the four configurations, including the p-values for each variable as a measure of its relevance. Beside noticing a strong improvement in the R^2 measure for those settings that take into account the Facebook SCI, this feature itself proves to be always strongly and positively significant.

Further detail of the evidence obtained during this phase can be observed in Figure 3, where the result of the best OLS model for migration data by residence is reported, corresponding to the linear regression described in the last column of Figure 2. The plot shows a comparison between the true and the estimated values of the BMI indicator related to 2019, also including information about SCI. Facebook strength of connectivity between two countries is strongly and positively related to the amount of migration flows they share. Moreover, the higher the connectivity the more accurate the migration prediction, as suggested by the match between colour scale and linear growth.

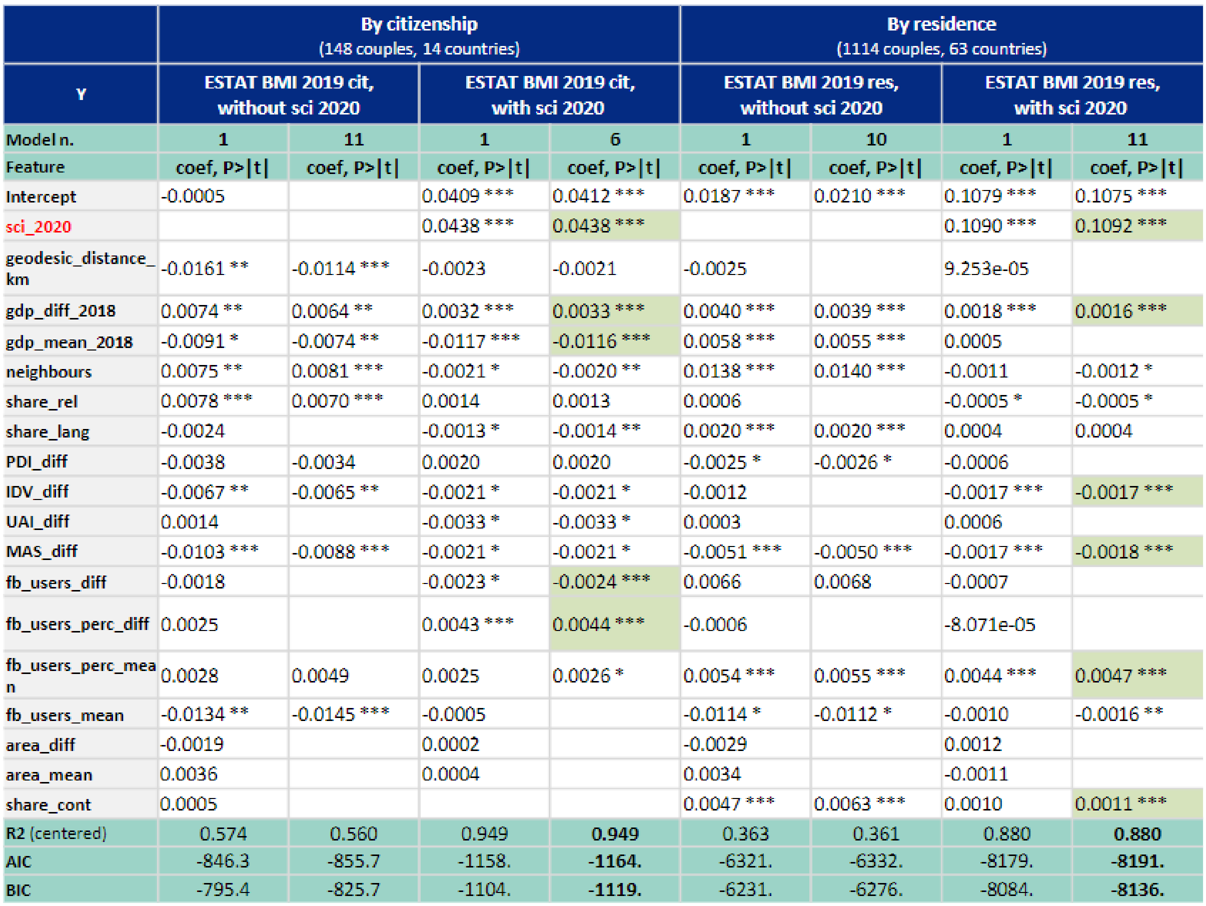
Future work

The ultimate goal of our analysis is to integrate migration drivers with knowledge about past migration flows to build models able to nowcast and forecast migration. We will investigate and test different kinds of Machine Learning models in order to determine the best one in terms of performance outcomes and suitability with respect to our data and task. The different architectures that will be explored in this context are Random Forests, Support Vector Machines and Artificial Neural Networks (e.g., Multilayer Perceptron). The linear regression model we presented here will be employed to rank model features and feed into a filter feature selection method for the upcoming Machine Learning phase. In this way the non-relevant variables resulting from the fit of the OLS model will not be directly excluded, but p-values and coefficients will be exploited for feature ranking.

Conclusion

All in all, our contributions lie in the need for new perspectives, methods, and analyses that can no longer prescind from taking into account a variety of new factors. The heterogeneous and multidimensional sets of data released with MIMI and exploited in the models with the aid of the BMI indicator offer a new overview of the characteristics of human migration, enabling a better understanding and a potential exploration of the relationship between migration and its drivers also through non-traditional sources of data.

[5] related to 2018, as the nearest available year with respect to the SCI reference period



	By citizenship (148 couples, 14 countries)				By residence (1114 couples, 63 countries)			
Y	ESTAT BMI 2019 cit, without sci 2020		ESTAT BMI 2019 cit, with sci 2020		ESTAT BMI 2019 res, without sci 2020		ESTAT BMI 2019 res, with sci 2020	
Model n.	1	11	1	6	1	10	1	11
Feature	coef, P>|t|	coef, P>|t|	coef, P>|t|	coef, P>|t|	coef, P>|t|	coef, P>|t|	coef, P>|t|	coef, P>|t|
Intercept	-0.0005		0.0409 ***	0.0412 ***	0.0187 ***	0.0210 ***	0.1079 ***	0.1075 ***
sci_2020			0.0438 ***	0.0438 ***			0.1090 ***	0.1092 ***
geodesic_distance_km	-0.0161 **	-0.0114 ***	-0.0023	-0.0021	-0.0025		9.253e-05	
gdp_diff_2018	0.0074 **	0.0064 **	0.0032 ***	0.0033 ***	0.0040 ***	0.0039 ***	0.0018 ***	0.0016 ***
gdp_mean_2018	-0.0091 *	-0.0074 **	-0.0117 ***	-0.0116 ***	0.0058 ***	0.0055 ***	0.0005	
neighbours	0.0075 **	0.0081 ***	-0.0021 *	-0.0020 **	0.0138 ***	0.0140 ***	-0.0011	-0.0012 *
share_rel	0.0078 ***	0.0070 ***	0.0014	0.0013	0.0006		-0.0005 *	-0.0005 *
share_lang	-0.0024		-0.0013 *	-0.0014 **	0.0020 ***	0.0020 ***	0.0004	0.0004
PDI_diff	-0.0038	-0.0034	0.0020	0.0020	-0.0025 *	-0.0026 *	-0.0006	
IDV_diff	-0.0067 **	-0.0065 **	-0.0021 *	-0.0021 *	-0.0012		-0.0017 ***	-0.0017 ***
UAI_diff	0.0014		-0.0033 *	-0.0033 *	0.0003		0.0006	
MAS_diff	-0.0103 ***	-0.0088 ***	-0.0021 *	-0.0021 *	-0.0051 ***	-0.0050 ***	-0.0017 ***	-0.0018 ***
fb_users_diff	-0.0018		-0.0023 *	-0.0024 ***	0.0066	0.0068	-0.0007	
fb_users_perc_diff	0.0025		0.0043 ***	0.0044 ***	-0.0006		-8.071e-05	
fb_users_perc_mean	0.0028	0.0049	0.0025	0.0026 *	0.0054 ***	0.0055 ***	0.0044 ***	0.0047 ***
fb_users_mean	-0.0134 **	-0.0145 ***	-0.0005		-0.0114 *	-0.0112 *	-0.0010	-0.0016 **
area_diff	-0.0019		0.0002		-0.0029		0.0012	
area_mean	0.0036		0.0004		0.0034		-0.0011	
share_cont	0.0005				0.0047 ***	0.0063 ***	0.0010	0.0011 ***
R2 (centered)	0.574	0.560	0.949	**0.949**	0.363	0.361	0.880	**0.880**
AIC	-846.3	-855.7	-1158.	**-1164.**	-6321.	-6332.	-8179.	**-8191.**
BIC	-795.4	-825.7	-1104.	**-1119.**	-6231.	-6276.	-8084.	**-8136.**

Figure 2: Results of OLS Backward Elimination Stepwise Linear Regression. Asterisks represent the relevance of each variable according to p-values, as illustrated above.

Acknowlegdements

This work is supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics", and by the Horizon2020 European project "HumMingBird – Enhanced migration measures from a multidimensional perspective", Grant Agreement n. 870661.

References

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018, August). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, *32*(3), 259–80. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.32.3.259 doi: DOI: 10.1257/jep.32.3.259

Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F., & Spyratos, S. (2022). Data innovation in demography, migration and human mobility.
doi: DOI: 10.2760/958409

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). Using twitter data to estimate the relationship between short-term mobility and long-term migration. New York, NY, USA: Association for Computing Machinery.

Gendronneau, C., Yıldız, D., Hsiao, Y., Stepanek, M., Abel, G., Hoorens, S., . . . Weber, I. (2019). *Measuring labour mobility and migration using big data - exploring the potential of social-media data for measuring eu mobility flows and stocks of eu movers*. Publications Office of the European Union.

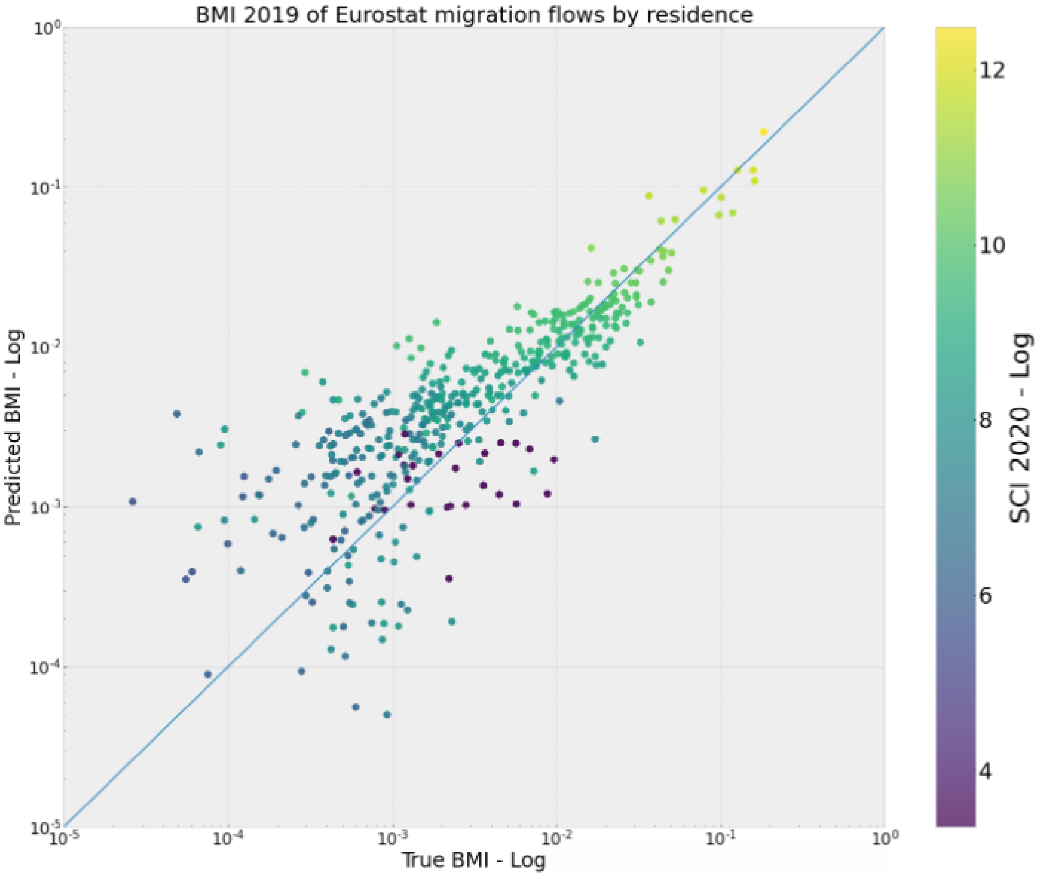


Figure 3: BMI 2019 of EUROSTAT migration flows by residence: comparison between true values and linear regression predictions. This plot refers to the best fit obtained with the linear regression, i.e. the OLS model with $R^2 = 0.88$ corresponding to the last column in Figure 2. Each data point in the plot represents a country pair.

Goglia, D. (2022, March). *Multi-aspect Integrated Migration Indicators (MIMI) dataset.* Zenodo. doi: DOI: 10.5281/zenodo.6493325

Goglia, D., Pollacci, L., & Sirbu, A. (2022). Dataset of multi-aspect integrated migration indicators. doi: DOI: 10.48550/arXiv.2204.14223

Hofstede, G. (1980). Culture's consequences: International differences in work related valuese.

Kaasa, A., Vadi, M., & Varblane, U. (2016). A new dataset of cultural distances for european countries and regions. *Research in International Business and Finance*, *37*, 231-241. Retrieved from https://www.sciencedirect.com/science/article/pii/S0275531915300751 doi: DOI: https://doi.org/10.1016/j.ribaf.2015.11.014

Kaasa, Anneli, V., Vadi, Maaja. (2014). Regional cultural differences within european countries: Evidence from multi-country surveys. *Management International Review*, *54*, 825-852. Retrieved from https://doi.org/10.1007/s11575-014-0223-6 doi: DOI: 10.1007/s11575-014-0223-6

Kim, J., Sîrbu, A., Rossetti, G., Giannotti, F., & Rapoport, H. (2021). *Home and destination attachment: study of cultural integration on twitter.* arXiv. doi: DOI: 10.48550/ARXIV.2102.11398

Meta. (2021). *Social connectedness index.* https://bit.ly/SCIdataset. ([Online; accessed December 2021.])

Pollacci, L., Milli, L., Bircan, T., & Rossetti, G. (2022). Academic mobility from a big data perspective. doi: DOI: 10.21203/rs.3.rs-1510153/v1

Salah, A. A. (2021). Chapter 8: Mobile data challenges for human mobility analysis and humanitarian response. In *Research handbook on international migration and digital technology.* Cheltenham, UK: Edward Elgar Publishing. doi: DOI: 10.4337/9781839100611.00017

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). *Migration data using social media: a european perspective.* Publications Office of the European Union.

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., … Sharma, R. (2021). Human migration: the big data perspective. *International Journal of Data Science and Analytics*, *11*(4), 341-360. doi: DOI: 10.1007/s41060-020-00213-5

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from twitter data. New York, NY, USA: Association for Computing Machinery.

Zagheni, E., Polimis, K., Alexander, M., Weber, I., & Billari, F. C. (2018, 6). Combining social media data and traditional surveys to estimate and predict migration stocks. EAPS.

Zagheni, E., Weber, I., & Gummadi, K. (n.d.). Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*(4), 721–734.