Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data

Feasibility study on Inferring (labour) mobility and migration in the European Union from big data and social media data

Christina Hughes[1], Emilio Zagheni*[1], Guy J. Abel[2], Arkadiusz Wi´sniowski[3], Alessandro Sorichetta4, Ingmar Weber[5], and Andrew J. Tatem[4,6]

1. University of Washington, Seattle (*Corresponding author: emilioz@uw.edu)
2. Wittgenstein Centre, Vienna Institute of Demography
3. University of Manchester
4. University of Southampton
5. Qatar Computing Research Institute, Doha
6. Flowminder Foundation, Stockholm

January – 2016

Social Europe

Report prepared for the European Commission project #VT/2014/093.

This report has received financial support from the European Union Programme for Employment and Social Innovation "EaSI" (2014-2020). The information contained in this publication does not necessarily reflect the official position of the European Commission.

EUROPEAN COMMISSION

Directorate-General for Employment, Social Affairs and Inclusion
Directorate A — Employment and Social Governance
Unit A4 — Thematic Analysis
Contact: Filip Tanay

E-mail: filip.tanay*@ec.europa.eu*

European Commission
B-1049 Brussels

Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data

Feasibility study on Inferring (labour) mobility and migration in the European Union from big data and social media data

Directorate-General for Employment, Social Affairs and Inclusion

2016

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

LEGAL NOTICE

This document has been prepared for the European Commission however it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

More information on the European Union is available on the Internet (http://www.europa.eu).

Luxembourg: Publications Office of the European Union, 2016

ISBN: 978-92-79-59742-8
doi: 10.2767/61617

© European Union, 2016
Reproduction is authorised provided the source is acknowledged.

Table of Contents

SUMMARY

This report addresses the question of whether it is technically, financially and legally feasible to estimate geographic mobility and migration flows in the European Union. Our assessment indicates that the feasibility is dependent on a number of factors:

1. It depends on the data that one can have access to. Some data sources can be accessed by anyone with the appropriate technical skills (e.g., samples of Twitter tweets); some can be purchased (e.g., historical tweets); some are not for sale and require partnerships with companies (e.g., Yahoo!, Facebook, LinkedIn, and mobile phone providers); some are not shared by companies (Google does not share data, except for some aggregate indexes, like the ones in Google Trends).

2. It depends on the outcome desired. Estimating trends or changes in trends in migration flows is feasible and can be done in a timely manner. Getting accurate and precise estimates for special populations, like refugees, may or may not be feasible depending on the context: it would require further research. Likewise, obtaining estimates of short-term migration by education, gender or employment status is feasible. Obtaining unbiased estimates of short-term mobility from a single, non-representative source would be more difficult. It may be feasible in some circumstances (e.g., when the data set is rich enough for the use of post-stratification techniques), but not in others.

3. It depends on legal obstacles. Companies may have terms and conditions or non- disclosure agreements for data sharing that may or may not include inconsistencies with the rules governing universities and funding agencies. We have not identified major issues in this area, but each individual collaboration across units would require some careful examination of the terms and conditions in order to resolve any potential lack of consistency.

There is no ideal data set that fits all needs. To estimate professional migrations or other labour market indicators for a specific segment of the workforce, LinkedIn is probably the best data source available. However, if the goal is to obtain information on low-skilled labour migration, then LinkedIn is not appropriate.

Different data sources would give slightly different estimates for the same quantities, similarly to what different surveys do. Some sources may be more reliable/less biased than others for specific goals. Some sources may exist and be available now, but may disappear in a few years. Some new sources may emerge in the meanwhile. Some sources may have more demographic information about the users, and may add complementary dimensions to the study of the same phenomenon.

In this report we discuss some methodologies that have been applied mainly to demographic issues. After examining strengths and weaknesses of various approaches and data, our main conclusion is that it is key to develop methods that leverage all existing sources and that are robust to the lack of a specific data source either at some point in time or for a specific geographic area (either because the service disappears, or because the terms of service for the provider change, or because of any other unexpected reason). Bayesian methods can be used effectively to combine different migration data in a consistent way. Various sources of information can be incorporated into the estimation of the true flows as prior probabilities in a hierarchical Bayesian model. Although there is no known example of such a study in the context of big data and migration processes, we believe that it is a promising approach and we will work on developing a framework to incorporate traditional and new data sources for migration within a Bayesian model that can be easily adapted to combine data from a range of sources and control for a variety of measuring issues (including those that arise from big data sources).

INTRODUCTION

International migration is an important driver of demographic change in the European context, and a relevant source of uncertainty for population and fiscal projections. The availability of timely and accurate migration statistics is key for enabling informed policy decisions. Although there have been efforts to produce harmonized statistics in Europe, there is still substantial uncertainty about migration flows, partially because of lack of timely and comparable data for most countries. Moreover, traditional data sources are often not appropriate to measure short-term mobility and may suffer from underestimation problems, due to under-registration issues.

The rapid and global spread of Internet applications, social media and mobile phones has transformed the way in which we communicate with each other. It has also had a transformative impact on the way in which we study our societies. As a result of the digital revolution, new forms of data collection as well as new ways of harvesting so-called "digital breadcrumbs" have emerged. A key feature of the new, often big, data sources is the increased availability of geo-located information. Digital records with spatial attributes hold the promise of exceptional development of new demographic knowledge, particularly in the context of migration and mobility.

Big data for demographic research offers important new opportunities, but it also comes with a number of challenges. In this study we present and review state-of-the-art approaches in the area of migration estimations with traditional and new data sources. The main goals of this paper are: (i) to provide an overview of traditional and new data sources for the study of migrations; (ii) to review methods and results related to the use of new and innovative data sources; (iii) to discuss the feasibility of scaling existing approaches, in particular with regards to potential legal or technical barriers.

The first section reviews the state-of-the-art methods to measure stocks and flows of migrants using traditional data sources. This section includes a presentation of the available data sources as well as a discussion of indirect methods and the respective limitations. The second section provides a description and classification of new and innovative data sources that have been used recently in the migration and mobility literature. This section includes some discussion of forms of access to the data and issues of privacy protection. The third section presents relevant substantive results described in the literature about using data from mobile phones, social media, and other Web data to infer patterns of migration and mobility. This section also includes a schematic summary of key features for the main new data sources used in the recent literature about migrations and geographic mobility. More specifically, a summary table lists the main data sources, the type of access to the data that researchers can potentially have, the cost, the geographical coverage, the key indicators that can be extracted from the data, and the relevant literature that has either relied specifically on the data sources or can be hypothetically applied to the data sources mentioned. The fourth section summarizes and discusses methodological aspects related to the analysis of new and promising data sources that are typically not representative of the underlying population. The fifth section focuses on the feasibility of using new data sources to complement traditional ones, and the potential barriers related to data access, data sharing, as well as technical and legal issues. The article ends with concluding remarks about existing bottlenecks in the analysis of Internet, social media and mobile phone data for migration research along with a discussion of a potential approach to combine existing data sources within a unified framework.

MEASURING INTERNAL AND INTERNATIONAL MIGRATION WITH TRADITIONAL DATA

Migration is an event with two main dimensions: spatial and temporal (Willekens, 2008). That is, for a person to be considered a migrant, he or she needs to cross a predefined border (i.e. either a country or other administrative borders) and stay in a new place for a specified amount of time. Additionally, a person often has to fulfil various requirements to be formally considered a migrant in a given country. These criteria are typically defined by national statistical authorities when carrying out censuses, population registers, administrative data collection and surveys from which measurements of migrant stocks and flows are traditionally monitored.

Sources of data for migration

Censuses

Censuses are the most comprehensive source of information about the entire population of a given country at a given point in time. They usually take place every five or ten years and the UN actively promotes carrying out a census in each country of the world (United Nations, 2008). They provide crucial demographic and socio-economic characteristics of the population, such as age, sex, education, and occupation (Bilsborrow et al., 1997; United Nations, 2008). Summaries of data are given at varying geographic distributions. Typically these have been at regional levels. Many modern censuses and population registers provide outputs for relatively small spatial units.

Internal migration is measured from questions about the residence duration in the current location or the previous place of residence prior to the census night (usually either 1, 2, 5 or 10 years earlier). Some censuses directly ask for the date of the last change in usual residence (Bell et al., 2015). International migration is measured from census questions on nationality and/or country of birth, which can then be used to calculate the size of migrant stocks in a given country and their geographic distribution. Information about the previous country of residence, or the period of arrival in the current country of residence, allows for identification of migration flows in a transitory approach (i.e. comparing population at two points in time). Further, censuses are often used as a benchmark population, which is updated every year with vital events recorded in the population register, and migrations, to produce estimates of the population. Finally, censuses serve as sampling frames for many large-scale surveys which may be later used to estimate migration. The main limitations of censuses, with regard to monitoring migration patterns, are two-fold. First they are expensive to carry out. Second they are undertaken infrequently. As a result, census data on migrants are rapidly outdated, especially for day-to-day policy making, administration, and allocation of public funds. Due to the complexity in collating results, there is typically a lengthy period between the date(s) of data collection and the publication of results. It also precludes a thorough analysis of short- and mid-term migration dynamics, as well as its reasons and consequences. Further, their non-continuous nature neglects all international and internal migration events that take place between censuses. Migrants leaving the country immediately before the census are also not captured.

Population registers

Population registers are databases maintained by the authorities at central or local level, providing an inventory of the population living in the given area. They are continuously updated with information on vital events which can include migration events (i.e. flows) (Rees et al., 2000). Population registers tend to cover the entire population including migrants who are often legally obliged to register in or deregister from it (Bilsborrow et al., 1997). Furthermore, especially in Europe, their importance

as sources of population information has recently increased as censuses are replaced with register-based censuses (Coleman, 2013). However, although population registers are largely used to generate internal migration data in Europe and some parts of Asia, differences on how the registers are designed (e.g. definition of residence) and on the population that they should cover (e.g. exclusion of foreign citizen) can complicate their use for making comparisons among different countries (Bell et al., 2015). In some countries, such as the Netherlands, there is evidence that many Central and Eastern European citizens did not register, leading to an underestimation of their numbers (see for instance (van Ostaijen et al., 2015). A special case of a population register is a register of foreigners (e.g. Central Register of Foreigners in Germany), which records only nationals of other countries. While population registers in Europe are generally considered to provide high quality statistics on vital events, there are several reasons why the measurement of international migration flows and resulting stocks is problematic.

First, registers do not necessarily cover the entire population (e.g. nationals of a given country may not be required to register such as EU nationals in France or the UK) and different definitions can be applied to various subpopulations (e.g. different duration of stay criteria for nationals and foreigners) (Kupiszewska and Wiśniowski, 2009).

Second, final statistics may be distorted by the data processing and dissemination procedures (Kupiszewska and Nowok, 2008).

Third, application of various legal criteria in the register in various countries leads to lack of comparability of the migration statistics (Poulain et al., 2006; Kupiszewska and Nowok, 2008; Abel, 2010; De Beer et al., 2010; Raymer et al., 2013).

Finally, population registers are likely to undercount the number of migrants, especially emigrants. Undercount of immigrants usually results from the lack of legal requirements and incentives to register, after which a migrant would gain access to public services in the destination country. For example, freedom of movement in the EU allows relocations without being recorded in the registers. The emigrants have even fewer incentives to deregister upon leaving the origin country. They may actually prefer to not cut the ties with their origin country due to risks involved in relocating to another country (Bilsborrow et al., 1997). Incentives to de-register may also exist and have a positive impact on the measurement of emigration, for example in the case of Lithuania, which has a compulsory health insurance for its residents.

Surveys

National survey programmes are widely used to collect information about internal migration (Bell et al., 2015). For example the USAID's Demographic and Health Survey (DHS) and the World Bank's Living Standards Measurement Study (LSMS) are used in developing countries Elsewhere, the European Union Labour Force Surveys, the American Community Survey and the Canadian National Household Survey are used in developed countries. The latter two are the only sources of internal migration data, recently replacing the census questionnaires in the US and Canada. As with censuses, these surveys usually contain information about the place of birth, previous place of residence (in some cases referring to a fixed time interval) and residence duration in the current location within the country. Additionally, DHS and LSMS also contain information about migration from rural to urban areas. Although national surveys are carried out more frequently than censuses, some may cover only parts of a country.

Some countries, where regular population registers do not exist (e.g. the United Kingdom) or do not record migration events (e.g. Ireland), surveys to measure international migration flows, migrant stocks, or both are utilized. Three types of surveys are commonly used to identify and measure migration: (i) passenger surveys (flows), (ii) large-scale house- hold surveys (stocks), and (iii) specialized surveys. For

example, migration flows to and from the UK are computed by using the International Passenger Survey (IPS) which surveys passengers arriving in the UK in all major airports and sea routes, Eurostar terminals and on Eurotunnel shuttle trains. Migrant stocks in the UK are estimated using large-scale surveys, such as Labour Force Survey and the Annual Population Survey (Singleton et al., 2010). Specialized surveys, such as the one conducted by (Groenewold and Bilsborrow, 2008), usually target the migrant population and are capable of providing detailed and reliable characteristics of migrants.

Each type of survey suffers from various weaknesses when monitoring migration patterns. Passenger surveys can be very limited when measuring migration. They are primarily designed to capture data on travelling people rather than migrants. Further, data on potential migrants are based on respondents' intentions rather than the actual deeds. Finally, they may not be truly representative, as surveyed travellers cannot be treated as a random sample from the pre-specified population. Large-scale surveys are likely to suffer from issues related to non-response and to relatively rare migrant events. However, the non-response rates vary across countries. This is partially related to the fact that in some countries the questionnaires are available in more than one language. For examples from the UK see (Thomas, 2008) and (Barnes, 2008). For these reasons, specialized surveys may include non-probabilistic samples.

Administrative sources

Administrative sources, such as data from healthcare databases, can be used to estimate internal migration flows. However, in doing so, it has to be considered that administrative sources usually cover only part of the population and that, often, there are no legal obligations ensuring both a complete and timely registration (Bell et al., 2015).

Data on international migration events and migrants can also be collected during various administrative procedures, such as issuing visas, work and residence permits, and registrations of foreigners with a general health practitioner. They provide information relevant to the authorities issuing these documents and can be used to measure migration flows and migrant stocks. An example here comes from the UK data on National Insurance Numbers (NINo) provided by the Department of Work and Pensions, which can provide information on international immigration, but not emigration. However, once a person is issued a visa or a permit, the authorities may not track renewals or changes of citizenship (Bilsborrow et al., 1997). Further, important characteristics such as nationality or country of birth may be removed during data processing and dissemination. Methods for overcoming these limitations include linkage of micro data between registers and other sources, such as surveys (Raymer et al., 2012).

Border statistics

Border statistics contain information on all persons entering or departing a country, regardless of the purpose of their visit. Typically these statistics do not distinguish between migration movements and other types of mobility, such as tourism or visiting. Further, the greater the possibility of entering a country via illegal channels, i.e. evading the border control, the less reliable border statistics is. In the Schengen Area of the EU, there is no systematic border control between the countries due to the Schengen Agreement's clause of freedom of movement, which renders this method of measuring migration infeasible. However, a notable exception is the UK, outside the Schengen Area, where exit checks of all passengers leaving the country have been reintroduced in April 2015 (Border Force, 2015).

Cross National Data

During the last three decades there have been various efforts to collect internal migration data from multiple countries. The first attempt to establish a global inventory of internal migration datasets was made by the (United Nations, 1978) involving moves within 121 countries. More recently, collections of internal migration data focused on specified regions or group of countries (Nam et al., 1990; Rees et al., 1996; Rees and Kupiszewski, 1999; United Nations, 2000; Vignoli and Busso, 2009; United Nations, 2009). Further collections of migration data are presently available from the Integrated Public Use Micro data Series International (IPUMSI[1]) project, the UN Economic Commission for Latin America and the Caribbean[2] and Eurostat[3]. One of the newest and largest collections of internal migration data was used in the International Migration Around the GlobE project (Bell et al., 2015). The project established an inventory of internal migration datasets for 193 UN member states (including all European countries). As well as compiling together data from each country, the project provides strengths, limitations and ways of using the different measures of internal migration as well as suggests how to harmonize the data for comparing internal migration levels and patterns among different countries (Stillwell et al., 2014).

International migration flow data from all EU Member States as well as EFTA members and several others like the US, Russia and Armenia, are accessible from a number of international organizations, including Eurostat (the statistical office of the EU). Availability has been aided by policy makers of the European Parliament who have introduced legislation for the supply of international migration flow data. In 1976, Community Regulation No 311/76 required members to supply migration statistics annually to Eurostat. In 2007, Regulation No 862/07, obliged members to provide migration statistics which comply with a harmonized definition. Research projects such as Towards the Harmonisation of European Statistics on International Migration (THESIM) and MIgration MOdelling for Statistical Analyses (MI- MOSA) fully documented differences between data collection methods and measurements used by national statistics institutes. The United Nations and OECD regularly publish a similar collection of international migration flow data from predominantly developed nations over the last two decades. As with the Eurostat data, migration measures are based on the individual countries' collection methods and definitions which limit direct comparisons. In principle, the 2007 Regulation No 862/07 should lead to migration data that are submitted to Eurostat being harmonised by the member states to the common definition. However, countries apply harmonisation procedures in their own respects and using their own data resources and techniques, without necessarily exchanging information with other countries, which may lead to discrepancies[4] The United Nations population division also provides data on net migration flows for all countries during five-year periods back to 1950 as part of their World Population Prospects, published every two years.

As migrants stocks are relatively easier to measure than flows, estimates are available across more countries and with fewer comparability issues. The World Bank (Ozden et al., 2011) provides foreign-born migration stock tables at the start of each decade, from 1960 to 2000, for 226 countries. Data are primarily based on place of birth responses to census questions or details collected from population registers. Where no data were available, alternative stock measures such as citizenship or ethnicity are used. For countries where no stock measures were available, missing values were imputed using various propensity and interpolation methods, typically dependent on

[1] https://international.ipums.org/international/
[2] http://www.cepal.org/en
[3] http://ec.europa.eu/eurostat/data/database
[4] See 2009 annexes to the migration metadata http://ec.europa.eu/eurostat/cache/metadata/en/migr_immi_esms.htm. Metadata for some countries (e.g. Poland) are missing. Further, currently provided harmonised statistics also contain large discrepancies

foreign-born distributions from available countries in the region. The United Nations Population Division (2013) provides foreign- born migrant stock tables at the start of each of the last three decades (1990, 2000, and 2010) covering 230 countries. As with the World Bank estimates, data are primarily based on place of birth responses to census questions and register information. Abel (Abel, 2015) compares the World Bank and UN stock data and highlights potential weaknesses which tend to originate in alternative methods used to impute missing migrant stocks or the use of a mix of place of birth and citizenship data. The OECD has compiled migrant stocks from 2000 onwards in over 100 countries including all OECD member countries and some other non-OECD countries. The datasets include information on demographic characteristics (age and gender), duration of stay, labour market outcomes (labour market status, occupations, sectors of activity), fields of study, educational attainment and the place of birth.

Synthetic Estimates of Migration from Traditional Data

When traditional data are missing, indirect methods have been developed to estimate migrations based on traditional data sources. Such methods are used to infer migration flows in time periods, such as non-census years, where no migration data are available or to calculate flows between areas where no data have been collected. These are typically based on fitting a model to available data and using the relationships with known auxiliary data to update or predict a synthetic estimate of migration.

Internal migration

There is a vast literature on the modelling of different types of internal migration at various spatial and temporal scales (Stillwell, 2005). At a macro level, models to either predict internal migration flows and understand their determinants have evolved from the gravity model developed by (Zipf, 1946) to explain people's moves given the population in both the origin and destination of each region and the distance between them. Advanced spatial interaction models incorporate other demographic, socio-economic and environmental factors considered to be determinants of migration. For example, Garcia et al (2014) used harmonized IPUMSI micro census internal migration data and a suite of gravity type spatial interaction models that included additional demographic, socio-economic and environmental covariates to estimate subnational migration in 10 countries in sub-Saharan Africa and determine if these models could be applied where internal migration data are not available. Additional covariates used in the models includes: contiguity between origin and destinations, proportion of urban population, proportion of males, median population age, proportion of active population, and a long- and a short-term index of rainfall variability at origin and destination. A similar approach is now being used by (Tatem et al., 2015) to model subnational migration at global scale in all malaria endemic countries.

International migration flows

Motivated by the historical low quality of the official international migration statistics in Europe (Poulain et al., 2006), various endeavours have been undertaken to improve the availability and comparability of data. Using the differences between the reports of the same bilateral flow from both the sending and receiving country, methodologies have been developed (Raymer, 2007; Abel, 2010) to produce synthetic estimates of harmonised flow data between European countries. A Bayesian model for harmonising and correcting the inadequacies in the available data and for estimating the completely missing flows was proposed in the project Integrated Modelling of European Migration (IMEM) (Raymer et al, 2013; Wiśniowski, 2013). It provides estimates of the international migration flows amongst 31 countries in the EU and EFTA in 2002-2008, using the official data on migration flows published by Eurostat. The methodology integrates the data on migration from the sending and receiving

countries, covariate information and elicited expert judgement to produce a database of migration flows with measures of uncertainty. The IMEM model reconciles various measurement problems: undercount, varying duration of stay criteria, coverage and accuracy of the data collection method applied in various countries. A similar idea was employed by Wiśniowski (2013) to estimate migration flows from Poland to the UK in 2002-2008. In the Bayesian model, the Labour Force Survey data from both countries are combined to estimate migration flows for less than twelve months, with measures of uncertainty. These results are further combined with the IMEM output to produce estimates of total migration from Poland regardless of the duration of stay in the UK.

At the global level, Abel (Abel, 2013; Abel, 2015) and Abel and Sander (Abel and Sander, 2014) developed a methodology to estimate bilateral migrant flows between all countries from changes in the migrant stock data from the World Bank and UN. Research on integrating internal and international migration data to estimate migration flows between subnational administrative units located in different countries is beginning to emerge (Tatem et al., 2015; Dennett et al., 2013).

NEW DATA: DESCRIPTION OF MAIN SOURCES RECENTLY USED IN THE LITERATURE

This section deals with a description and classification of new and innovative data sources that have been recently used in the migration and mobility literature. The goal of this section is to provide some background information about new sources, such as mobile phone data and social media data, which have been produced for purposes other than research. These digital 'breadcrumbs' have been leveraged to improve our understanding of migration processes. The next section presents some of the main substantive results in the literature.

Mobile phone data

Data 4 Development

Data 4 Development is a research innovation challenge in which the telecommunications companies Sonatel and Orange Group made anonymized mobile phone call detail records (CDR) available to researchers. The goal of the challenge is to allow researchers to draw on call detail records to research topics in health, agriculture, transportation, urban planning, energy, and national statistics. In addition to these substantive areas, the challenge also called on researchers who wished to contribute to more technical data challenges surrounding the use of call records, including advances in anonymization, data mining, and cross-referencing data. The last challenge was for Senegal and ran from April 2014 to April 2015. A previous challenge provided anonymized CDR for the Ivory Coast.

Reality Commons Project

Sponsored by the MIT Human Dynamics Lab, the Reality Commons project focuses on addressing some of the common challenges of using mobile phone data for research. Since call detail record data are typically anonymized and difficult to leverage for social science research, the project collects cell and other data on communities of about 100 people each. The Friends and Family dataset collects cell phone and other data on decision-making from members of young families. The Reality Mining dataset followed the mobile phone use patterns (including locations) of 75 MIT students in 2004 in order combine mobile phone data with highly granular personal data. Lastly, the Social Evolution dataset tracked the daily routines of an entire dormitory at MIT

with mobile phones, recording the locations, proximities, and calls of participants from October 2008 to May 2009. All of the data from the Reality Commons project are now available as open source projects, which can be accessed via the funf open-source sensing platform for Android phones at http://funf.media.mit.edu or at cost at http://sociometricsolutions.com.

Social media data

Twitter

Twitter is an online social networking platform through which users share, send, and read short, 140-character messages, called 'tweets.' 'Retweets' are tweets that are shared verbatim by other users. Users must register and set up profiles, which can either be publicly shared with both users and non-users or privately managed so that only those approved by the user can view the user's tweets. Twitter can be accessed through a website interface, SMS, or mobile device app. Created in 2006, Twitter announced that it had more than 500 million users by its 8th birthday.

Twitter data have been used to study mobility in various ways, both through geocoded tweets and through natural language processing tools. In order to access the data, Twitter allows low latency access to its streaming APIs. However, to acquire historical Twitter data, researchers must pay for access through a managing service like GNIP (https://gnip.com/). The cost of the data varies since each data request is customizable.

Foursquare

Foursquare is a service that allows users to create a personalized search experience. The service itself has undergone several iterations. Present throughout these changes has been the personalized search and recommendation feature in which the service takes into account the places the user has gone, the things the user told the app that they like, and the other users whose advice the user trusts in order to make recommendations on the best places to go near the user's current location. Until 2014, Foursquare offered a social networking feature that allowed users to 'check in' to places and share this with their friends. In contrast to an earlier version of the service, this latest version allowed check-ins to draw on smart phones' GPS capabilities to record exact longitude and latitude coordinates. This feature was later removed from Foursquare, re-tooled, and launched as its own separate app called 'Swarm' in 2014. As of 2013, Foursquare counted around 45 million registered users.

Foursquare allows researchers to stream their check-in data, but details are to be arranged through direct contact with Foursquare itself at api@foursquare.com. Access to historical Foursquare data can also be arranged on a customizable, case-by-case basis, through GNIP or other similar service.

Flickr

Flickr is a Web and mobile phone service that provides image and video hosting. Users utilize Flickr to share and embed media, creating an online community that can view, upload, and share images and videos. As of 2013, Flickr had a total of 87 million registered users and over 3 million uploads per day. Though non-users may access content on Flickr, registration is required in order to upload content. Geotagging features are available through Flickr, which can be especially useful for those interested in studying mobility. Access to the Flickr API[5] is granted through the service itself. Potential researchers must apply for a key code, which provides them with access to the Flickr API and various methods to pull data from it. Flickr data can also be acquired through the reseller GNIP[6] or other similar services, again, at a customizable price.

Facebook

Facebook is a social network service through which users creates profiles with multidimensional socio-demographic information on sex/gender, employment history, education, tastes, interests, place of birth, place of residence, friendship circles, family members, and so on.

Users primarily interact through direct posts to friends' 'walls' (i.e. personal profile pages) or through 'status updates' in which users post content to their entire network of friends. Created in 2004, the site was first exclusive to students at Harvard University, and then expanded to colleges in the Boston area, the Ivy League, and Stanford University. It eventually opened to all college students, then high school students, and finally to all users older than 13 years of age. As of 2014, Facebook reported having over 1.3 billion active users. Facebook provides limited access to data via its public API, which includes the stream of user status updates and page status updates as they are posted to Facebook, though this only includes content from profiles which have their privacy setting set to 'public.' However, even acquiring access to this feed is limited to a restricted set of media publishers and requires approval from Facebook. Facebook data are not publicly available and cannot be purchased. However, Facebook Research has been relatively open to collaborations with scholars in academia. Collaborations have taken different forms, from data sharing regulated by non-disclosure agreements (NDAs), to visiting positions at the Facebook headquarters. See for instance the Facebook Academic programs[7]. Researchers at Facebook have shown preliminary results about coordinated migrations between cities, using information about users' reported "hometown" and "current city."[8] More recently, Facebook Data scientist Aude Hofleitner and colleagues have been investigating the possibility of using IP address geolocation of Facebook users to understand travel mobility and migration movements[9].

LinkedIn

LinkedIn is a career development and professional social networking service through which users can upload biographical information about themselves regarding work history, skills and expertise, and professional connections. The site was founded in 2002 and has since accumulated more than 259 million active users. Though originally founded in the US, the service has expanded globally to over 200 combined countries

---

[5] An application-programming interface (API) is a set of programming instructions and standards for accessing, for example, a database, via a Web-based software application.
[6] https://qnip.com
[7] https://research.facebook.com/programs/
[8] https://www.facebook.com/notes/facebook-data-science/coordinated-migration/10151930946453859
[9] Aude Hofleitner, "Understanding Travel and Migration Movements at a Global Scale", Brownbag presentation, Department of Demography, UC Berkeley, Nov. 4, 2015.

and territories. The service itself allows users to upload their resumes, curricula vitae, and any other relevant information.

Researchers can use the data from this service to analyse retrospective employment histories alongside other demographic data (e.g. education, country and city of residence, volunteer work, etc.). Very limited data can be accessed via the LinkedIn API. Research on migration using LinkedIn data mainly benefit from direct collaborations with scientists working at the company.

Google Latitude

Google Latitude is a feature of Google Maps, which tracks the location of users and shares this information with people chosen by the user. It was announced in 2013 that Google intended to shut down Latitude that year, so the feature no longer exists. Linking to the user's Google account, the feature would map the user's cell location. The level of geographic detail of the information could be controlled and customized by the user ranging anywhere from the exact location to just the city. Users could also enable privacy settings, which would turn off location services or allow the user to manually input locations. The service used geolocation API, user input, and automated location detection to map user locations, drawing on cellular positioning, Wi-Fi positioning, and GPS.

MAIN RESULTS FROM NEW DATA SOURCES: A REVIEW OF THE LITERATURE

Call Detail Records

Mobile phones have become one of the most valuable sources of data on geographic mobility in recent years. Reaching a penetration rate of 77% worldwide and 68% in industrializing countries, mobile phones represent one of the most pervasive forms of technology in use by humans in the contemporary world (Blumenstock, 2012). The widespread adoption of mobile phones uniquely situates them as a source of information on human behaviour, allowing researchers to track population density, location, mobility, routine patterns, and, on occasion, basic demographic traits. In one case, Deville et al. (Deville et al., 2014) obtained call records from telecommunications companies in Portugal and France to estimate the population density of each area using the information on cell tower pings. In another study, Bengtsson et al. (Bengtsson et al., 2011) sought to track population movement following disasters. Focusing first on the Haiti earthquake in 2010, the researchers looked at pre- and post-locations and movements in order to later compare these results to a restricted time sample of movements following a cholera outbreak. The general approach is particularly relevant to obtain timely information about sudden increases in migration flows, like in the context of asylum crises.

Call detail records have been shown to be useful in illuminating mobility trends. Although some information may be lost in the process of anonymization to protect user privacy, these data still contain useful information on the time of mobile phone calls and texts as well as the cell tower location associated with the call or text.

Using such information, past projects have studied mobility by inspecting (1) the number of cell towers used, (2) the maximum distance travelled, (3) the radius of gyration[10], (4) and inferred mobility (Phithakkitnukoon et al., 2012; Berlingerio et al., 2013; Becker et al., 2013; Bayir et al., 2009; Calabrese et al., 2010; Blumenstock,

[10] The root mean square distance of the locations visited from the barycentre of the distribution

2012; Csaji et al., 2013; Gonzalez et al., 2008; Dobra et al., 2014; Williams et al., 2014; Phithakkitnukoon et al., 2010; Farrahi and Gatica-Perez, 2010). In doing so, they have leveraged this information to construct origin-destination matrices, capture the breadth of unique location mobility, calculate total mobility according to varying metrics (e.g. total distance travelled, extent of distance travelled away from primary location, etc.), and estimate spatial and temporal dimensions of routine behaviours. Despite the shortcomings of anonymization, mobile phone records can address some of the limitations of traditional survey or census data used to study mobility. Traditional data are often difficult and costly to collect, which commonly limits data sets in terms of sample size and temporal scope. In contrast, call detail data have the potential to capture the entire population of mobile phone users, which is very large. They can also offer real-time information on caller locations spanning a longer period of observation. Additionally, they can provide the GPS coordinates of cell tower locations, which are often more accurate than the reported locations provided by surveys. As a potential method of addressing common concerns regarding sampling hard to reach populations, mobile phone and other big data may better capture the behaviours of these groups, including undocumented immigrants, temporary workers, circular migrants, etc. (Neubauer et al., 2015). For instance, Lu et al. (Lu et al., 2013) obtained call detail data on a large, random sample of mobile phone users in Cote d'Ivoire, which may traditionally lack granular data on the mobility of its population. Applying a Markov Chain-based estimation algorithm, the researchers found that they could estimate a potential predictability in user mobility as high as 88%.

Mobile Phone Records with Supplementary Information Sources

Though most call detail record data contain little to no demographic information on users, mobile phones are becoming increasingly sophisticated, integrating a multitude of applications and services that produce other valuable data like automatic location recording, details on proximity to services, and motion sensors, all of which can be leveraged to produce more multidimensional information on user behaviour (Ferrari and Mamei, 2011; Farrahi and Gatica-Perez, 2010). Researchers have also sampled smaller populations, supplementing mobile phone data with daily activity logs to increase data accuracy (Andrew et al., 2013; Ferrari et al., 2011a).

Some supplementary apps or features are compatible with mobile phones, though are not associated with the information obtained from call records. For example, Google Latitude and similar mobile phone applications allow users to choose more active tracking of their behaviour, integrating with multiple platforms within the phone to switch between using Wi-Fi, GPS, and GSM localizations to track user movements. Unlike solely capturing locations through the readings of GPS coordinates, this method significantly reduces the strain on the battery power of mobile devices and can potentially yield more multidimensional information on the users themselves. Additionally, Bluetooth capabilities within phones also have the potential to be leveraged for information on mobility (Versichele et al., 2014; Delafontaine et al., 2012; Laharotte et al; Yoshimura et al, 2014). A study by Delafontaine et al. (Delafontaine et al., 2012) set up Bluetooth receivers throughout a convention space, which recorded the unique footprints of nearby Bluetooth devices as they passed within the range of their signals. This approach, however, is currently limited to small spaces in which receivers must already be in place before the mobility event occurs.

Some researchers have begun to collect or link survey data with call records, though these studies are still in the minority due to privacy issues, protection of human subjects, funding, and logistical limitations. However, some efforts have been made to acquire more information on mobile phone users. For example, in order to obtain more granular information on their subjects, Blumenstock and Fratamico (Blumenstock and Fratamico, 2013) obtained call detail data from a telecommunications company in Rwanda and randomly called a subsample from this group in order to obtain more information about their demographic characteristics. As a whole, integrating mobile phone call detail data with other sources of data that bring to bear some specific

measures of demographic traits, would greatly improve the potential utility of this data source.

Despite the advantages offered by call detail record data, the use of this information still comes with its own array of limitations and shortcomings. One of the most common critiques levelled against using call record data aims at the unreliability of cell tower information (Bayir et al., 2009; Csaji et al., 2013; Bayir et al., 2010). Due to the load balancing features of cell tower allocation, the nearest cell tower does not always precisely capture users' locations. Though some researchers have attempted to mitigate these concerns by clustering around the oscillation between towers, the solution is imperfect (Bayir et al., 2009). Additional concerns question the computational feasibility of increasing call detail record data sample sizes, issues of selection bias regarding mobile phone usage and ownership, and data paucity or inconsistency across the rural to urban landscape (Bayir et al., 2009; Blumenstock and Fratamico, 2013; Blumenstock, 2012). Lastly, another level of criticism surrounding the use of mobile phone data for research is ethical in nature (Shilton, 2009; Friedland and Sommer, 2010). Though the data are anonymized in order to protect users' privacy, anonymization itself does not necessarily address all the issues related to informed consent and privacy. As debates continue surrounding surveillance and privacy, these ethical considerations may become increasingly important. For example de Montjoye et al. (de Montjoye et al., 2013) found that just 4 spatio-temporal points of observation were sufficient to uniquely identify 95% of people in a mobility database of 1.5 million people. Properly anonymizing the data while providing the most possible information to the research community is a subject of ongoing research and debate.

Geotagged Social Media Data

Beyond the potential data to be gathered from mobile phones, technological advances in Internet access and usage have opened up unprecedented channels through which data can be collected from its users. Social media data, in particular, have been the focus of several studies on how best to leverage this kind of information. Defined as computer- or phone- mediated tools through which people create, share, and exchange information in virtual communities, social media is a broad categorization to classify the usage of popular sites like Facebook, Twitter, Instagram, LinkedIn, YouTube, Foursquare, Tumblr, Flickr, Reddit, and so on. Other services are harder to classify but are similar to social media, like email, which is used as a communication tool that also yields geotagged information via unique IP addresses. Given the multifaceted nature of shared information from these sites, the potential of social-media-generated data to illuminate aspects of human life ranging from health to voting behaviour to social norm adherence can, in many respects, be huge.

In the context of human mobility and migration, efforts to collect and model these patterns via social media data have broadly taken two forms: first relies directly on geotagged information generated from these sites and the second infers locations and mobility patterns without explicitly reported location information. Existing studies have attempted to study mobility in various ways. Having access to geotagged information, researchers have tried to extract mobility patterns for varying temporal and spatial scopes by calculating the density of tweets sent from various locales within a city, from the location of landmarks featured in photo streams on Flickr, and from the reported location of users across administrative boundaries of cities, states, and countries (Hawelka et al., 2014; Compton et al.,2014; Lenormand et al., 2014; Jurdak et al., 2014; Liu et al., 2014; Zagheni et al., 2014; De Choudhury et al., 2010; Zheng et al., 2012; Yin et al., 2011; Naaman, 2011; Ferrari et al., 2011b; Grinberg et al., 2013).

The benefits of using social media data can be vast although currently untapped given data access and methodological restrictions. Unlike traditional sources of mobility data, social media data offer real-time information on users' locations, and depending

on the social media source, can contain information on social networks and other aspects of social behaviour not commonly captured by survey data (e.g. user attitudes towards popular culture, current events, local job market, policy changes, etc.). In contrast to mobile phone data, some social media data offer for more information on specific users' demographic traits and behaviours. In a study by Lenormand et al. (Lenormand et al., 2014), Twitter data and mobile phone call record data were compared to census data in Spain to crosscheck results from each of these sources. Though each data source offers varying degrees of granularity in terms of the information it can yield, the results showed consistency across various types of traditional and non-traditional data, bolstering claims concerning the reliability of using social media data for demographic research.

Non-Geotagged Social Media Data

Despite the exciting potential of using social media data for inferring information on human behaviour, this area of research is still in its infancy and, therefore, subject to several limitations. In terms of geo-location, such data only accounts for a small portion of social media information. Hawelka et al. (Hawelka et al., 2014) estimate that only around 1% of the total Twitter feed is geotagged, though new ways of automatically recording locations are currently being developed and adopted to increase tracking. When this information is not present, researchers must infer users' locations and mobility patterns, which can be especially problematic for users who travel often, who live in multiple spaces, or who tweet selectively from novel locations.

In thinking through how to model location when no geotagged information is available, researchers have attempted to triangulate several types of information in the tweet content itself (Kinsella et al., 2011; Stefanidis et al., 2013; O'Hare and Murdock, 2013; Ikawa et al., 2012; Graham et al., 2014; Gelernter and Mushegian, 2011; Ryoo and Moon, 2014). Cheng, Caverlee, and Lee (Cheng et al., 2010) attempted to predict users' locations based on their use of place-specific language in Twitter posts. Assuming that users will more frequently tweet about certain topics (e.g. sports teams, local events, political figures) and/or use specific expressions, idioms, or slang specific to particular locales (e.g. 'Howdy,' 'gnarly,' 'soda pop'), the researchers attempted to infer users' locations based on the frequency of their use of place-based language. Other researchers have attempted to locate users through their participation in locatable social networks (Li et al., 2012; Chandra et al., 2011; Jurgens, 2013).

In general, however, the data themselves are inherently noisy, especially when researchers attempt to scrape information from their content. Users may use shorthand or slang, the content may contain information that spans multiple locations (e.g. rooting for sports teams or political candidates), or it may include words or expressions that vary in meaning depending on context. Such hurdles are common for those seeking to use natural language processing tools to gather information from social media data. Beyond these issues, there remains the concern of selection bias since it is commonly asserted that usage of social media platforms cater especially to a younger and more urban population. Work by Zagheni and Weber (Zagheni and Weber, 2015) is presently attempting to address some of these concerns regarding selection bias in social media data.

Web Searches, and other Internet and Mobile Device Data

Though mobile call record data and social media data are two of the most active mediums through which big data has been collected and studied, other sources still remain that do not fall directly into either of those categories. Alongside the potential for mobile phone-generated data, Internet data expands beyond just that offered by social media platforms. Another way researchers have attempted to model user behaviour is through analysing user Web searches (Choi and Varian, 2012; Goel et al.,

2010; Vosen and Schmidt, 2011; D'Amuri and Marcucci, 2010; Ripberger, 2011; Yang et al., 2010; Tefft, 2011; Ayers et al., 2011; Reilly et al., 2012; Mccallum and Bury, 2013; Pelc, 2013; Yuan et al., 2013). For example, in a study by Askitas and Zimmerman (Askitas and Zimmermann, 2009), the authors pulled and analysed Google search keywords regarding job hunts in order to produce more timely estimates of monthly unemployment rates in Germany. By creating a catalogue of possible search terms that users would use to find work in the event of a layoff or prolonged period of unemployment, researchers predicted the unemployment rate and compared this to official estimates from the state. Other scholars have used the biographical information of users from résumés or curricula vitae in order to obtain time-identified demographic information. In one case, State et al. (State et al., 2014) estimated relative migration flows of professionals to the United States using job history information reported on users' LinkedIn profiles. By using educational and employment histories, they could evaluate the residential history of users, starting from the early 1990s. In a different study, Hadiji et al. (Hadiji et al., 2013) used bibliographic information from the DBLP computer science bibliography to study migrations of academic researchers. Since journal publications identify the associated institution belonging to each respective author, the study observed the movement of researchers from one institution to another to study the aggregate mobility of academic researchers. Beyond mobile phone and Internet data, there are other sources of big data that can be used to study mobility, though suggestions of what these could be and what information they could potentially yield is more limited. However, one study utilizes time series data on global light pollution from DMSP OLS images as a proxy for relative population density (Bharti et al., 2011). By comparing these images across seasons, the researchers argued that seasonal fluctuations in population density as measured by changes in light pollution was associated with the seasonal fluctuations of measles cases. For aggregate, de-identified data such as this, it is much more difficult to make specific claims about population processes.

Table 1 offers a summary of key features for the main new data sources used in the recent literature about migrations and geographic mobility. More specifically, the table lists the data sources, the type of access to the data that researchers can potentially have, the cost, the geographical coverage, the key indicators that can be extracted from the data, and the relevant literature that has either relied on the data source or could be hypothetically applied to the data. Table 2 provides a summary of social media penetration rates across European countries. Although there may be differences across platforms within countries, in general in the countries with lower penetration rates the population of social media users may be more selected for a number of socio-demographic characteristics. Thus estimates of geographic mobility for those countries may come with higher levels of uncertainty. In the appendix we provide some statistics regarding penetration rates broken down by type of social media (Facebook, LinkedIn and Twitter).

Table 1: Summary of key features for the main new data sources used in the recent literature about migrations and geographic mobility.

Data Source	Access Level	Cost	Geographic coverage	Indicators	Relevant Studies
Mobile Phone					
Data 4 Development	Application and research proposal required	Free	Senegal	Unique individual IDs, cell tower location pings	Becker et al. (2013); Berlingerio et al. (2013); Blumenstock (2012); Blumenstock and Fratamico (2013); Calabrese et al. (2010); Candia et al. (2008); Csáji et al.(2013); Dobra, Williams, and Eagle (2014); Deville et al. (2014); Gonzalez, Hidalgo, and Barabasi (2008); Iqbal et al. (2014); Lenormand et al. (2014); Lu et al. (2013); Bengtsson et al. (2011); Phithakkitnukoon (2010); Williams et al. (2014)
Reality Commons Project	Registration required, public access	Free	Boston, Chicago, anonymized North American city	Unique individual IDs, location, basic socio-demographic traits, social relationships	Bayir, Demirbas, and Eagle (2009); Farrahi and Gatica-Perez (2010)
Internet/Social Media					
Twitter	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; From ~ $500+ for historical data	Global	Unique individual IDs, tweet content, some geotagged locations	Chandra, Khan, and Muhaya (2011); Cheng, Caverlee, and Lee (2010); Compton, Jurgens, and Allen (2014); Ferrari, Rosi, Mamei, and Zambonelli (2011); Graham, Hale, and Gaffney (2014); Grinberg et al. (2013); Hawelka et al. (2014); Ikawa, Enoki, and Tatsubori (2012); Kinsella, Murdock, and OHare (2011); Lenormand et al. (2014); Naaman (2011); Neubauer (2015); Ryoo and Moon (2014); Yin et al. (2014); Zagheni et al. (2014)
Foursquare	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; From ~ $500+ for historical data	Global	Unique individual IDs, check-in location	Grinberg et al. (2013); Hawelka et al. (2014)
Tumblr	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; From ~ $500+ for historical data	Global	Unique individual IDs, microblog content, user likes	Chandra, Khan, and Muhaya (2011); Cheng, Caverlee, and Lee (2010); Graham, Hale, and Gaffney (2014); Ikawa, Enoki, and Tatsubori (2012)
WordPress	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; From ~ $500+ for historical data	Global	Unique individual IDs, blog content, selective geotagged location, other user metadata	Chandra, Khan, and Muhaya (2011); Cheng, Caverlee, and Lee (2010); Hawelka et al. (2014); Graham, Hale, and Gaffney (2014); Ikawa, Enoki, and Tatsubori (2012); Naaman (2011); Neubauer (2015); Yin et al. (2014)
Disqus	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; From ~ $500+ for historical data	Global	Unique individual IDs, comment content, upvote and downvote activity	Chandra, Khan, and Muhaya (2011); Cheng, Caverlee, and Lee (2010); Graham, Hale, and Gaffney (2014); Ikawa, Enoki, and Tatsubori (2012)
VK	Public API	Free for current streaming from API, limits based on user	Russia and Eastern Europe	Unique individual IDs, profile and activity information, social	Ferrari, Rosi, Mamei, and Zambonelli (2011); Cheng, Caverlee, and Lee (2010); Jurgens (2013)

		authorization, subject to rate limits		networks	
Flickr	Public API	Free for current streaming from API, subject to rate limits	Global	Photo and text information, selective geotagged location	De Choudhury et al. (2010); Hawelka et al. (2014); Naaman (2011); Neubauer (2015); Yin et al. (2014); Zheng, Zha, and Chua (2012)
Panoramio	Public API	Free for current streaming from API, subject to rate limits	Global	Photo and text information, selective geotagged location	De Choudhury et al. (2010); Hawelka et al. (2014); Naaman (2011); Neubauer (2015); Yin et al. (2014); Zheng, Zha, and Chua (2012)
Instagram	Public API	Free for current streaming from API, subject to rate limits	Global	Unique individual IDs, photo and text information, selective geotagged location	De Choudhury et al. (2010); Hawelka et al. (2014); Naaman (2011); Neubauer (2015); Yin et al. (2014); Zheng, Zha, and Chua (2012)
Reddit	Public API	Free for current streaming from API, subject to rate limits	Global	Unique individual IDs, user activity, user account preferences, site content	Chandra, Khan, and Muhaya (2011); Cheng, Caverlee, and Lee (2010); Graham, Hale, and Gaffney (2014); Ikawa, Enoki, and Tatsubori (2012)
Yelp	Public API, Dataset Challenge	Free for current streaming from API, subject to rate limits; Free access to dataset challenge data upon approval	Global	Unique individual IDs, user activity, site content	Chandra, Khan, and Muhaya (2011); Cheng, Caverlee, and Lee (2010); Naaman (2011); Neubauer (2015); Yin et al. (2014)
Google Trends	Public Search	Free	Global	Search activity, aggregated trends	Askitas and Zimmerman (2009); D'Amuri and Marcucci,(2010); Goel et al. (2010); Vosen and Schmidt (2011)
LinkedIn	Economic Graph Challenge; collaborations With LinkedIn researchers	Not available for purchase	Global	Professional histories	State et al. (2014)
Yahoo!	Collaborations with Yahoo! Researchers	Not available for purchase	Global	Geotagged location	Zagheni and Weber (2012)
Facebook	Academic Program	Not available for purchase	Global (no China)	Geotagged location; self-reported demographic information	Hofleitner et al (2013, http://on.fb.me/1fDRVZ6)

Table 2: Active social media penetration rate in European countries as of January 2014. The values indicate the percentage of Internet users who logged on to social media services at least once per month. Source: www.statista.com.

Country	Penetration rate	Country	Penetration rate
Iceland	70%	Norway	64%
Malta	58%	Denmark	58%
Sweden	57%	UK	57%
Serbia	52%	Netherlands	52%
Belgium	52%	Macedonia	51%
Luxembourg	50%	Ireland	50%
Montenegro	49%	Cyprus	48%
Hungary	48%	Portugal	48%
Finland	46%	Albania	45%
Bulgaria	43%	Estonia	43%
Switzerland	43%	France	42%
Italy	42%	Spain	41%
Czech Republic	41%	Greece	41%
Croatia	40%	Slovenia	40%
Slovakia	40%	Bosnia & Herzegovina	40%
Austria	39%	Lithuania	35%
Germany	35%	Russia	33%
Romania	32%	Poland	31%
Ukraine	27%	Latvia	23%
Belarus	23%	Moldova	10%

METHODOLOGICAL ASPECTS RELATED TO THE ANALYSIS OF NEW AND NON-REPRESENTATIVE DATA SOURCES

In the previous sections, we discussed how Internet, social media and mobile phone data hold many promises for research about migration processes. We showed that there exists a considerable body of literature about using new data sources to estimate geographic mobility and internal and international migrations. In this section, we would like to discuss some of the methodological issues related to the analysis of so-called "digital breadcrumbs", that is data that were generated for purposes other than research, but that can be leveraged for research purposes. More specifically, the key issue is selection bias: users of a given website or mobile phone provider are not representative of the underlying general population. Here we present some major approaches that have been developed to address this central problem.

Validation when official statistics exist

In a number of circumstances, estimates for the same quantity of interest are available from both new and non-representative sources and traditional sources like sample surveys. The typical challenge that researchers face is related to a trade-off between variance and bias. Estimates from big social data are expected to have low variance because the sample size is large. However, the bias due to selection may be substantial. In the context of probabilistic sample surveys or other official statistics, the bias is expected to be small, because of the nature of the research design. However, the variance may be quite big as relatively rare events like migrations require large samples.

Two main approaches have been used in order to correct for non-representativeness of digital records and to produce population-level estimates of quantities of interest. The first method is based on a calibration method that relies on a parsimonious model of the relationship between the quantity of interest and penetration rates of the media under consideration. The second approach relies on a combination of post-stratification and a relatively complex prediction model.

The calibration approach in the context of Internet data has been introduced by Zagheni and Weber (2012) in order to estimate age- and sex-specific profiles of international emigration using geo-located Yahoo! e-mail data. The underlying idea is that there is a relationship between the size of selection bias in each age- and sex-group and their respective use of a specific website or web service like Yahoo! In other words, if everybody in a specific demographic group uses the web service, then there should not be any selection bias. If only a fraction of the population in a given demographic group uses the web services, then it is likely that users are not representative of the entire population. For instance, they may be more highly educated or have higher income than the general population. In order to correct for selection bias, Zagheni and Weber (2012) developed a parametric model where a correction factor for each demographic group is expressed as a function of penetration rates for the respective group, and of a parameter. If the penetration rate is less than 1, then the correction factor is less than 1. This means that that quantity of interest would be corrected downwards, reflecting the underlying assumption that in demographic groups with lower penetration rates, users are likely to be more highly educated, have more income and be more mobile than the general population. The extent to which correction factors vary with penetration rates is determined by a parameter that can be estimated with statistical methods, for instance by finding the value that minimizes deviations between corrected values for the quantity of interest and respective official statistics. Typically, 'ground truth' data may be available only for a subset of regions of interest: the estimated parameter would thus be used to generate corrected estimates from digital records, in regions where no official statistics are available.

In the example described above, there are a number of assumptions embedded. Zagheni and Weber (2015) generalized the approach by expressing the quantity of interest for a demographic group yd as a function of the uncorrected estimate from digital records, μd and and the bias:

$$y_d \approx f(\mu_d) + bias_d \quad (1)$$

Then $bias_d$ would then be modelled as a general function of penetration rates and other covariates.

The existing literature indicates that online and mobile phone data contain meaningful information. Statistical models can be used to calibrate different sources in a way that is consistent with existing fragmentary data sources. The level of uncertainty for the estimates would be reflected in confidence intervals that may vary in width depending on the data source used. Appropriate statistical calibration would allow for comparability with existing sources. For example, the definition of migration might be fixed for surveys (e.g. a migrant is someone who has lived for at least x months in a country different from the one of previous residence). When online data are available, the researcher can estimate quantities using different definitions of migration. For instance, the number of migrants can be estimated as a function of the choice of number of x months in the receiving country. This flexibility in defining migration events, given online data, offers opportunities to harmonize data and compare results across countries.

Approaches to address selection bias in absence of 'ground truth' data

In the example that we discussed in the previous section, data about the quantity of interest, say migration rates, can be estimated from geo-located trajectories of a website's users. The sample size may be quite large, but biased. Limited demographic information may or may not be available to calibrate a model against official statistics.

In other situations, like in the case of Web surveys, more detailed demographic information about users of a given platform may be available. That information can be leveraged to generate appropriate weights for the respondents and correct for selection bias. For example, Wang et al. (Wang et al., 2014) showed that, with the proper statistical adjustments, non-representative polls can be used to forecast elections. More specifically, they used a survey on the Xbox gaming platform and adjusted the responses via multilevel regression and post-stratification. Post-stratification is a well-known sampling survey technique to correct for known differences between sample and target populations. The general idea is to partition the population into cells based on combinations of a number of attributes (e.g., age, sex, race, residence). The quantity of interest is then estimated for each cell. Finally, the cell-level estimates are aggregated to population-level estimates by weighting each cell by its relative proportion in the population. A major constraint with post-stratification is that, even with a relatively small number of attributes, the number of cells may be very large, and thus for a number of combinations of attributes there may be very few observations, or none at all. Wang et al. (2013) fit nested multilevel logistic regression models to predict the value of each cell. The underlying idea is that, within a Bayesian framework, estimates for relatively sparse cells can be improved by borrowing strength from cells that have similar attributes, but larger samples.

The approach that we just summarized implies that good-quality demographic attributes for the users are available. That is true for a number of Web surveys, but it is not necessarily the case for a number of other 'digital breadcrumbs', like Twitter data. When little or no demographic attributes are available, and when there is no 'ground truth' data to calibrate a model, then a valuable approach is to choose the model that relies on the most sensible assumptions to reduce bias. For example, Zagheni, Garimella et al. (2014) noted that estimates of migration rates from Twitter data could not be produced for a single point in time, as there was not enough information about the size and direction of the bias. However, using a difference-in-

difference approach, estimates for relative trends could be obtained. The underlying assumption is that the bias may be changing relatively slowly and thus the underlying relationship between the quantity of interest in the biased data set and in the population of interest may be fairly stable over short periods of time. In this context, using a difference-in-difference model results in the bias being cancelled out. Thus estimates of relative trends can be obtained. See Zagheni and Weber (2015) for a formal discussion.

Feasibility and scalability

In the previous sections, we discussed the state-of-the-art about estimation of migrations with traditional and emerging data sources. Although there are a number of methodological issues that have to be accounted for and addressed, the existing body of literature shows that it is feasible to extract relevant information about migration patterns from data sources like mobile phone data, social media data, and other Internet data.

A number of articles published in the literature can be considered feasibility studies for various types of data sources. The main question that arises is whether these studies can be scaled to meet our societies' need for reliable and timely available migration statistics. There are a number of barriers that need to be taken into account.

The first barrier is related to data access. In a number of situations, social media data and other Internet data are not publicly available. Portions of the databases can be accessed via public APIs. For example, 1% of the streaming of Twitter tweets can be accessed via the streaming API. However, larger and historical Twitter data sets have to be purchased via resellers like GNIP. The cost to purchase social media data varies. As a ballpark estimate, a request for historical Twitter data costs around $500 for a coverage of 10 days and up to 1 million tweets. The price may be subject to a number of parameters and may vary, depending on the reseller (e.g., GNIP, Topsy, Datasift) and the volume of data purchased[11].

Access to the live stream of 10% of all the tweets costs in the order of $11,000 per month. This access is guaranteed by a subscription to the so-called "Decahose". In general, the data cost may be relatively small for projects that focus on a specific geographic area or temporal frame, but may be substantially bigger for larger-scale projects, especially if they rely on multiple data sources. For some data sources, like LinkedIn or Facebook, there are no formalized mechanisms to purchase the data. Thus, in those circumstances data use for research purposes is subject to ad-hoc agreements with the respective companies.

Even when data can be purchased, like in the case of Twitter data, there are a number of limitations related to using and sharing the data. These limitations are detailed in the Terms of Service of the companies involved. For example, Twitter does not allow sharing more than 50,000 "objects", which can be tweets or user profiles. However, researchers can share an unlimited number of object IDs, i.e. numeric tweet IDs or user IDs. These would then have to be "re-hydrated"[12] via the Twitter API. Among other things, this ensures that users can delete their data, which will then disable the re-hydration and remove the user from the analysis. The re-hydration process would be subject to the API's rate limits[13], meaning that it may be a long and cumbersome task. If the dataset is particularly large, then sharing may not be feasible, unless the data is re-purchased, given the legal and technical constraints set out by a number of social media companies.

---

[11] Personal communication from a colleague who purchased the data.

[12] Re-hydration indicates the process by which the data for specific user IDs is accessed again via the API, so that only those tweets that are public available at that time, and that had not been deleted by the user.

[13] https://dev.twitter.com/rest/public/rate-limits

Social media and other Internet companies have two main concerns. The first one is that personally identifiable information (e.g., e-mail addresses, phone numbers, IDs) are protected, not shared and used in accordance with the Terms of Service. The second concern is that business sensitive information, like number of users in a particular region, would not be shared. Researchers working with proprietary data should become familiar with the Terms of Service of the specific provider. The Terms of Service set out important legally-binding rules regarding how the users' information may be used and shared, as well as limitations regarding data acquisition. For example, although it is technically feasible to collect LinkedIn data (e.g. some information about users' profiles) using various types of web scraping techniques, the user agreement states that the practice is not allowed, unless LinkedIn offers permission.

Although sharing individual-level data may not be feasible in a number of circumstances, sharing aggregate data is typically allowed. As an example, Google does not share individual- level data, but does provide aggregate-level indexes of Web searches that can be leveraged for migration research. These aggregate-level data are publicly available via Google Trends and Google Correlate. For a number of practical applications related to the estimation of migration flows, only aggregate-level information may be needed. Thus, the model of Google Flu Trends, a tool that provides almost real time estimates of flu activity in the US based on search queries, could be potentially adopted to provide provisional estimates of migration trends.

Call detail records (CDRs) present unique challenges in terms of privacy protection and anonymization. This type of data is generally less accessible to the research community. The reason why mobile phone companies are often unwilling to share mobile phone CDRs are diverse and include the fact that these data contain the complete customer database and thus critical market information for the mobile network operators. Furthermore, operators are concerned with the privacy of their customers and in many settings the legal basis for sharing data is either unclear or absent. Thus, In order to access CDRs, it is extremely important to address all these concerns in detail, by working in close collaboration with operators on a case-by-case basis. This could also require the involvement of international organizations, governments, and operator parties in order to establish trust and make a clear case for the use of the data (e.g., how the data will be used to support development, disaster would be available for download.relief, and or to improve jobs).

Although access and analysis of CDRs are limited by the highly sensitive nature of the data, there are two important examples of success in access and use of CDRs. The non-profit organization Flow minder has been successful in accessing and using CDRs for humanitarian and research purposes. The organization has relied on building long-term and mutually beneficial collaborations with operators. The Flow minder's strategy is to use only anonymized data at a resolution level that decreases the sensitivity of the data. The data are also managed in accordance with non-disclosure agreements signed with each operator. In the European context, Telefonica has been a leading partner in research using CDRs.

The Data 4 Development challenge offers an alternative model of collaboration between mobile phone providers and the research community. CDRs are anonymized and organized in a way to minimize the risk of users' identifiability. The data are then shared with applicants that propose relevant projects. The data granularity is lower than the one available to the mobile phone provider. However, when the data are organized in this way, they can be accessed by a large number of researchers and generate substantial innovative methods and results.

Conclusions

In this article, we offered first a systematic review of the literature about measuring internal and international migration with traditional and new data sources. Then we discussed main methodological aspects related to the analysis of new big data sources, which are often non-representative of the underlying population. Finally we evaluated the feasibility and scalability of current approaches. In doing so, we emphasized some of the main technical and legal barriers related to data access, and data sharing, as well as some of the idiosyncrasies of the new data sources.

There are a number of bottlenecks in the analysis on Internet data, social media data and mobile phone data. Some of the bottlenecks are related to legal barriers that result from the Terms of Service of the providers, and the sensitivity of the data in terms of privacy protection and business interests. However, an often-overlooked barrier is related to the rapid proliferation of data types and services that can be used for migration studies. New providers and new services are constantly emerging, with users switching from one mobile phone, social media or email provider to a different one at a rapid pace. Just a few years ago, the advent of Internet data was greeted with enthusiasm as a unifying force. It was seen as new data that was able to cross national borders. Today, more and more services are becoming available, including some that may gain large popularity only in a few countries, but not in others.

The proliferation of services and databases that have information relevant to the study of migration processes is a critical scientific challenge in the context of producing an appropriate infrastructure that generates reliable information about migration patterns over time and across countries. We believe that developing methods to combine existing data sources, both traditional and emerging ones, is key to generating a robust infrastructure that can leverage innovative data sources as they evolve over time.

Is it technically, financially and legally feasible to estimate geographic mobility and migration flows in the European Union? As for most interesting questions, the answer is 'it depends'. First, it depends on the data that one can have access to. Some data sources can be accessed by anyone with the appropriate technical skills (e.g., samples of Twitter tweets); some can be purchased (e.g., historical tweets); some are not for sale and require partnerships with companies (e.g., Yahoo!, Facebook, LinkedIn, and mobile phone providers); some are not shared by companies (Google does not share data, except for some aggregate indexes, like the ones in Google Trends). Second, it depends on the outcome desired. Estimating trends or changes in trends in migration flows is feasible and can be done in a timely manner. Getting accurate and precise estimates for special populations, like refugees, may or may not be feasible depending on the context. Likewise, obtaining estimates of short-term mobility that could be useful to inform, for example, natural experiment related to labour migration policies, is feasible. Obtaining unbiased estimates of short-term mobility from a single, non-representative source would be more difficult. It may be feasible in some circumstances (e.g., when the data set is rich enough for the use of post-stratification techniques), but not in others. Third, it depends on legal obstacles associated to specific cases. Companies may have terms and conditions or non-disclosure agreements for data sharing that may or may not include inconsistencies with the rules governing universities and funding agencies. We have not identified major issues in this area, but each individual collaboration across units would require some careful examination of the terms and conditions in order to resolve any potential lack of consistency.

There is no ideal data set that fits all needs. To estimate professional migrations or other labour market indicators for a specific segment of the workforce, LinkedIn is probably the best data source available. However, if the goal is to obtain information on low-skilled labour migration, then LinkedIn is not appropriate.

New and traditional data sources do not substitute for each other, they complement each other. There are three main aspects related to combining new and traditional data sources that we would like to emphasize. First, traditional data sources have a number of drawbacks, but without a benchmark it is difficult to assess the validity of new sources and build trust in new and innovative approaches. Second, the proliferation of data sources comes with a potentially large number of data sets that provide complementary information across countries, over time and at different levels of granularity. These data sets may vary in size and their populations may be selected in different ways. Borrowing strength from a number of data sources is key to generate the most reliable and comprehensive estimates. Third, new data sources are highly dynamic and compositional changes in the user base may change fairly rapidly. Combining data sources is key to produce an infrastructure that is robust to unanticipated changes in the use of technology. Building that infrastructure would be a gradual and incremental process where increasing data production and access, together with the development of methods, would sustain each other.

We believe that Bayesian statistical models for migration count data hold the promise of addressing the issue of unifying traditional and emerging data sources. Recently Raymer et al. (2013) developed frameworks for modelling international migration flows in the context of un-harmonized migration data in Europe. That is an example of how Bayesian methods can be used effectively to combine different migration data in a consistent way. Various sources of information can be incorporated into the estimation of the true flows as prior probabilities in a hierarchical Bayesian model. Although there is no known example of this type of study in the context of big data and migration processes, we believe that it is a promising approach and we will work on developing a framework to incorporate traditional and new data sources for migration within a Bayesian model that can be easily adapted to combine data from a range of sources and control for a variety of measuring issues (including those that arise from big data sources).

REFERENCES

Abel, G. J. (2010). Estimation of international migration flow tables in europe. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173(4):797–825.

Abel, G. J. (2013). Estimating global migration flow tables using place of birth data. Demographic Research, 28:505–546.

Abel, G. J. (2015). Estimates of global bilateral migration flows by gender between 1960 and 2010. Technical report, Vienna Institute of Demography Working Papers.

Abel, G. J. and Sander, N. (2014). Quantifying global international migration flows. Science, 343(6178):1520–1522.

Andrew, A. H., Eustice, K., and Hickl, A. (2013). Using location lifelogs to make meaning of food and physical activity behaviors. In Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on, pages 408–411. IEEE.

Askitas, N. and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. German Council for Social and Economic Data (RatSWD) Research Notes, (41).

Ayers, J. W., Ribisl, K., and Brownstein, J. S. (2011). Using search query surveillance to monitor tax avoidance and smoking cessation following the united states' 2009 "schip" cigarette tax increase. PLoS One, 6(3):e16777.

Barnes, W. (2008). Improving migrant participation in the Labour Force Survey: A review of existing practices in European Union member states. Survey Methodology Bulletin, 63:25–38.

Bayir, M. A., Demirbas, M., and Eagle, N. (2009). Discovering spatiotemporal mobility profiles of cellphone users. In World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a, pages 1–9. IEEE.

Bayir, M. A., Demirbas, M., and Eagle, N. (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. Pervasive and Mobile Computing, 6(4):435–454.

Becker, R., C´aceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. Communications of the ACM, 56(1):74–82.

Bell, M., Charles-Edwards, E., Kupiszewska, D., Kupiszewski, M., Stillwell, J., and Zhu, Y. (2015). Internal migration data around the world: Assessing contemporary practice. Population, Space and Place, 21(1):1–17.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., and Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. PLoS medicine, 8(8):e1001083.

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., and Sbodio, M. L. (2013). Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In Machine learning and knowledge discovery in databases, pages 663–666. Springer.

Bharti, N., Tatem, A. J., Ferrari, M. J., Grais, R. F., Djibo, A., and Grenfell, B. T. (2011). Explaining seasonal fluctuations of measles in niger using nighttime lights imagery.Science, 334(6061):1424–1427.

Bilsborrow, R. E., Hugo, G., Oberai, A. S., and Zlotnik, H. (1997). International migration statistics: Guidelines for improving data collection systems. International Labour Organization.

Blumenstock, J. and Fratamico, L. (2013). Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data. In Proceedings of the 4th Annual Symposium on Computing for Development, page 11. ACM.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. Information Technology for Development, 18(2):107–125.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L., and Ratti, C. (2010). The geography of taste: analyzing cell-phone mobility and social events. In Pervasive computing, pages 22–37. Springer.

Chandra, S., Khan, L., and Muhaya, F. B. (2011). Estimating twitter user location using social interactions–a content based approach. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pages 838–843. IEEE.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 759–768. ACM.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. Economic Record, 88(s1):2–9.

Coleman, D. (2013). The twilight of the census. PoPulation and develoPment review, 38(s1):334–351.

Compton, R., Jurgens, D., and Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In Big Data (Big Data), 2014 IEEE International Conference on, pages 393–401. IEEE.

Csaji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., and Blondel, V. D. (2013). Exploring the mobility of mobile phone users. Physica A: Statistical Mechanics and its Applications, 392(6):1459–1473.

D'Amuri, F. and Marcucci, J. (2010). 'google it!'forecasting the us unemployment rate with a google job search index.

De Beer, J., Raymer, J., Van der Erf, R., and Van Wissen, L. (2010). Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. European Journal of Population/Revue européenne de Démographie, 26(4):459–481.

De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., and Yu, C. (2010). Automatic construction of travel itineraries using social breadcrumbs. In Proceedings of the 21st ACM conference on Hypertext and hypermedia, pages 35–44. ACM.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3.

Delafontaine, M., Versichele, M., Neutens, T., and Van de Weghe, N. (2012). Analysing spatiotemporal sequences in bluetooth tracking data. Applied Geography, 34:659–668.

Dennett, A., Wilson, A., et al. (2013). A multi-level spatial interaction modelling framework for estimating inter-regional migration in europe. Environment and Planning A, 45(6):1491–1507.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45):15888–15893.

Dobra, A., Williams, N. E., and Eagle, N. (2014). Spatiotemporal detection of unusual human population behavior using mobile phone data. arXiv preprint arXiv:1411.6179.

Farrahi, K. and Gatica-Perez, D. (2010). Probabilistic mining of socio-geographic routines from mobile phone data. Selected Topics in Signal Processing, IEEE Journal of, 4(4):746–755.

Ferrari, L. and Mamei, M. (2011). Discovering daily routines from google latitude with topic models. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pages 432–437. IEEE.

Ferrari, L., Mamei, M., and Zambonelli, F. (2011a). All-about digital diaries.

Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. (2011b). Extracting urban patterns from location-based social networks. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pages 9–16. ACM.

Friedland, G. and Sommer, R. (2010). Cybercasing the joint: On the privacy implications of geo-tagging. In HotSec.

Gelernter, J. and Mushegian, N. (2011). Geo-parsing messages from microtext. Transactions in GIS, 15(6):753–773.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with web search. Proceedings of the National Academy of Sciences, 107(41):17486–17490.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. Nature, 453(7196):779–782.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? geolocation and language identification in twitter. The Professional Geographer, 66(4):568–578.

Grinberg, N., Naaman, M., Shaw, B., and Lotan, G. (2013). Extracting diurnal patterns of real world activity from social media. In ICWSM.

Groenewold, G. and Bilsborrow, R. (2008). 14 design of samples for international migration surveys: Methodological considerations and lessons learned from a multi-country study in Africa and Europe. International migration in Europe, page 293.

Hadiji, F., Kersting, K., Bauckhage, C., and Ahmadi, B. (2013). Geodblp: Geo-tagging dblp for mining the sociology of computer science. arXiv preprint arXiv:1304.7984.

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

Ikawa, Y., Enoki, M., and Tatsubori, M. (2012). Location inference using microblog messages. In Proceedings of the 21st international conference companion on World Wide Web, pages 687–690. ACM.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2014). Understanding human mobility from twitter. arXiv preprint arXiv:1412.2154.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. In ICWSM.

Kinsella, S., Murdock, V., and O'Hare, N. (2011). I'm eating a sandwich in Glasgow: modeling locations with tweets. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, pages 61–68. ACM.

Kupiszewska, D. and Nowok, B. (2008). Comparability of statistics on international migration flows in the European union. In Raymer J., Willekens F. (eds.), International Migration in Europe: Data, Models and Estimates. Wiley, London.

Kupiszewska, D. and Wiśniowski, A. (2009). Availability of statistical data on migration and migrant population and potential supplementary sources for data estimation. mimosa deliverable 9.1 a report.

Laharotte, P.-A., Billot, R., Come, E., Oukhellou, L., Nantes, A., and El Faouzi, N.-E Spatiotemporal analysis of bluetooth data: Application to a large urban network.

Lenormand, M., Tugores, A., Colet, P., and Ramasco, J. J. (2014). Tweets on the road. PloS one, 9(8):e105407.

Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1023–1031. ACM.

Liu, J., Zhao, K., Khan, S., Cameron, M., and Jurdak, R. (2014). Multi-scale population and mobility estimation with geotagged tweets. arXiv preprint arXiv:1412.0327.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., and Bengtsson, L. (2013). Approaching the limit of predictability in human mobility. Scientific reports, 3.

Mccallum, M. L. and Bury, G. W. (2013). Google search patterns suggest declining interest in the environment. Biodiversity and conservation, 22(6-7):1355–1367.

Naaman, M. (2011). Geographic information from georeferenced social media data. SIGSPATIAL Special, 3(2):54–61.

Nam, C., Serow, W., and Sly, D. (1990). International Handbook on Internal Migration. Greenwood Press: New York.

Neubauer, G., Huber, H., Vogl, A., Jager, B., Preinerstorfer, A., Schirnhofer, S., Schimak, G., and Havlik, D. (2015). On the volume of geo-referenced tweets and their relationship to events relevant for migration tracking. In Environmental Software Systems. Infrastructures, Services and Applications, pages 520–530. Springer.

O'Hare, N. and Murdock, V. (2013). Modeling locations with social media. Information Retrieval, 16(1):30–62.

Pelc, K. J. (2013). Googling the wto: what search-engine data tell us about the political economy of institutions. International Organization, 67(03):629–655.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., and Ratti, C. (2010). Activity-aware map: Identifying human daily activity pattern using mobile phone data. In Human Behavior Understanding, pages 14–25. Springer.

Phithakkitnukoon, S., Smoreda, Z., and Olivier, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. PloS one, 7(6):e39253.

Poulain, M., Perrin, N., and A, S. (2006). THESIM: Towards Harmonised European Statistics on International Migration. UCL-Presses Universitaires de Louvain, Louvain-La-Neuve.

Raymer, J. (2007). The estimation of international migration flows: A general technique focused on the origin-destination association structure. Environment and Planning A, 39(4):985.

Raymer, J., Rees, P., Blake, A., Boden, P., Brown, J., Disney, G., Lomax, N., Norman, P., and Stillwell, J. (2012). Conceptual Framework for UK Population and Migration Statistics. Office for National Statistics, United Kingdom.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., and Bijak, J. (2013). Integrated modeling of European migration. Journal of the American Statistical Association, 108(503):801–819.

Rees, P., Bell, M., Duke-Williams, O., and Blake, M. (2000). Problems and solutions in the measurement of migration intensities: Australia and Britain compared. Population Studies, 54(2):207–222.

Rees, P., Durham, H., and Kupiszewski, M. (1996). Internal migration and regional population dynamics in Europe: United kingdom case study. WORKING PAPER-SCHOOL OF GEOGRAPHY UNIVERSITY OF LEEDS.

Rees, P. and Kupiszewski, M. (1999). Internal migration: what data are available in Europe? Journal of Official Statistics, 15(4):551.

Reilly, S., Richey, S., and Taylor, J. B. (2012). Using google search data for state politics research an empirical validity test using roll-off data. State Politics & Policy Quarterly, 12(2):146–159.

Ripberger, J. T. (2011). Capturing curiosity: Using internet search trends to measure public attentiveness. Policy Studies Journal, 39(2):239–259.

Ryoo, K. and Moon, S. (2014). Inferring twitter user locations with 10 km accuracy. In Proceedings of the companion publication of the 23rd international conference on World wide web companion, pages 643–648. International World Wide Web Conferences Steering Committee.

Shilton, K. (2009). Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. Communications of the ACM, 52(11):48–53.

Singleton, A., Lenoël, A., and Gora, O. (2010). Country Report United Kingdom. Promoting Comparative Quantitative Research in the Field of Migration and Integration in Europe (PROMINSTAT).

State, B., Rodriguez, M., Helbing, D., Zagheni, E., et al. (2014). Migration of professionals to the us. In Social Informatics, pages 531–543. Springer.

Stefanidis, A., Crooks, A., and Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. GeoJournal, 78(2):319–338.

Stillwell, J. (2005). Inter-regional migration modelling: A review and assessment. In 45th Congress of the European Regional Science Association. Vrije Universiteit Amsterdam, The Netherlands, 23-27 August 2005.

Stillwell, J., Daras, K., Bell, M., and Lomax, N. (2014). The image studio: A tool for internal migration analysis and modelling. Applied Spatial Analysis and Policy, 7(1):5–23.

Tatem, A., Sorichetta, A., Ruktanonchai, N., Bird, T., Tejedor, N., Strano, E., and Viana, M. (2015). Defining regional blocks for malaria elimination. Mid-project progress report for grant OPP1115575 with the Bill and Melinda Gates Foundation.

Tefft, N. (2011). Insights on unemployment, unemployment insurance, and mental health.Journal of Health Economics, 30(2):258–264.

Border Force (2015). Exit checks fact sheet. Policy Paper. Border Force, Home Office.United Kingdom.

United Nations (1978). Statistics of Internal Migration: A Technical Report. Studies in Methods, Series F23, ST/ESA/STAT/SER F/23. Department of International Economic and Social Affairs, UN, New York.

United Nations (2000). World population monitoring 1999: population growth, structure and distribution. Population Division, ST/ESA/SER A/183; Department of Economic and Social Affairs, UN: New York.

United Nations (2008). Principles and Recommendations for Population and Housing Censuses, Revision 2. Department of Economic and Social Affairs, Statistics Division, United Nations, New York.

United Nations (2009). World Development Report 2009: Reshaping Economic Geography. The World Bank: Washington DC.

Thomas, M. (2008). Improving migrant participation in the Labour Force Survey: Non-response and attitudes of non-English speaking migrants to participation. Survey Methodology Bulletin, 63:39–51.

van Ostaijen, M., Faber, M., Engbersen, G., and Scholten, P. (2015). Social consequences of cee migration.

Versichele, M., De Groote, L., Bouuaert, M. C., Neutens, T., Moerman, I., and Van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. Tourism Management, 44:67–81.

Vignoli, J. R. and Busso, G. (2009). Migración interna y desarrollo en América Latina entre 1980 y 2005: un estudio comparativo con perspectiva regional basada en siete países, volume 102. United Nations Publications.

Vosen, S. and Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. google trends. Journal of Forecasting, 30(6):565–578.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2014). Forecasting elections with non-representative polls. International Journal of Forecasting.

Willekens, F. (2008). Models of migration: Observations and judgement. International migration in Europe: Data, models and estimates, pages 117–147.

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., and Dobra, A. (2014). Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. arXiv preprint arXiv:1408.5420.

Wiśniowski, A. (2013). Bayesian modelling of international migration with Labour Force Survey data. PhD thesis. Collegium of Economic Analysis, Warsaw School of Economics. Poland.

Yang, A. C., Huang, N. E., Peng, C.-K., and Tsai, S.-J. (2010). Do seasons have an influence on the incidence of depression? the use of an internet search engine query data as a proxy of human affect. PloS one, 5(10):e13728.

Yin, Z., Cao, L., Han, J., Luo, J., and Huang, T. S. (2011). Diversified trajectory pattern ranking in geo-tagged social media. In SDM, pages 980–991. SIAM.

Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J. P., Blat, J., and Sinatra, R. (2014). An analysis of visitors? behavior in The Louvre Museum: a study using Bluetooth data. Environment and Planning B: Planning and Design, 41:1113–1131.

Yuan, Q., Nsoesie, E. O., Lv, B., Peng, G., Chunara, R., and Brownstein, J. S. (2013). Monitoring influenza epidemics in china with search query from baidu. PloS one, 8(5):e64323.

Zagheni, E., Garimella, V. R. K., Weber, I., et al. (2014). Inferring international and internal migration patterns from twitter data. In Proceedings of the companion publication of the 23rd international conference on World wide web companion, pages 439–444. International World Wide Web Conferences Steering Committee.

Zagheni, E. and Weber, I. (2015). Demographic research with non-representative internet data. International Journal of Manpower, 36(1).

Zheng, Y.-T., Zha, Z.-J., and Chua, T.-S. (2012). Mining travel patterns from geotagged photos. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3):56.

Zipf, G. (1946). The p1p2/d hypothesis: on intercity movement of persons. American Sociological Review, 11:677–686.

APPENDIX

Appendix Table 1: Facebook penetration rates (percentage of the population who subscribed to Facebook) in the European Union, by country, as of November 15, 2015. Source: internetworldstats.com.

Country	FB penetration rate	Country	FB penetration rate
Austria	40.8%	Italy	46.1%
Belgium	52.4%	Latvia	32.7%
Bulgaria	44.4%	Lithuania	47.9%
Croatia	42.6%	Luxembourg	49.7%
Cyprus	69.7%	Malta	62.9%
Czech Republic	42.7%	Netherlands	56.2%
Denmark	61.8%	Poland	36.8%
Estonia	44.9%	Portugal	54.0%
Finland	47.5%	Romania	40.8%
France	48.4%	Slovakia	42.4%
Germany	35.7%	Slovenia	41.2%
Greece	44.4%	Spain	47.4%
Hungary	51.8%	Sweden	57.5%
Ireland	56.2%	United Kingdom	58.7%

Appendix Table 2: Approximate LinkedIn penetration rates (percentage of the population who subscribed to LinkedIn) in selected countries as of 2015. Source: Linkedin.com.

Country	LI penetration rate	Country	LI penetration rate
United Kingdom	30%	Ireland	22%
France	15%	Norway	19%
Italy	13%	Portugal	10%
Netherlands	35%	United States	38%
Spain	15%	Canada	31%
Belgium	18%	Brazil	11%
Sweden	21%	China	1%
Denmark	18%	India	3%

Appendix Table 3: Percentage of Internet users aged 16-64 who had visited Twitter during the last month (Q1 2015), for selected countries. Source: Globalwebindex.

Country	% who visited Twitter	Country	% who visited Twitter
Spain	39%	Sweden	22%
Ireland	36%	France	18%
Italy	31%	Netherlands	17%
United Kingdom	30%	Poland	16%
Germany	16%	Belgium	16%

Appendix Table 4: Multi-networking behavioUr: Percentage of active Twitter users who have accounts also on other social networking websites. Source: Globalwebindex, 2015.

Account	% of Twitter users with accounts on
Facebook	93%
Google+	78%
YouTube	76%
LinkedIn	48%
Instagram	44%
Pinterest	42%

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);

- more than one copy or posters/maps:
  from the European Union's representations
  (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries
  (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service
  (http://europa.eu/europedirect/index_en.htm) or calling 00 800 6 7 8 9 10 11
  (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

Priced subscriptions:

- via one of the sales agents of the Publications Office of the European Union
  (http://publications.europa.eu/others/agents/index_en.htm).

KE-02-16-632-EN-N

Publications Office

doi: 10.2767/61617