The Potential of the Computational Linguistic Analysis of Social Media for Population Studies

Letizia Mencarini
Bocconi University, Dondena Centre for Research on Social Dynamics and Public Policy
via Roentgen 1, 20136 Milan, I.
letizia.mencarini@unibocconi.it

Abstract

The paper provides an outline of the scope for synergy between computational linguistic analysis and population studies. It first reviews where population studies stand in terms of using social media data. Demographers are entering the realm of big data in force. But, this paper argues, population studies have much to gain from computational linguistic analysis, especially in terms of explaining the drivers behind population processes. The paper gives two examples of how the method can be applied, and concludes with a fundamental caveat. Yes, computational linguistic analysis provides a possible key for integrating micro theory into any demographic analysis of social media data. But results may be of little value in as much as knowledge about fundamental sample characteristics are unknown.

1 The incomplete data revolution in demography

Demography is the study of population. Traditionally, demography is concerned with measuring and estimating population change by births, deaths and migration. Demography is rooted in quantitative methods, with data at its heart. As the field moved through different epochs of data availability, in demography data have always been "big" (Billari and Zagheni, 2017). Starting with the exercise of mapping macro-level trends through population level parameters, based largely on census and administrative records, the field became more theory driven as individual data became available. It is fair to say that with the explosion in available survey data, a revolution in demographic studies took place. Rather than simply describing demographic patterns, today demographers are equally concerned in understanding both the drivers and the consequences of demographic processes. In doing so, demographers have assembled an enormously rich set of data for explaining not only population processes, but also the motivational and behavioral drivers behind these processes. However, data generated by surveys may have peaked. As survey and polling agencies struggle with increasing costs and declining survey response rates, statistic producers are increasingly looking towards big data. Still, given their quantitative pedigree, demographers are perhaps better placed than most other social scientists to take on the challenge of the new big data revolution.

Demographers are, in fact, already using big data to describe demographic processes, including data derived from social media. But there are challenges. Big data is messy and unstructured, and this is a considerable challenge for a scientific field acutely concerned with representativeness and unbiased estimation. Social media provides a promising avenue, however, as demographers are interested not only in describing population processes, but also in the motivations that individuals have for their behavior, which, ultimately, generates observed population processes. For demographers in search of the determinants and consequences of demographic behavior, the linguistic analysis of social media texts can offer a precious and rich new source. Caution – as this paper highlights – is necessary, since its non-representativeness and partiality makes it problematic in social-science terms.

The rapid emergence of big data from social media outpaced social scientists' capacity for using and analyzing them. That having been said,

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 62–68
NAACL HLT 2018, New Orleans, Louisiana, June 6, 2018.

demographers have made a start in exploiting social media data. For example, Reis and Brownstein (2010) show that the volume of Internet searches for abortion is inversely proportional to local abortion rates and directly proportional to local restrictions on abortion. Billari et al. (2013) show that Google searches for fertility-related queries, like 'pregnancy' or 'birth', can be used to predict fertility intentions and consequently fertility rates, several months ahead of them being made public through other data sources. Ojala et al. (2017) use Google Correlate to detect evidence for different socio-economic contexts related to fertility (e.g., teen fertility, fertility in high income households, etc.). Email data have been used to track migrants (Zagheni and Weber, 2012); Facebook data to monitor migrant stocks (Zagheni et al., 2017); patterns of short- and long-term migration (Zagheni et al., 2014); and family change have been derived from Twitter data (Billari et al., 2017). These applications are important, and have demonstrated that the combination of survey and internet data improve predictive power and the accuracy of the described demographic phenomena. Billari and Zagheni (2017) triumphally affirm that the Data Revolution is already here for the study of population processes. However, these studies are all ultimately about describing demographic processes. So far, progress in exploiting content analysis of texts and corpora has been limited, and existing studies have not yet tackled how social media data can explain the behavioral motivations that drive observed population processes. On this point, there is massive potential for synergy between demography and computational linguistics. Certain strands of the social sciences have started looking in this direction, as there are several examples in political science and political economy.

2 Why people's opinions matter

In order to exploit social media data to explain the determinants of population processes, one has, perforce, to delve into the behavioral theories commonly invoked in demographic studies. For population studies, there is no single theory. Instead, being an interdisciplinary science, demographers borrow from a host of theoretical concepts from across the social sciences. One example is the Second Demographic Transition theory (Van de Kaa, 1987; Lesthaeghe, 2010), which has been a point of reference in family demography in recent decades. The theory stems from Inglehart's work (1971). He argued that with the onset of modernization, individuals now cared more about self-realization and less about traditional family life, which consequently fostered new demographic behaviors, such as out-of-wedlock childbearing, cohabitation replacing marriage and fertility decline. In other words, values, attitudes and opinions, play a critical role. Another example concerns the theoretical concept of gender equality and equity. As women increasingly attain the same levels of higher education as men their attitudes change. Other than having children, they also want fulfilling work careers (Esping-Andersen and Billari, 2015; Aassve et al., 2015). The sense of gender equity (Mencarini, 2014) changes as women reach men's level in terms of education, but traditional attitudes may prevail within households. If so, there is a mismatch between gender equity and actual equality, which, McDonald (2000) argues, creates a gender conflict, which eventually leads to lower fertility. Yet, another important theoretical concept originates in economics. Economic models are used to explain changes in divorce, migration drivers, and fertility and so forth. Starting with individual preferences, behavior come out through a process of decision making, where individuals' (presumed) rational evaluations are made in order to maximize their wellbeing. As one moves from survey data to a social media *corpus*, these theoretical concepts offer both challenges and opportunities. On the one hand, new methods, not always familiar to demographers, must be implemented. On the other, there is opportunity in the fact that social science theories can show us what one should be looking for in an otherwise complex and sometimes overwhelming amount of data.

3 Social media linguistic analysis as a middle ground between qualitative and quantitative analysis

One important reason behind the slow progress in the field, is, perhaps, that demographers are more confident with the analysis of numbers than with text: i.e. with quantitative rather than qualitative analysis. Or, perhaps, there is still uncertainty and suspicion about the extent to which social media data can be used to properly infer theoretical concepts for demography. Developments are being made elsewhere in the social sciences. However,

the most prominent examples are based on digitized historical texts. The approach taken is similar to what is being done with social media data, in the sense that one exploits distributional semantic techniques. This is a 'usage-based' theory of meaning built upon similarities of linguistic distributions in a corpus (Lenci 2008), and it allows for the extraction of (near-) synonyms in a context-dependent way, for each document and period under consideration. As we discuss below, the key lies in defining, and coding, the concepts that are to be captured (Kenter et al., 2015, Betti and van den Berg, 2016; Fokkens et al., 2016).

The challenge lies in how theoretical concepts commonly used in demographic analysis (such as the ones mentioned earlier) can be integrated into computational linguistic analysis. Social media has created an extraordinary quantity of potential research material that would have unimaginable even just a few years ago. This material, especially those texts where individuals express opinions through conversation and other ways of communications, where they reveal subjective perceptions, expression of feelings and reasons for their actions, are of tremendous value. Those spontaneous texts are very similar in their nature to certain kinds of qualitative data collection. Texts are the central form of data in qualitative research, in the form of interview transcripts, observations, field notes and primary documents (Mills, 2017). Compared to classical qualitative text analysis, social media texts are much more disordered, but they have two important positive features: they are spontaneous and they are enormous in quantity. These are important issues. The sheer quantity of social media data effectively deals with one of the most frequent criticisms of classic qualitative studies, i.e. the small number of observations. Moreover, classical qualitative studies, do not lend themselves easily to tracking how concepts change over time.

The fact that social media texts are the product of conversations between individuals, groups, and organizations, instead of responses to questions created by researchers (who usually have only *post-hoc* intuitions about the relevant factors in making meaning) is relevant, and gives hints of how perceptions, values, etc., evolve in real time. The quantity of material can, instead, create challenges for social scientists. Often linguistic analysis looks for positive or negative expressions of sentiment. This, though, in itself is not enough. The challenge lies in *how* text data can be investigated for research questions which require closer analysis and nuanced interpretation. But neither traditional qualitative approaches requiring the manual screening and classification of all the material, nor quantitative statistical analysis, can be applied. In this sense, social media data texts provide a middle ground between qualitative studies and more standard quantitative approaches. Some studies have recently and successfully used a mixture of manual coding and machine learning techniques (as discussed next).

4 The analytical approach: the importance of coding

When the concepts of interest are theory driven, they are often complex, multifaceted, and not always directly measurable. Therefore, considerably more effort is needed in annotating texts so as to get meaningful classification results. This, note, is also the case for demographic analysis and for family research.

One method is to combine a conventional classification method in qualitative social science (i.e. manual coding), with algorithmic classification using supervised machine learning. After having collected social media texts over a given period and in a given geographical area, the first step is to get at the texts that contain relevant topics for the research question. This kind of research cannot rely simply on hashtags or other similar holistic tools that allow for the identification of texts and posts. Usually one encounters situations where the potentially relevant data are broad in scope. Consequently, it becomes difficult to identify the presence of information related to the topics one is interested in. The filtering should be based on theoretical guided keywords (using hashtags when available), or by users: i.e. in some cases we are interested in individuals but not companies, institutions or newspapers. Duplicates (e.g. re-tweets) can be deleted. As a result of the filtering, a *corpus* of potentially relevant texts is obtained. The idea is to first manually examine the texts, according to a pre-defined and theoretically-based semantic scheme, thus creating an annotated *corpus* (e.g. of tweet messages). Then an annotation model should be created and operationalized as a clear guide for manual annotators. The approach needs then to be tailored to the specific research question, which may require tweaks. As noted in

Karamshuk et al. (2017), if, for instance, crowdsourcing is used to increase the set of manual labels, slightly different approaches or different decision trees may need to be developed to enable adequate levels of agreement amongst crowd workers. The coding scheme that can be interpreted and applied by crowd workers to create reliable high quality labels is central in this process and clear guidance should be provided for crowd workers. Karamshuk et al. (2017) used a decision tree to help to create greater consistency in labelling. As a result of this fundamental step, what is known as a *gold standard corpus* of annotated texts (with sentiment but also with topics labels) is created. This will constitute the base for the algorithmic classification of the rest of the texts using machine learning, thereby mimicking the human researcher in coding the texts. It is, naturally, important to see how the machine algorithm is able to generate labels in agreement with the crowd labels, i.e. with what levels of accuracy. An acceptable percentage of accuracy from a linguistic point of view, may not be satisfactory to social scientists.

Examples of this analytical approach include Karamshuk et al. (2017) and Mencarini et al. (2017 and 2018), two works from quite different fields with different research questions. Karamshuk et al. (2017) use a case study approach, applying semi-automated coding, for public social media empathy in the context of high-profile deaths by suicide. Five cases were chosen which had a high rate of public response on Twitter, with the aim of exploring what types of response were more or less common in the public Twitter space, and what factors might affect these responses. The analysis suggests that the combination of qualitative analysis with machine learning can offer both a big picture view of public events and a close analysis of particular turning points or key moments in discussions of such events, yielding new insights that were not easily achievable with traditional qualitative social science methods. The paper develops semi-automated coding, where the authors first manually bootstrap a coding scheme from a micro-scale sample of data, then use a crowdsourcing platform to achieve a meso-scale model, and finally apply machine learning to build a macro-scale model.

In Mencarini et al. (2017) the aim is to investigate how computational linguistic techniques can be used to explore opinions and semantic orientations related to fertility and parenthood. There was a two-step approach: first, we developed a Twitter Italian corpus annotated applying a novel multi-layered semantic annotation scheme for marking information not only about sentiment polarity, but also about the specific semantic areas/sub-topics which are the target of sentiment in the fertility-SWB domain. As a reference dataset, we collected all the tweets posted in Italian language in 2014 from the TWITA collection[1]. Then we applied a multi-step thematic filtering, which included a keyword-based filtering stage through the inflection of a list of hashtags and keywords resulting from a combination of a manual content analysis on 2,500 tweets sampled at completely random (taken as a starting point) and a linguistic analysis on synonyms (see Sulis et al. 2017 and Mencarini et al. 2018 for the more details on the development of the corpus). A random sample of about 6,000 tweets has been manually annotated by using the CrowdFlower platform. The annotator's task was, first, to mark if the post is *in-* or *off-topic*[2] (or unintelligible), and then to mark for *in-topic* posts, on the one hand, the polarity and presence of irony, on the other hand, the sub-topics. An analysis of the manually annotated tweets to highlight relationships between the use of affective language and sub-topics of interest has been carried out. This step sheds lights on the social media content of messages related to fertility domains. The end product of this phase is a *gold standard corpus*, TW-SWELLFER, available to the community, which is essentially a body of trustworthy texts used for training and for meaningful evaluation in the next stage. The second phase consisted of a supervised machine learning experiment carried out on the overall dataset and based on the annotated tweets from the previous stage. Employing well-known algorithms from NLP, messages concerning children, parenthood or fertility (*in-topic*) from others (*off-topic*) were distinguished. Also sentiment *polarity,* with a standard annotation (as provided for the Sentipolc shared task in Basile et al. 2014) was de-

[1] http://valeriobasile.github.io/twita/about.html

[2] Topics related to fertility and parenthood. are somehow spread in the dataset and it is not an easy task to filter messages which contain relevant information on such subjects. Then, we decided to apply this manual check to identify and remove noise.

tected. This step was devoted to infer to what extent social media users report negative or positive affects on topics relevant to the fertility domain. The prevalence of positive tweets was then correlated with relevant regional characteristics regarding fertility. Data was derived from tweets in Italian and, since there is currently no up-to-date survey data on individual subjective well-being that can be connected to childbearing and parenthood for Italy, this material is, thus, potentially of real value for socio-demographic research.

5 Features and caveats in the study of demographic behavior

The growing deluge of digitally-generated texts and the development of computational algorithms to analyze them, create an unprecedented opportunity for the study of socio-demographic behavior. First, social media texts allow for the harvesting of opinions which are expressed spontaneously, not responding to a specific question and often as a reaction to some emotional driven observation. Second, social media coverage in time and space offers a continuity that surveys cannot provide. These two features are very important and offer a unique opportunity for learning about social media users and, therefore, for providing new perspectives on socio-demographic behavior.

Still, a fundamental question is who the users are. Which population do they represent? As data is generated from social media platforms, one is necessarily relying on a biased, or non-representative base of users. Despite using data with millions of data points, we are focusing on small biased subsets of the population, which otherwise, should be sampled through parameters such as gender, race, geography, age, income and education. For instance, there are studies suggesting that Twitter users in the Netherlands are young and female with specific personality traits (Nguyen et al., 2013; Plank and Hovy, 2015; Verhoeven et al., 2016). Individuals from such groups, will necessarily provide different kinds of information. In other words, despite the massive quantities of social media data available, we risk ignoring parts of the population, relevant to policy makers and social scientists. There are now efforts being made to overcome this issue. Studies attempt to calibrate non-representative digital data against reliable official statistics, thereby evaluating and modeling possible biases, or, when official statistics are not available, relative trends are compared (Zagheni and Weber, 2015). Some have suggested retrieving information on the socio-demographic traits of Twitter users with the crowd-sourcing platform CrowdFlower and the image-recognition software Face++ (Yildiz et al., 2017) or by manually inspecting data that they have published elsewhere, e.g. on LinkedIn profiles. When age is not given, it could be estimated by taking into account, if present, the information included, say, in the LinkedIn education section, such as the starting date of a degree. Gender could be inferred from profile photos and names, by following a methodology similar to that in Rangel et al. (2014). In particular, the idea of extracting information about the age and gender of users by automatically analyzing their pictures, relying on advanced face-recognition techniques, might allow a novel methodological framework for a demographic-oriented analysis of social media and an assessment of theoretical ideas. Another fundamental piece of information for demographic studies, refers to the geographical location where social media users live or operate. Geocoded texts are available of course, but again, not universally so (e.g., in Mencarini et al. 2017 only one out of four messages were geo-tagged), and establishing residence is difficult since a large number of social media texts are generated on portable devices. Nevertheless, these stable or semi-stable socio-demographic traits of users are fundamental in making sense of social media data for demographic purposes, not least because they are instrumental in judging the representativeness of the social media sample applied.

6 The end of theory is not here, yet

The message of this paper is twofold. First, computational linguistic analysis offers great potential in advancing social science and demographic analysis. To do so successfully, however, one must develop an annotation procedure to incorporate the key theoretical concepts from the social sciences. On this point, social sciences and demography have the potential to provide huge advances in computational linguistics analysis. Second, there is no way (yet), to ignore the issue of representativeness. For social media data to make sense for demographic analysis, or more generally, for the social sciences, one needs to know something about the sample used for one's analysis. Perhaps one day we will reach the point where the quantity

of big data is so huge, so all encompassing, and so comprehensive, that it will capture and answer all possible social questions. In the defense of the classical approach, however, one can always argue that such data will produce biases; and that there will be digital divides, both in the way information and technology is produced (Graham, 2012). Despite the enormity of digital data and the development of statistical tools designed to crunch data, social scientists will, at least for the foreseeable future, set the research questions and agendas, search for causation, and contribute useful theories for demographic analysis. As such, we are still some distance away from the supremacy of unsupervised machine learning, where the power of correlation supersedes causation, and where an epistemological revolution will effectively end social theory simply by letting data speak for themselves (Anderson, 2008; Chandler, 2015). At least for research into socio-demographic behavior, sociologists and demographers, with computer scientist colleagues, will still, for some time yet, be in the business of torturing the data until they talk.

References

Arnstein Aassve, Letizia Mencarini, and Maria Sironi. 2015. Institutional change, happiness and fertility. *European Sociological Review*, 31(6), 749-765. https://doi.org/10.1093/esr/jcv073

Anderson Chris. 2008. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, 23 June.

Ben Y. Reis and John S. Brownstein. 2010. Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC public health*, 10(1): 514. https://doi.org/10.1186/1471-2458-10-514

Arianna Betti, and Hein van den Berg. 2016. Towards a Computational History of Ideas. In *Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age. CEUR Workshop Proceedings.* CEUR-WS.Org, edited by Lars Wieneke, Catherine Jones, Marten Düring, Florentina Armaselu, and René Leboutte. Vol. 1681. Aachen.

Francesco C. Billari, Nicolò Cavalli, Eric Qian, and Ingmar Weber. 2017. Footprints of Family Change: A Study Based on Twitter, Paper presented Annual Meeting of the Population Association of America, Chicago, IL, April 27-29 2017.

Francesco C. Billari, Francesco D'Amuri, and Juri Marcucci. 2013. Forecasting births using google. Paper presented at Annual Meeting of the Population Association of America, New Orleans, LA, April 11-13 2013.

Francesco C. Billari, and Emilio Zagheni. 2017. Big Data and Population Processes: A Revolution?. SocArXiv. July 1, published also in: Alessandra Petrucci, Rosanna Verde (edited by), SIS 2017. Sta tistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). *Proceedings of the Conference of the Italian Statistical Society*, Firenze University Press, 2017, pages 167–178. https://doi:10.17605/OSF.IO/F9VZP

David Chandler. 2015. A world without causation: big data and the coming of age of posthumanism. *Millennium: Journal of International Studies,* 43(3): 833–851. https://doi.org/10.1177/0305829815576817

Gosta Esping Andersen, and Francesco C. Billari. 2015. Retheorizing Family Demographics. *Population and Development Review*, 41(1), 1-31. https://doi.org/10.1111/j.1728-4457.2015.00024.x

Antske Fokkens, Serge ter Braake, Isa Maks, and D. Ceolin. 2016. On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change, paper presented at "Drift-a-LOD", Detection, Representation and Management of Concept Drift in linked Open Data, Workshop EKAW, Bologna, Italy, 20th November, 2016.

Mark Graham. 2012. Big data and the end of theory?, *The Guardian*, 9 March 2012.

Ronald Inglehart. 1971. The Silent Revolution in Europe: Intergenerational Change in Postindustrial Societies. *American Political Science Review*, 65: 991–1017. https://doi.org/10.2307/1953494

Dmytro Karamshuk, Frances Shaw, Julie Brownlie, and Sastry Nishanth, 2017. Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide, *Online Social Networks and Media*, 1: 33-43. https://doi.org/10.1016/j.osnem.2017.01.002

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad Hoc Monitoring of Vocabulary Shifts over Time. *CIKM '15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Pages 1191 – 1200.

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics*, 20: 1–31.

Ronald Lesthaeghe. 2010. The unfolding story of the second demographic transition. Population and

Development Review, 36(2): 211-251. https://doi.org/10.1111/j.1728-4457.2010.00328.x

Peter McDonald (2000). Gender equity, social institutions and the future of fertility. *Journal of population Research*, 17(1), 1-16. https://doi.org/10.1007/BF03029445

Letizia Mencarini. 2014. Gender equity, In: Michalos AC (Ed*.). Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht, Netherlands: Springer, Pages 2437-2438.

Letizia Mencarini, Viviana Patti, Mirko Lai, and Emilio Sulis, Happy parents' tweets. 2017. In: Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges*, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society, Firenze University Press, 2017, 693-700. https://doi:10.17605/OSF.IO/F9VZP

Letizia Mencarini, Delia Irazú Hernández-Farías, Mirko Lai, Viviana Patti, Emilio Sulis, Daniele Vignoli. 2018. *Italian happy parents in Twitter*, Dondena WP 117, Bocconi University.

Kathy A. Mills. 2017. What are the threats and potentials of big data for qualitative research?, *Qualitative Research*, First Published November 30. https://doi.org/10.1177/1468794117743465

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.*

Jussi Ojala, Emilio Zagheni, Francesco C Billari, and Ingmar Weber. 2017. Fertility and its meaning: Evidence from search behavior. *Proceedings of the International Conference on Web and Social Media* (ICWSM) 2017.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–98.

Francisco Rangel Pardo, Paolo Rosso, Irina Chugur,Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, Walter Daelemans. 2014. Overview of the 2nd author profiling task at PAN 2014, in: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (eds.), CLEF 2014 Labs and Workshops, *Notebook Papers*, 1180, CEUR-WS.org, pages 898-927.

Emilio Sulis, Cristina Bosco, Viviana Patti, Mirko Lai, Delia Irazú Hernández Farías, Letizia Mencarini, Michele Mozzachiodi, Daniele Vignoli. 2016. Subjective Well-Being and Social Media. A Semantically Annotated Twitter Corpus on Fertility and Parenthood. *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, Napoli, Italy, December 5-7, 2016. CEUR Workshop Proceedings volume 1749, CEUR-WS.org.

Dirk J. Van de Kaa. 1987. Europe's second demographic transition. *Population Bullettin,* 42(1):1–59.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TWISTY: a Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016).

Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, and Jennifer A. Holland. 2017. Using Twitter data for demographic research. Demographic Research, 37 (46). https://doi.org/10.4054/DemRes.2017.37.46

Emilio Zagheni, and Ingmar Weber. 2012. You are where you E-mail: Using E-mail Data to Estimate International Migration Rates. *Proceedings of the 4th Annual ACM Web Science, Evanston, IL*, pages 348-351.

Emilio Zagheni, Kiran Garimella, and Ingmar Weber. 2014. Inferring international and internal migration patterns from twitter data. In *Proceedings of the23rd International Conference on World Wide Web*. ACM, pages 439-444.

Emilio Zagheni, Ingmar Weber. 2015. Demographic research with non-representative internet data. *International Journal of Manpower*. 36(1): 13-25. https://doi.org/10.1108/IJM-12-2014-0261

Emilio Zagheni, Ingmar Weber, Krishna Gummadi. 2017. Estimate stock of migrants using Facebook's advertising platform, *Population and Development Review*, on-line first. https://doi.org/10.1111/padr.12102