Demographic research with non-representative internet data

Emilio Zagheni
Department of Sociology, University of Washington, Seattle, Washington, USA, and
Ingmar Weber
Department of Social Computing, Qatar Computing Research Institute, Doha, Qatar

Abstract

Purpose – Internet data hold many promises for demographic research, but come with severe drawbacks due to several types of bias. The purpose of this paper is to review the literature that uses internet data for demographic studies and presents a general framework for addressing the problem of selection bias in non-representative samples.

Design/methodology/approach – The authors propose two main approaches to reduce bias. When ground truth data are available, the authors suggest a method that relies on calibration of the online data against reliable official statistics. When no ground truth data are available, the authors propose a difference in differences approach to evaluate relative trends.

Findings – The authors offer a generalization of existing techniques. Although there is not a definite answer to the question of whether statistical inference can be made from non-representative samples, the authors show that, when certain assumptions are met, the authors can extract signal from noisy and biased data.

Research limitations/implications – The methods are sensitive to a number of assumptions. These include some regularities in the way the bias changes across different locations, different demographic groups and between time steps. The assumptions that we discuss might not always hold. In particular, the scenario where bias varies in an unpredictable manner and, at the same time, there is no "ground truth" available to continuously calibrate the model, remains challenging and beyond the scope of this paper.

Originality/value – The paper combines a critical review of existing substantive and methodological literature with a generalization of prior techniques. It intends to provide a fresh perspective on the issue and to stimulate the methodological discussion among social scientists.

Keywords Demography, Internet data, Digital breadcrumbs, Non-representative samples, Selection bias

Paper type Research paper

Introduction

The global spread of internet and digital technologies has radically transformed the way in which we communicate with each other. As a consequence of the digital revolution, individuals leave an increasing quantity of traces online. These records can be aggregated and mined to advance our understanding of social processes. Web-based research has been rapidly increasing its prominence in the areas of epidemiology, economics, statistics, demography and sociology. Social scientists are increasingly using internet data in their research and, over the last years, have increasingly offered methodological contributions to the field of web data mining. In this article, we discuss the opportunities and challenges for the study of populations with digital records. We argue that demographers can gain a lot of relevant information from digital

The authors would like to thank Klaus Zimmermann, Nikos Askitas and two anonymous reviewers, for their very helpful comments and suggestions.

International Journal of Manpower
Vol. 36 No. 1, 2015
pp. 13-25

0143-7720
DOI 10.1108/IJM-12-2014-0261

records. At the same time, population scientists, with their traditional attention to data quality, statistical analysis and formal modeling, can play an important role in the methodological development of web data mining.

Demography has been a data-driven discipline since its birth, about 350 years ago. In 1662, John Graunt published the very first life table after compiling data that had been collected partially for "marketing" purposes: data on number of deaths was originally collected at the request of the merchants of London who wanted to evaluate the number of potential customers (i.e. live people by age) in London at times of epidemics (Graunt, 1662). Since then, data collection and development of formal methods have sustained most of the major advances in our understanding of population processes.

Today's increasing availability of digital records holds the promise of an exceptional development of new demographic knowledge, which will lead to theoretical advances in population studies. "Big data" is widely seen as a sea change. With this paper, we emphasize the importance of developing methods to extract information from massive (but also messy) data. There is a general debate about whether statistical inference can be made from non-representative samples. There is not a definite answer to this question. In this paper we discuss our perspective on an approach based on calibration and one based on a difference in differences method. Under certain circumstances, we believe that we can extract signal from noisy and unstructured data. We discuss under which assumptions the approaches are appropriate and remind the reader that it is always important to be aware of the limitations and of the data at hand, since the same analytical approach may work in some circumstances, but not others. Although the focus of the paper is on demographic research, the discussion of methods and substantive issues is quite general and relevant to a broader audience of social scientists.

The main contribution of this paper is about methodological aspects related to the analysis of large, but non-representative data sets, like the ones that are generated from activities on social media platforms. In the first section, we review the most relevant literature about using internet data for population studies. The second section is the core of the paper. We discuss two situations: when "ground truth" data from a census or a large survey exist, we propose an approach to adjust for selection bias that is closely related to the literature on calibration of stochastic microsimulation. When no independent and reliable empirical evidence is available for calibration of web data, we suggest a difference in differences technique in order to evaluate trends over time. Finally, we conclude by providing a general discussion intended to further stimulate the methodological debate among social scientists and to indicate new avenues for research in this area.

Mining and monitoring demographic variables

Monitoring and surveillance are essential tools to evaluate trends in demographic rates. Traditional data sources used by demographers include vital statistics, censuses, population-based surveys and health records. New and innovative data sources, like web engine search queries or social media posts could complement existing practices and provide new insights into demographic behavior.

Demography is a discipline that focuses on the study of populations. It is a broad field, without clearly defined boundaries. Demographic research is inherently interdisciplinary. For instance, it is often relevant to, and borrows tools from, sociology, economics, geography, statistics, public health and public policy. Although the scope of demographic research is quite broad, ultimately the fundamental objects of population studies are processes related to fertility, mortality and migration. In this section, we discuss the literature that uses internet data to offer insights into these core demographic processes.

The use of internet data for fertility research is still at an early stage, but holds promise. The two examples of research that stand out in this area use web search engine queries to understand and monitor demographic behavior. Reis and Brownstein (2010) showed that the volume of internet searches for abortion is inversely proportional to local abortion rates and directly proportional to local restrictions on abortion (Reis and Brownstein, 2010). Billari *et al.* (2013) showed that Google searches for fertility-related queries, like "pregnancy" or "birth", can be used to predict fertility intentions and fertility rates several months ahead. One of the most important messages of the paper is that combining traditional data sources with new data, like web searches, can improve the predictive power of demographic models. Critics of the use of web searches for understanding decisions of individuals and couples point to the potential problem that correlations between aggregate web searches and individual intentions may not persist for long periods of time. For example, the widely known Google flu approach to track influenza symptoms and detect potential outbreaks using web searches (Ginsberg *et al.*, 2008) has been very useful and successful. However, at times, it also produced largely erroneous estimates, typically when the nature of the relationship between searches, news and behaviors changed (Lazer *et al.*, 2014). Thus the results of these models have to be interpreted carefully and with caution. Nonetheless, there is a widespread need for timely indicators of fertility intentions and recent fertility trends. For example, the Vienna Institute of Demography has developed the "Fertility barometer" to monitor fertility in Austria and Vienna using administrative data sources [1]. Using web search queries could potentially lead to the creation of fertility indicators for a number of countries, provided that research in this area can reveal the underlying relationship between web searches, fertility intentions and actual behavior.

Marriage and union formation are demographic processes that are closely related to fertility. The study of marriage and the family is an important area of interest for social demographers. In this field, data from online dating web sites have been the most relevant internet data source so far. There is a fairly large literature: a few examples include works on matching algorithms (Hitsch *et al.*, 2010), the demography of online dating users (Valkenburg and Peter, 2007) and the social demography and economics of online dating (Sautter *et al.*, 2010; Oyer, 2014). There is a large potential for social scientists to understand the mechanisms that lead to union formation, and to evaluate the demographic consequences of usage of online dating web sites in terms of assortative mating, inter-racial marriage and the implications for transmission of inequality.

In the context of mortality research, the introduction, in developing countries, of verbal autopsies performed via cellphones is probably one of the projects that may have the largest impact for the study of mortality. In a large number of countries where civil registration systems are not fully developed, vital statistics are not routinely collected. As a result, it is hard to monitor health metrics over time. Cause of death data are collected by field workers who ask relatives about the circumstances of the death of a family member. This technique is called "verbal autopsy". The use of mobile phones to collect data from remote areas and store information in a central database has the potential to improve considerably our understanding of mortality in developing countries (Tamgno *et al.*, 2013). Research is this area is still at a relatively early stage, but holds promise.

Of all the major areas of demographic research, the study of migrations has likely benefited the most from the increasingly large availability of geo-located data from internet sources. Travel itineraries and short-term mobility have been inferred using various data sources, including geo-tagged pictures in Flickr (De Choudhury *et al.*, 2010),

recommendations posted on Couchsurfing (Pultar and Raubal, 2009), Twitter (Ferrari *et al.*, 2011), Google Latitude (Ferrari and Mamei, 2011) and Foursquare (Noulas *et al.*, 2011). Cellphone data have also been used to evaluate patterns and regularities of internal mobility, for a country or movements within a relatively small region (Bayir *et al.*, 2009; Gonzalez *et al.*, 2008; Candia *et al.*, 2008; Blumenstock, 2012; Deville *et al.*, 2014).

Trends in international flows of migrants have been estimated by tracking the locations, inferred from IP addresses, of users who repeatedly login into Yahoo! services (Zagheni and Weber, 2012; State *et al.*, 2013). Recently, geolocated Twitter tweets have been used to integrate the dimensions of internal and international migration (Zagheni *et al.*, 2014) and to study global mobility patterns (Hawelka *et al.*, 2014). LinkedIn data have proven useful to evaluate trends in migration by educational attainment and sector of employment (State *et al.*, 2014).

From a commercial point of view, data on geographic mobility is often valued for use in recommender systems, for example to suggest tourist travel routes. From the social science viewpoint, the initial motivation to use digital records for migration studies was the lack of official statistics. Data on migration flows are typically outdated and inconsistent across countries. For a number of countries, they do not exist at all (De Beer *et al.*, 2010). The first works in this area were mainly feasibility studies to assess the opportunities that internet data opened (Zagheni and Weber, 2012; State *et al.*, 2013). Gradually, we are observing a shift towards more substantive and theoretical contributions. For instance, State *et al.* (2014) look into labor migration, and used LinkedIn data to provide empirical evidence supporting the theory that the relative importance of the USA in the global network of professional migration is decreasing, mainly as a result of the rise of high-skilled migrations in other areas of the world. Zagheni *et al.* (2014) used geo-located Twitter data to document the recent rise of out-migration rates in countries like Ireland, and to show the potential of digital records to bridge the gap between the study of international and internal migration. The existent gap between these two areas of migration studies is a reflection of lack of data and the difficulty of collecting data for different types of migration studies. Internet does not generally have borders. Thus internet data has the potential to be transformative for migration studies.

Overall, the use of internet data for demographic research has been increasing. We expect that data from social media will continue to be used to drive substantive discoveries in the area of demography and, more broadly, in the social sciences. At the same time, there are a number of challenges related to the analysis of digital breadcrumbs. The next section focuses on selection bias and on strategies that can be adopted to address the issue.

Statistical inference from internet data

One of the most serious limitations to the study of demographic processes using internet data is the problem of selection bias. Internet and social media users are not representative of the whole population. Thus, without addressing the problem of selection bias, inference from populations of social media users may lead to unreliable results.

Social scientists have traditionally used surveys with probability samples to quantify social processes and demographic phenomena. Recently, concerns about coverage and non-response have revived the debate about whether non-probability sampling methods might be viable and cost-effective alternatives. For example, in 2011 the American Association of Public Opinion Research (AAPOR) appointed a task force to evaluate the extent to which non-probability samples can be used for statistical inference (Baker *et al.*, 2013b).

A number of approaches could be pursued to extract signal from non-probability samples, including network sampling (Heckathorn, 1997), and weight adjustment methods like propensity score matching (Rosenbaum and Rubin, 1983; Schonlau *et al.*, 2009). It has been shown that in some situations, like in the case of predicting election results, post-stratification methods, combined with multilevel regression models, can be effective (Wang *et al.*, 2014). In other situations, analyses with non-representative samples may perform poorly. There is an important and growing body of literature that deals with aspects related to data quality in web surveys (de Pedraza García *et al.*, 2010). The AAPOR report stresses that making statistical inferences requires some reliance on modeling assumptions, which have to be made explicit by the researcher. For instance, "post-stratification eliminates the bias due to selection or coverage problems if, within each adjustment cell, the probability that each case completes the survey is unrelated to the case value on the survey variable of interest" (Baker *et al.*, 2013a).

In this section, we discuss a new perspective about addressing the issue of selection bias in non-representative samples like the ones obtained from social media data. More specifically, we offer a generalization of the methods that we have proposed in the context of migration studies (Zagheni and Weber, 2012; Zagheni *et al.*, 2014), and we discuss how these methods are related to the literature on calibration of stochastic microsimulaton (Ševčíková *et al.*, 2007). In previous research, we tailored these approaches to the study of migration. Here, we would like to present them in a more general framework.

Estimating quantities of interest

We claim that it is useful to think of the problem of estimating quantities of interest from internet data as a calibration problem. This approach can be applied when there exist historic "ground truth" data that can be used to estimate the parameters of models that evaluate the extent of bias in internet data. Moreover, the approach requires assumptions about the functional form of the relation between quantities of interest and their online proxies, as well as internet penetration and socio-demographic variables. In the following, we explain the details.

Consider a quantity of interest, y_{ij}^t, for the geographic location i and demographic group j, at time t. The definition is intentionally broad: y could be any quantity of interest. It could be the number of births, deaths, abortions, marriages, unemployed people, measures of trust, attitudes towards marriage, etc. Assume that, for the time period t, this quantity can be considered "ground truth". In other words, y_{ij}^t comes from a reliable source, such as a census or a large representative survey.

For the same time period t, we have data from an internet source (like web searches for a particular keyword, or geo-located tweets) that are expected to approximate the quantity of interest. In other words, for each y_{ij}^t we have a respective quantity μ_{ij}^t that comes from digital breadcrumbs.

If the digital breadcrumbs were drawn from a representative sample, we would expect that:

$$y_{ij}^t = \left(k \times \mu_{ij}^t\right) + \epsilon_{ij}^t \tag{1}$$

where k is a constant that rescales the quantities in all groups and locations equally (that is why there are no subscripts to k). ϵ_{ij}^t is an error term, assumed to be normally distributed with mean equal to 0. In other words, we would expect that the quantities obtained from internet data would closely approximate the "ground truth" data,

provided that the model is correct and that a constant adjusts all the values in order to account for different units of measurement, or different propensities in the online population, compared to the offline population of interest.

Data from the internet (e.g. web searches or geo-located tweets) typically come from non-representative samples. We thus need to correct for selection bias beyond a constant adjustment for all groups. To account for that, Equation (1) can be re-written as:

$$y_{ij}^{t} = \left(k \times \mu_{ij}^{t}\right) + a_{ij}^{t} + \epsilon_{ij}^{t} \tag{2}$$

where a_{ij}^{t} is the bias for the location i and demographic group j.

We thus have that:

$$a_{ij}^{t} \approx y_{ij}^{t} - \left(k \times \mu_{ij}^{t}\right) \tag{3}$$

The main problem that the researcher faces is to model the selection bias. This entails identifying the best statistical model. In general terms, the bias terms can be expressed as a function of a number of covariates that are deemed relevant. The model does not have to be linear, but a simple example would be:

$$a_{ij}^{t} = \beta_0 + \beta_1 p_{ij}^{t} + e_{ij}^{t} \tag{4}$$

where p_{ij}^{t} is the internet penetration rate for the geographic location i, demographic group j and time period t, e_{ij}^{t} is a random error with mean 0. β_0 and β_1 are parameters that can be estimated, in simple situations, with the ordinary least squares method.

Depending on the context of the analysis, the statistical model can be more complex and include fixed effects, non-linearities, etc. More advanced models that account for the fact that, for instance, penetration rates are bounded between 0 and 100 per cent could be used (see e.g. Zagheni and Weber, 2012).

The main idea is simple. Estimates of the bias for specific locations and groups, $\hat{a}_{ij}$, can be obtained using existing traditional data sources and internet data for the respective time periods. Once the estimates for the parameters have been generated, the model can be used to make extrapolations for countries for which we have internet data, but no ground truth data about the quantity of interest. The calibration model can also be used anytime internet data become available before official statistics. In other words, the model would be calibrated at time $(t-1)$, when both official statistics and internet data are available. At time t, when only internet data are available, the "true" values for the quantity of interest can be estimated using the same model, thus assuming that the bias has remained constant over the short time period. That would be an example of "nowcasting". This formula would apply:

$$\hat{y}_{ij}^{t+1} = \left(k \times \mu_{ij}^{t+1}\right) + \hat{a}_{ij}^{t} \tag{5}$$

It is important to note that the general framework in Equation (2) is analogous to the one of calibration models for stochastic microsimulations (Ševčíková *et al.*, 2007). The underlying idea in the microsimulation literature is that simulations may generate estimates of quantities of interest that are biased. Identifying and modeling the bias is thus key to make statistical inference. We can consider social media and the internet as "laboratories" that produce estimates of quantities of interest that are biased, but in a systematic way. Here, "systematic" means that there are hidden, potentially stochastic

rules that determine the relationship between the online data and the offline quantities of interest. Note that the user base of the online services can grow and even change in composition as long as factors such as the relationship between the market penetration of the service and the age distribution of its users are part of the model. Thus, conditional on the model for the bias, we can make statistical inference for the quantities of interest using Bayesian techniques like the Bayesian melding (Poole and Raftery, 2000; Ševčíková *et al.*, 2007; Alkema *et al.*, 2008). In some situations, the relationship described by the model may not be robust. For example, the relationship may hold only for a limited number of geographic locations, or may hold only for a limited period of time, or may be spurious and mediated by variables not considered in the model. In those situations, the relationship is not "systematic" and thus the assumption would not hold.

In Equation (2) we showed a framework in which we use an additive model for the bias: the quantity a_{ij}^t, for each specific demographic group and location is added to $\left(k \times \mu_{ij}^t\right)$. We chose to present an additive model to emphasize the symmetry with simulation models where the standard modeling framework is additive (Ševčíková *et al.*, 2007). However, in some situations, a multiplicative model, or a combination of the two models, may be preferable, depending on the patterns of bias in the population. This is an empirical question, which could be evaluated on a case by case situation. Nonetheless, the same general approach could be pursued. For example, Equation (2) could be expressed in multiplicative terms as:

$$y_{ij}^t = \left(k \times \mu_{ij}^t\right) \times a_{ij}^t \times \in {}_{ij}^t \tag{6}$$

In this situation, a logarithmic transformation of the data would bring us back to an additive model:

$$\log y_{ij}^t = \log k + \log \mu_{ij}^t + \log a_{ij}^t + \log \in {}_{ij}^t$$

An illustrative example of models with multiplicative bias can be found in Zagheni and Weber (2012), who deal with the estimation of age-specific out-migration rates using IP-geo-located e-mail data. Essentially, they consider the bias in age-specific out-migration rates as a multiplicative factor. They model the bias as a function of age- and country-specific internet penetration rates. They then use "ground truth" data for European countries from Eurostat, to calibrate the model and estimate the parameters. Finally, they use the estimates of the bias, as a function of internet penetration rates, to correct the raw estimates from e-mail data. The underlying assumption, that is verified empirically, is that migration rates, for age groups and countries where internet penetration is low, may be overestimated. This is because those people who actually use e-mails, especially in developing countries, may be more highly educated, have more income and be more mobile than the general population. The model thus corrects migration rates downwards. The correction is bigger for the elderly and for developing countries.

The application in Zagheni and Weber (2012) is instructive for our purposes, not only because it is a specific case of the general approach described above, but also because it offers the opportunity to reflect on one of the limitations of modeling the bias in the context of a calibration model. Typically, "ground truth" data are available for more developed countries (i.e. European countries). These countries are also the ones with higher internet/social media penetration rates. Thus, the variability in the independent variable (penetration rate) is quite small as very few data points, if any, are available for

countries with low internet penetration rates. As a result, the uncertainty associated to predictions of the bias or correction factors at low internet penetration rates would be quite high. If this uncertainty were too large, that would undermine our ability to extract meaningful signal from the data. Zagheni and Weber (2012) show that, in some African countries, the combination of very low internet penetration rates and relatively small samples of users prevent the researchers from producing reliable estimates. Conversely, in South American countries, the calibration process is feasible, although it comes at the cost of higher uncertainty, compared to European and North American countries.

There is a general debate about whether statistical inference can be made from non-representative samples. There is not a definite answer to this question. Here we discussed our perspective on an approach based on calibration. Under certain circumstances, we believe that we can extract signal from noisy and messy data. However, it is always important to be aware of the limitations and the data at hand, since the same analytical approach may work in some circumstances, but not others.

Evaluating trends

In the previous section, we discussed a situation where ground truth data exist. In those cases, it is possible to address the issue of selection bias as a calibration problem. However, in some cases, ground truth data do not exist. Without any knowledge about the size and the direction of the bias, providing a reliable picture for the quantity of interest at one point in time is challenging. In such cases, instead of estimating the absolute value of variables of interest, we attempt to accomplish a more modest task: we aim at estimating trends, i.e. relative changes in quantities. In the following, we discuss this approach.

Concretely, we propose a difference in differences approach to estimate relative trends in the quantity of interest. Let y_i^t be the quantity of interest for a geographic location i at time t. Consider then a baseline location, z, that can be treated as a "control" group. This quantity could also be an average over several locations. The quantity of interest for that location at time t is y_z^t.

We are interested in relative trends for the true values y_i^t and y_z^t. However, what we observe are proxies from internet data, μ_i^t and μ_z^t, respectively. Under different assumptions, different models would be the most appropriate ones. We will start by discussing the situation of additive bias. Then we will consider the case of multiplicative bias.

Assume that:

$$\mu_i^{t-1} \approx y_i^{t-1} + n \tag{8}$$

and:

$$\mu_z^{t-1} \approx y_z^{t-1} + m \tag{9}$$

where n and m are additive biases for location i and z, respectively.

If between time $(t-1)$ and time t, the true value of interest y_i^t, increases by α units, and y_z^t increases by θ units, then the following difference in differences formula gives $(\alpha-\theta)$:

$$\hat{\delta}^t \approx \left(\mu_i^t - \mu_z^t\right) - \left(\mu_i^{t-1} - \mu_z^{t-1}\right) \tag{10}$$

In other words, if the two different biases, n and m stay constant, or if they change by the same number of additive units in both locations, then Equation (10) filters them out,

and we are left with the estimate $\hat{\delta}^t = (\alpha - \theta)$, which is the value of the true differential variation over the time period considered, for the two locations. Note that n and m could be any type of bias, including bias related to change in online behavior. They would cancel out as long as their additive change is equal in the two locations considered.

The estimation process can be dealt with standard regression models. For example to estimate $\hat{\delta}^t$, we could use the following model:

$$\mu_i^t = \beta_0 + \beta_1 G_i + \beta_2 T_t + \beta_3 G_i T_t + e_{it} \tag{11}$$

where G_i is an indicator variable that is equal to 1 if the observation is for the region i (our group of interest), and 0 if the observation is in the region z (our baseline group). T_t is an indicator variable that takes the value 1 in the "post-treatment" period t, and 0 at time $(t-1)$. It can be shown that the difference in differences estimator $\hat{\delta}^t$ is equal to the OLS estimate of β_3.

If in the data set at hand there are reasons to assume that the bias has a multiplicative effect, instead of an additive one, then we could write:

$$\mu_i^{t-1} \approx n \times y_i^{t-1} \tag{12}$$

and:

$$\mu_z^{t-1} \approx m \times y_z^{t-1} \tag{13}$$

where n and m would be the multiplicative biases for location i and z, respectively.

If the underlying variable of interest increases by a factor of α in location i and by a factor of θ in location z, from time $(t-1)$ to time t, we would have:

$$\mu_i^t \approx \alpha \times n \times y_i^{t-1} \tag{14}$$

and:

$$\mu_z^t \approx \theta \times m \times y_z^{t-1} \tag{15}$$

Taking a logarithmic transformation of Equations (12)-(15) will generate a model where the bias is additive in logarithms. Applying Equation (10) to the logarithmic transformations would yield $\hat{\delta}^t = \log\alpha - \log\theta = \log\frac{\alpha}{\theta}$. In this context, $e^{\hat{\delta}^t} = \frac{\alpha}{\theta}$, which is the true ratio of change for the underlying variables of interest. It is important to note that, in this case, if the two biases change in multiplicative ways by the same factor across locations, then they would cancel out in the difference in difference approach applied to logarithms. However, if n and m change by a different factor between time $(t-1)$ and t, then the difference in differences approach would not filter out the bias completely. This is not a one-fits-all approach. Instead, we are suggesting that careful examination of the nature of the bias, and development of the appropriate methods, would allow for ways to filter out the bias in some circumstances.

In those situations where no ground truth data are available, selection bias prevents us from making statistical inference for single points in time. In addition, changes in the composition of the samples (e.g. internet users, social media users, etc.) prevent us from using time series of internet-based estimates to make statistical inference about changes over time. However, if changes in the composition of social media users have similar trends across the geographic locations of interest, then the comparison of relative

changes for a single location with relative changes for the group of reference can be used to provide information about relative underlying trends.

Zagheni *et al.* (2014) proposed a difference in differences method to estimate trends in geographic mobility rates in OECD countries using geo-located Twitter data. Their study can be considered an illustrative example of the general difference in differences approach described above. More specifically, the main problem that Zagheni *et al.* (2014) address is that the composition of Twitter users changes over time. Thus, trends in time series of out-migration rates may be driven by compositional changes in the Twitter population. For example, if young people sign up for Twitter at a higher rate than relatively old individuals, even if the rates of out-migration in the general population remain constant, over time we would observe an increase in the out-migration rates obtained from Twitter data. This would happen because younger people have higher mobility rates than older ones. Thus, if the population of Twitter users becomes younger over time, the rate of migration of Twitter users would increase. Similar effects can also be caused by changes in usage pattern. For instance, as Twitter matures as a platform, even relatively old individuals might start using geo-tagging for their tweets. Using the difference in differences approach, where one country, or a group of countries, is the control, allows the researcher to evaluate relative differences in trends across countries and to asses which countries experienced increases or decreases in migration rates, in relative terms.

Zagheni *et al.* (2014) used the difference in differences approach described above to filter out compositional changes in the population of users of a specific social media. The example is relevant because it shows a potential use of the difference in differences method. Nonetheless, it is important to reflect about the assumptions and limitations of the approach. A critical underlying assumption is that bias was assumed to be additive. If that is not the case, data transformations may be needed. For instance, as we showed earlier, if there is evidence that the bias affects the estimates in a multiplicative way, then a logarithmic transformation would be appropriate to revert to an additive model.

An additional assumption is that compositional changes in the population of users are similar across the countries of interest. If in a given country Twitter subscriptions are growing exponentially, whereas in a different one they are growing linearly, a difference in differences approach would not be effective. Equally, if in one country young people are signing up for Twitter, and in another country old individuals are mainly signing up, then comparisons across countries would not be reliable. Finally, a key assumption is that relative changes for selected groups of the population may be indicative of trends in the general population. If the users of a given social media outlet are predominantly teenagers, we would have very little information about other people. Using a difference in differences approach would imply that there is a structured relationship between a subgroup of the population and the rest of the population, so that trends in the quantity of interest for a subgroup are indicative of trends for the whole population.

All the assumptions described above constrain the use of difference in differences to a limited number of cases. Our goal here is to discuss the potential and limitations of the approach. There are a number of situations where the method can be used to analyze social media and web data. In some cases, the approach may be more effective than in others. A careful evaluation of the assumptions is needed in order to assess whether the difference in differences method is suitable to extract signal from biased data.

Discussion and conclusion

As more and more of human activity leaves digital traces on the web, the use of these data for demographic research will become more and more common. In this paper, we reviewed the state of the art of demographic research with web data and described methods to address issues particularly pertinent to web data: non-representativeness and selection bias. We proposed a general framework to extract information from biased web data. We offered an estimation method that is closely related to the literature on calibration of stochastic microsimulation. When no independent and reliable empirical evidence is available for calibration of web data, we suggested a difference in differences technique in order to evaluate trends over time.

At the heart of the paper is the question of whether statistical inference from non-representative samples is possible. Our approach shows that there are a number of social science tools that can be leveraged to extract signal from noisy, messy and biased data. All of the methods have limitations and rely on assumptions that could be quite strong and that should be considered carefully. There is not a single theoretical framework like for the case of surveys from probabilistic samples. Instead, a number of approaches should be taken into consideration for different situations: the most appropriate one should then be selected. This paper is intended to stimulate methodological discussions among social scientists who use web data.

We believe that, for a number of studies, including the ones related to the measurement of migrations, models that address the issue of selection bias could be tested using administrative records from Nordic countries, where complete population registries exist. For example, Sweden keeps tracks of internal and international migrants. These data could be used as "ground truth" for calibration procedures. Analogously, the data could be used to test the validity of the difference in differences approach and the goodness of the assumptions. For instance, one could select a biased sample that mimics the population of social media users, and test whether the estimates are in line with the trends for the overall population. More importantly, we could experiment with the data and test the sensitivity of the results to the extent to which the assumptions are met, and to compositional changes. We believe that this could be an important line of future research.

This paper focused on selections bias. However, there are many other issues that may complicate statistical inference from web data. For example, there could be measurement errors. A notable example is the estimation of geographic location from IP addresses: the estimates could be noisy. Plus, the use of proxy servers may add additional noise and complexities to studies of geographic mobility based on IP address geo-location. Often, there is not a one-on-one correspondence between users of social media services and accounts. Some users may have more than one account. Some accounts may be open on behalf of institutions. All of these issues increase the complexity of the analysis substantially.

With this paper we did not intend to offer definitive solutions to very complex problems. We described the existing approaches and discussed them in a critical way. We believe that social scientists are well positioned to address a number of issues that we raised in this paper. We hope that this paper would help to stimulate the development of methods for estimating demographic quantities of interest from non-representative data.

Note

1. www.oeaw.ac.at/vid/barometer/index.html

References

Alkema, L., Raftery, A.E. and Brown, T. (2008), "Bayesian melding for estimating uncertainty in national HIV prevalence estimates", *Sexually Transmitted Infections*, Vol. 84 No. 1, pp. i11-i16.

Baker, R., Brick, J., Bates, N., Battaglia, M., Couper, M., Denver, J., Gile, K. and Tourangeau, R. (2013a), "Non-probability Sampling", report of the AAPOR Task Force, American Association for Public Opinion Research, Boston, MA.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013b), "Summary Report of the AAPOR Task Force on Non-probability Sampling", *Journal of Survey Statistics and Methodology*, Vol. 1 No. 2, pp. 90-105.

Bayir, M.A., Demirbas, M. and Eagle, N. (2009), "Discovering spatiotemporal mobility profiles of cellphone users", *World of Wireless, Mobile and Multimedia Networks & Workshops, WoWMoM 2009, IEEE International Symposium*, *IEEE*, pp. 1-9.

Billari, F., D'Amuri, F. and Marcucci, J. (2013), "Forecasting births using google", *Annual Meeting of the Population Association of America*, New Orleans, LA.

Blumenstock, J.E. (2012), "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda", *Information Technology for Development*, Vol. 18 No. 2, pp. 107-125.

Candia, J., Gonzalez, M.C., Wang, P., Schoenharl, T., Madey, G. and Barabási, A.-L. (2008), "Uncovering individual and collective human dynamics from mobile phone records", *Journal of Physics A: Mathematical and Theoretical*, Vol. 41 No. 22, p. 224015, available at: http://iopscience.iop.org/1751-8121/41/22/224015

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*, Vol. 26 No. 4, pp. 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", *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, ACM*, pp. 35-44.

de Pedraza García, P., Tijdens, K., de Bustillo Llorente, R.M. and Steinmetz, S. (2010), "A Spanish continuous volunteer web survey: sample bias", *Weighting and Efficiency. Reis: Revista Espan˜ola de Investigaciones Sociológicas*, Vol. 131, pp. 109-130.

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*, Vol. 111 No. 45, pp. 15888-15893.

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

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

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2008), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457 No. 7232, pp. 1012-1014.

Gonzalez, M.C., Hidalgo, C.A. and Barabasi, A.-L. (2008), "Understanding individual human mobility patterns", *Nature*, Vol. 453 No. 7196, pp. 779-782.

Graunt, J. (1662), "Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality", reprinted by Ayer Company Pub, 1975.

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*, Vol. 41 No. 3, pp. 260-271.

Heckathorn, D.D. (1997), "Respondent-driven sampling: a new approach to the study of hidden populations", *Social Problems*, Vol. 44 No. 2, pp. 174-199.

Hitsch, G.J., Hortaçsu, A. and Ariely, D. (2010), "Matching and sorting in online dating", *The American Economic Review*, Vol. 100 No. 1, pp. 130-163.

Lazer, D.M., Kennedy, R., King, G. and Vespignani, A. (2014), "The parable of google flu: traps in big data analysis", *Science*, Vol. 343 No. 6176, pp. 1203-1205.

Noulas, A., Scellato, S., Mascolo, C. and Pontil, M. (2011), "An empirical study of geographic user activity patterns in foursquare", *ICWSM*, Vol. 11, pp. 70-573.

Oyer, P. (2014), *Everything I Ever Needed to Know about Economics I Learned from Online Dating*, Harvard Business Review Press, Boston, MA.

Poole, D. and Raftery, A.E. (2000), "Inference for deterministic simulation models: the bayesian melding approach", *Journal of the American Statistical Association*, Vol. 95 No. 452, pp. 1244-1255.

Pultar, E. and Raubal, M. (2009), "A case for space: physical and virtual location requirements in the couchsurfing social network", *Proceedings of the 2009 International Workshop on Location Based Social Networks, ACM*, pp. 88-91.

Reis, B.Y. and Brownstein, J.S. (2010), "Measuring the impact of health policies using internet search patterns: the case of abortion", *BMC Public Health*, Vol. 10 No. 1, pp. 514.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, Vol. 70 No. 1, pp. 41-55.

Sautter, J.M., Tippett, R.M. and Morgan, S.P. (2010). The social demography of internet dating in the United States", *Social Science Quarterly*, Vol. 91 No. 2, pp. 554-575.

Schonlau, M., Van Soest, A., Kapteyn, A. and Couper, M. (2009), "Selection bias in web surveys and the use of propensity scores", *Sociological Methods & Research*, Vol. 37 No. 3, pp. 291-318.

Ševčíková, H., Raftery, A.E. and Waddell, P.A. (2007), "Assessing uncertainty in urban simulations using bayesian melding", *Transportation Research Part B*, Vol. 41 No. 6, pp. 652-669.

State, B., Rodriguez, M., Helbing, D. and Zagheni, E. (2014), "Migration of professionals to the US: evidence from linkedin data", *Proceedings of the 6th International Conference on Social Informatics (SocInfo)*.

State, B., Weber, I. and Zagheni, E. (2013), "Studying inter-national mobility through IP geolocation", *WSDM* Barcelona, pp. 265-274.

Tamgno, J.K., Faye, R.M. and Lishou, C. (2013), "Verbal autopsies, mobile data collection for monitoring and warning causes of deaths", *Advanced Communication Technology (ICACT), 2013 15th International Conference on, IEEE*, pp. 495-501.

Valkenburg, P.M. and Peter, J. (2007), "Who visits online dating sites? Exploring some characteristics of online daters", *CyberPsychology & Behavior*, Vol. 10 No. 6, pp. 849-852.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2014), "Forecasting elections with non-representative polls", *International Journal of Forecasting*, doi: http://dx.doi.org/10.1016/j.ijforecast.2014.06.001.

Zagheni, E. and Weber, I. (2012), "You are where you e-mail: using e-mail data to estimate international migration rates", *Proceedings of Web Science (WebSci)*, pp. 348-351.

Zagheni, E., Garimella, V.R.K., Weber, I. and State, B. (2014), "Inferring international and internal migration patterns from twitter data", *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web (WWW)*, pp. 439-444.

Corresponding author
Professor Emilio Zagheni can be contacted at: emilioz@uw.edu