International Statistical Review

International Statistical Review (2020) doi:10.1111/insr.12404

On Two Existing Approaches to Statistical Analysis of Social Media Data

Martina Patone[1] and Li-Chun Zhang[1,2,3]

[1] *Department of Social Statistics and Demography, University of Southampton, Southampton, UK*
E-mail: M.Patone@soton.ac.uk

[2] *Statistisk sentralbyrå, Oslo, Norway*

[3] *Department of Mathematics, University of Oslo, Oslo, Norway*

Summary

Using social media data for statistical analysis of general population faces commonly two basic obstacles: firstly, social media data are collected for different objects than the population units of interest; secondly, the relevant measures are typically not available directly but need to be extracted by algorithms or machine learning techniques. In this paper, we examine and summarise two existing approaches to statistical analysis based on social media data, which can be discerned in the literature. In the first approach, analysis is applied to the social media data that are organised around the objects directly observed in the data; in the second one, a different analysis is applied to a constructed pseudo survey dataset, aimed to transform the observed social media data to a set of units from the target population. We elaborate systematically the relevant data quality frameworks, exemplify their applications and highlight some typical challenges associated with social media data.

Key words: measurement; non-probability sample; quality; representation; test.

1 INTRODUCTION

There has been a notable increase of interest from researchers, companies and governments to conduct statistical analysis based on social media data collected from platforms such as Twitter or Facebook. At the same time, there is also a growing concern about various issues associated with these new types of data. For instance, Boyd & Crawford (2012) ask whether such data may alter what 'research' means, and they call for the need to question relevant assumptions and biases. Bright *et al.* (2014) argue that caution is needed when interpreting social media data, and major questions remain on how to employ such data properly. Hsieh & Murphy (2017) highlight what they call coverage error, query error and interpretation error in relation to Twitter data. Halford *et al.* (2017) urge to develop better understanding of the construction and circulation of social media data, to evaluate their appropriate uses and the claims that might be made from them.

The aim of this paper is to examine and summarise two existing approaches to statistical analysis based on social media data, when the analysis otherwise would have been possible

based on the traditional approach of survey sampling. To fix the scope, let $U = \{1, 2, \ldots, N\}$ be a target population of *persons*. Let y_i be an associated value for each $i \in U$. Let the parameter of interest be a function of $y_U = \{y_1, \ldots, y_N\}$, denoted by

$$\theta = \theta(y_U).$$

For instance, θ can be the population total or mean of the y-values. The quality of sample survey data can generally be examined with respect to two dimensions: representation and measurement (Groves *et al.*, 2004). The representation dimension concerns the relationship between U and the *observed* set of persons, denoted by s. For example, s suffers from undercoverage if there are persons in U who have no chance of being included in s. The measurement dimension concerns the potential discrepancy between y_i and the *obtained* measures, denoted by y_i^* for $i \in s$. For instance, y_i^* may be subjected to various causes of measurement error, such that $y_i^* \neq y_i$ for some persons in s.

Thus, when social media data are employed, one needs to address two basic obstacles with respect to each quality dimension. Firstly, social media data are initially organised around different units than persons; secondly, the relevant measures typically cannot be directly observed but need to be processed using algorithms or machine learning techniques. For example, one may like to make use of the relevant tweets to estimate the mean of a value associated with the resident population of a country. The directly observed unit (or data object) is then the tweet, whereas the statistical unit of interest is the resident. Next, instead of using designed survey instruments to measure the value of interest as one could in survey sampling, one will need to process a proxy to the target value from the Twitter texts by means of text mining.

Two existing approaches can be discerned in the literature. In what we refer to as the *one-phase approach*, statistical analysis is directly applied to the observed social media data that are organised around data objects other than persons. An example is Yan *et al.* (2019), who document statistical association between available drug-related tweets (processed by text mining techniques) in May–December 2012 and US county crimes rates (calculated against population size adjusted for non-residents) over 2012–2013. Next, in the *two-phase approach*, a different analysis is applied to a constructed *pseudo survey dataset*, after transforming the observed social media data to a set of persons from the target population. An example is Rampazzo *et al.* (2018), who document correlation between fertility rate published by the United Nations and that can be calculated for Facebook users. The pseudo survey dataset is collected directly from the Facebook Advertising Platform, which is assumed to be cleared of bots or other non-human accounts. The variable 'number of children' is also prepared by Facebook based on the information the company has about the users.

In this paper, we shall delineate these two approaches more generally and systematically than they have hitherto been treated in the literature, where the Social Media Index (SMI) for Dutch Consumer Confidence (Daas & Puts, 2014) serves as a typical case of the one-phase approach, and the Office for National Statistics study on residency and mobility data constructed from geolocalised tweets (Swier *et al.*, 2015) is used to illustrate the construction of pseudo survey dataset under the two-phase approach. We shall elaborate the relevant data quality frameworks and methodologies and highlight some typical challenges to statistical analysis.

The rest of the paper is organised as follows. In Section 2, we systematically describe the general issues of representation and measurement of social media data. In Sections 3 and 4, we delineate and examine the one-phase and two-phase approaches, respectively. Finally, some concluding remarks are provided in Section 5.

2 GENERAL ISSUES OF REPRESENTATION AND MEASUREMENT

2.1 Representation

A major concern about the use of social media for research is the non-representativeness of data, when the population of interest does not coincide with the social media population (Boyd & Crawford, 2012; Bright *et al.*, 2014; Halford *et al.*, 2017; Hsieh & Murphy, 2017). Meanwhile, when investigating the representativeness of a social media population, one often compares it to the resident population of a country, about which one has high-quality statistics. For instance, Pew Research Centre publishes every year a report on the use and participation in social media of the US population. It is shown that US users of Twitter and Facebook tend to be younger and more educated than the US resident population (Greenwood *et al.*, 2016). In the UK, Blank & Lutz (2017) find that Facebook users are more likely to be younger and female, while LinkedIn users are more likely to have a higher income than non-users. Mellon & Prosser (2016) examine how Twitter and Facebook users differ from the UK resident population in terms of demographics, political attitudes and political behaviour.

Twitter provides a typical example of online news and social networking site. Communication occurs through short messages, called *tweets*; the act of sending tweets is called *tweeting*. To be able to tweet, an account needs to be created. To register, a user has to provide an email address, a username and a password. A user can be a person, a business, a public institution, even softwares (bots) and so forth. In case of person, the user is not obliged to create an account reflecting her or his physical persona. Optional fields include a profile picture, a bio and a location, which are neither verified nor expected to accurately characterise the user. By default, tweets are publicly available, although the user may change the privacy setting to make it private. Each tweet can be original, a reply to another tweet or a copy of a different tweet, known as a retweet. It can mention a username account (@) to address a specific user, and it can contain hashtag (#) to declare the topic of the tweet. Hashtags offer a way to categorise tweets into specific topics (e.g. a tv show, a sport event and a news story). Some events such as football matches, film festivals or conferences may have an official hashtag under which the relevant tweets about the event are classified. Hashtags can also be user-specific and not intelligible to the general public.

As in the Twitter example, one can identify two directly observable units of data on most social media platforms, which we will refer to as the *post* and the *account*:

Post We use the generic term post to refer to the immediate packaging of social media content, which otherwise has a platform-specific name: Facebook has posts, Twitter has tweets, Instagram uses picture and so forth.

Account An account is the ostensible generator of a post. As in Twitter, the user(s) operating a social media account can be different entities including but not limited to persons. Moreover, the same user can have multiple accounts, but the connections between these accounts and the user are not publicly accessible.

Denote by P and A, respectively, the totality of all the posts and accounts on a given social media platform. There is a many-one relationship from posts to the active accounts, denoted by $A_P = a(P)$, and the inactive accounts $A \setminus A_P$ is non-empty in general. Next, there is a many-one relationship from accounts to the users, denoted by $b(A)$. The *observable* persons are given by the joint set of the target population U and $u_{AP} = b(A_P) = b(a(P))$, that is, via the active accounts. Moreover, $U \setminus u_{AP}$ is non-empty as long as there are persons not engaged with the given social media platform, and $u_{AP} \setminus U$ is non-empty as long as they are other users than persons. These relationships are summarised in Table 1.

Table 1. *Many-one relations a from post to account, and b from account to user*

	Post	Account	Person
Totality	P	A	U
Observable	P	$A_P = a(P)$	$U \cap u_{AP}, u_{AP} = b(A_P) = b(a(P))$
		$A \setminus A_P \neq \varnothing$	$U \setminus u_{AP} \neq \varnothing, u_{AP} \setminus U \neq \varnothing$
Sample	i. $s_P \subset P$	i. $s_A = a(s_P)$	$U \cap s_{AP}, U \setminus s_{AP} \neq \varnothing, s_{AP} \setminus U \neq \varnothing$
	ii. $s_P \subset a^{-1}(s_A)$	ii. $s_A \subset A$	i. $s_{AP} = b(a(s_P))$, ii. $s_{AP} = b(s_A)$

Next, a common way of collecting data from a given social platform is via the public APIs, either directly or indirectly through third-party data brokers; Web scraping provides another option, albeit with unclear legal implications at this moment. Via the APIs, a sample of posts or, less commonly, accounts is harvested directly from the social media company, and the obtainable sample depends on the company's terms and conditions. Depending on the API, the obtained datasets may differ in terms of being real-time or historical, or the amount of data that is allowed for.

Gaffney & Puschmann (2013) provide an overview of the tools available to extract Twitter data. For example, the `Streaming API` returns two possible samples: a 1% sample of the total firehose (the firehose is the totality of tweets ever tweeted), without specifying any filter; or a sample of posts on specific keywords or other metadata associated to the post. However, if the number of posts matching these filters is greater than 1% of the firehose, the Twitter API returns at most 1% of the firehose. In addition, historical tweets can be retrieved using the `Search API`, which provides tweets published in the previous 7 days, with a selection based on 'relevance and not completeness' (Twitter Inc.). For both APIs, Twitter does not provide the details of the process involved nor guarantees that the sampling is completely random. See, for example, studies that have been conducted to understand and describe how the data generation process works with Twitter (Morstatter *et al.*, 2013; González-Bailón *et al.*, 2014; Wang *et al.*, 2015).

Sampling of accounts is less common, which is only feasible if the usernames are known in advance. Consider the case where the interest is on the political candidates during an election. If a complete list of their usernames is available, sampling can be performed by the analyst; all the posts generated by the sample accounts on the social media platform can possibly be retrieved. The approach is only applicable when the group is made of 'elite' users (of known people), rather than 'ordinary' users; for instance, it is not always possible to identify all the eligible or potential voters. Rebecq (2018) and Berzofsky *et al.* (2018) use the user ID number to randomly select a set of users from Twitter. Both the authors also use the available connections between users to propagate the initial sample.

Thus, the actually observed units are generally either a subset of P or A to start with. An initial observed *sample* of posts, denoted by $s_P \subset P$, can lead one to a corresponding sample of accounts $s_A = a(s_P)$ and, then, in principle, a sample of users $s_{AP} = b(a(s_P))$. Given a sample s_A directly selected from A, we can possibly acquire a sample of users $s_{AP} = b(s_A)$ and a sample of associated posts, denoted by $s_P = a^{-1}(s_A)$. The observed sample of persons is given by the joint set of U and s_{AP}. Again, both $U \setminus s_{AP}$ and $s_{AP} \setminus U$ are non-empty in general. The relationships are summarised in Table 1 as well.

2.2 Measurement

Unlike in sample surveys, social media data are not generated for the purpose of analysis. They have been referred to as 'organic data' (Groves, 2011) to emphasise their non-designed

origin. One can only decide what is best to do with the data given the state in which they are 'found'. In light of the discussion of representation above, the obtained measures from social media data are associated with either the sample of posts or accounts. These may be based on the content of a post such as a text or an image, or the metadata of a post or account, such as the geolocation of a post or the profile of an account. According to Bright *et al.* (2014) and Japec *et al.* (2015), social media data are seen to provide the opportunity to study the following social aspects: (1) to capture what people are thinking, (2) to analyse public sentiment and opinion and (3) to understand demographics of a population. To this, one may add that social media data can obviously provide data about certain network relationships between posts, accounts and users.

Take Twitter, for example, of all the possibilities mentioned above. While Twitter does not provide the information whether a user is a parent or not, it may sometimes be possible to infer that the user behind a tweet is a parent based on its content. Similarly, while Twitter does not provide the location of a user, it is sometimes possible to infer this from the location (or content) of the relevant tweets. When opinions about a particular topic are of interest, sentiment analysis can be performed on each tweet. By analysing the frequency of different hashtags, it could be possible to investigate the major topics that capture people's attention at a given moment. Finally, retweeting or the inclusion of certain hashtags can reveal particular network connections between the different users.

Generally, we shall distinguish three types of data extraction from the sample posts and accounts, while at the same time noting the associated challenges in each respect:

Content Thought, opinion and sentiment provide typical examples of content extraction, which are the direct interest of study. Sentiment analysis is a common technique for extracting opinion-oriented information in a text. However, social media posts present some distinct challenges, because the expressions may be exaggerated or too subtle (Pang & Lee, 2008). Moreover, the posts on social media are public by nature, such that a user may easily be influenced by other opinions, or she may want to project an image of herself which does not necessarily represent the truth.

Feature Demographics, location and socio-economic standing are common examples of feature extraction, when these are not the direct interest of study but may be useful or necessary for disaggregation and weighting of the results. Various techniques of 'profiling' have been used for feature extraction. For instance, Daas *et al.* (2016) and Yildiz *et al.* (2017) consider the problem of estimating age and gender of Twitter users on the basis of the user's first name, bio, writing style and profile pictures. Or Swier *et al.* (2015) derive the likely place of residence of a user, from all the geo-located tweets that the user has posted. Completely accurate feature extraction is generally impossible regardless of the techniques.

Network Directional posting, reposting, sharing, following and referencing all provide the possibility of observing network relationships among the posts, accounts or users. Common interests regarding the pattern and interaction among social network actors include identifying the most influential actor and discovering network communities. Tabassum *et al.* (2018) provide an overview for social network analysis. However, it should be noted that the possibility and ease of network extraction are to a large extent limited by the APIs provided for a given social media platform.

In light of the above, whether by content, feature or network extraction from available social media data, one should generally consider the obtained measures as proxy values to the ideal target values. Of course, measurement errors are equally omnipresent in sample surveys. For instance, survey responses to questions of opinion may be subjected to mode effects, social

desirability effects and various other causes of measurement error (e.g. Biemer *et al.* 2004). So there is certainly scope for exploring social media data for relevant studies.

There is a noteworthy distinction between measurement errors in survey and social media data. In sample surveys, a measurement error does not affect the representation of the observed sample. The matter differs with social media data. For instance, when relevant accounts to a study are selected based on the metadata of an account, such as place of residence, errors can arise if the information recorded at the time of registration is not updated despite that there has actually been a change of the situation. Such an error can then directly affect which accounts are selected for the study, that is, the representation dimension of data quality. An initial measurement error in the description of the account can thus result in a coverage error with respect to the study population. Similarly, one may fail to include a post in a study if it is classified as not containing the relevant opinion of interest.

It may be envisaged that combining multiple platforms, such as Twitter and LinkedIn, can be useful for enhancing the accuracy of data extraction, although we have not been able to find any documented examples. This could be due to ethical reasons or the limitations imposed by the terms of conditions of the social media companies. An additional concern could be the 'interaction' between representation and measurement just mentioned above, where, for example, the accounts for which data combination is possible are subjected to an extra step of selection from the initially observed sample of accounts.

3 ONE-PHASE APPROACH

In the one-phase approach, one needs to estimate the target parameter $\theta = \theta(y_U)$ directly from the obtained measures, denote by z_j, associated with a different observed set of units s_P or s_A, despite the differences to y_i and U.

To see why this may be possible at all, consider the following example. Suppose one is interested in the totality of goods (θ) that have been purchased in a shop over a given time period. One could survey all the people who have been in the given shop during the period of interest and ask what they have purchased. The population U then consists of all the relevant persons, and y_i is the number of goods they have purchased (possibly over multiple visits to the shop). Alternatively, θ can be defined based on the transactions registered over the counter. The population P consists then of all the relevant transactions, and z_j is the number of goods associated with each transaction $j \in P$. Clearly, despite the differences in (y_i, U) and (z_j, P), either approach validly aims at the same target parameter θ.

Below, we reexamine the SMI (Daas & Puts, 2014) as an application, to formalise this approach and the relevant quality issues and methodological challenges.

3.1 Case: Social Media Index

Every month, Statistics Netherlands conducts a sample survey to compute the consumer confidence index (CCI). It is based on a questionnaire of people's assessment of the country economy and their financial situation. As part of the research on the use of social media data in official statistics (Daas & Puts, 2014; Daas *et al.*, 2015), the authors collected posts from different social media platforms and constructed the SMI from these posts. They observed and compared the CCI and SMI over time and concluded that the two series are highly correlated (Figure 1).

The SMI is constructed as an index that measures the overall sentiment of social media posts. The posts were purchased, in the time period between June 2010 and November 2013, from the

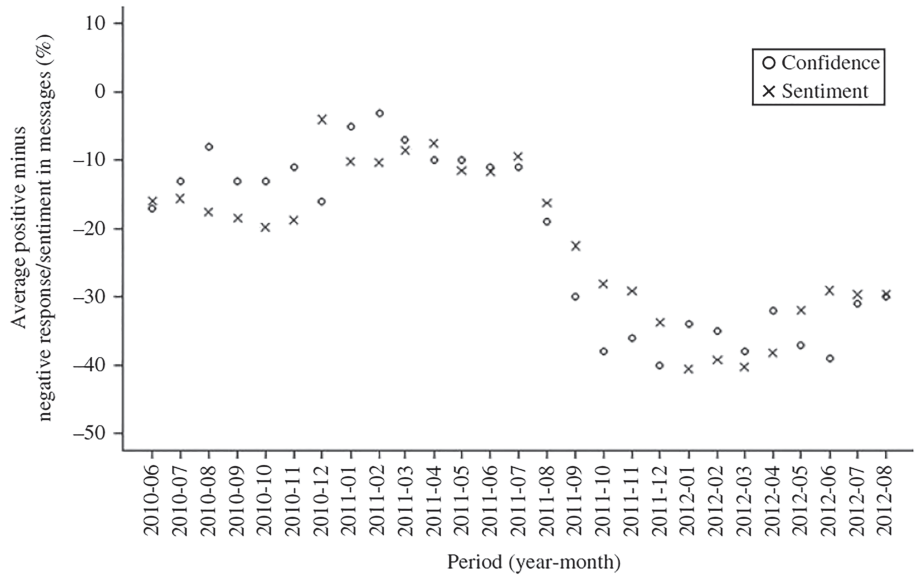

Figure 1. *Comparison of Dutch consumer confidence index (CCI) and Social Media Index (SMI) on a monthly basis. A correlation coefficient of 0.88 is found for the two series (Daas et al., 2015).*

Dutch company Coosto, which gathers social media posts written in the Dutch language on the most popular social media of the country (Facebook, Twitter, LinkedIn, Google+ and Hyves). Coosto also assigns a sentiment classification—positive, neutral or negative—to each post on the basis of sentiment analysis (Pang & Lee, 2008), which determines the overall sentiment of the combination of words included in the text of the post. A neutral label is assigned when the text does not show apparent sentiment.

Let P_t be the totality of all the observed posts in month t. Let $s_{P,t}$ be a subset of posts that are selected from P_t. Let m_t be the size of $s_{P,t}$. The posts included in $s_{P,t}$ can have positive, neutral, or negative sentiment value, respectively denoted by $z_j = 1,0,-1$, for $j \in s_{P,t}$. The SMI is calculated as the percentage difference between the positive and negative posts in $s_{P,t}$, that is, a function of $z_{s_{P,t}} = \{z_j; j \in s_{P,t}\}$:

$$\mathrm{SMI}_t = \mathrm{SMI}(z_{s_{P,t}}) = \frac{100}{m_t} \sum_{j \in s_{P,t}} z_j .$$

Daas & Puts (2014) experimented with different ways of selecting the sample $s_{P,t}$. The choices involve a decision about which social media platforms to include, and whether to accept all the posts from an included platform or only certain groups. The groups can be filtered using a set of keywords, such as posts containing personal pronouns like 'I', 'me', 'you' and 'us', or words related to the consumer confidence or the economy, or words that are used with high frequency in the Dutch language. The idea is that selecting only certain groups of posts could affect the association between the SMI and the CCI. For instance, from a previous study (Daas *et al.*, 2012), the same authors found that nearly 50% of the tweets produced in the Netherlands can be considered a 'pointless bubble'. In the end, $s_{P,t}$ is chosen to include all the Facebook posts and filtered Twitter posts, for which the resulting SMI achieved the highest correlation coefficient with the CCI (Figure 1).

Finally, considering the SMI as an estimator with its own expectation and variance, let

$$\mathrm{SMI}_t = \xi_t + d_t\,, \tag{1}$$

where ξ_t is the expectation of the SMI and d_t has mean 0 and variance τ_t^2.

3.2 Formal Interpretation

To assess the SMI as a potential replacement of the CCI, let us now formalise the CCI and its target parameter. Let U_t be the Dutch *household* population in month t, which is of the size N_t. Let y_i, for $i \in U_t$, be a consumer confidence score for household i based on positive, neutral or negative responses to five survey questions. The target parameter of the CCI is given by

$$\theta_t = \theta(y_{U_t}) = \frac{100}{N_t} \sum_{i \in U_t} y_i\,.$$

The CCI based on the sample survey is an estimator of θ_t, which can be given by

$$\mathrm{CCI}_t = \theta_t + e_t\,, \tag{2}$$

where e_t is the sample survey error of the CCI. For our purpose here, we shall assume that $e_t \sim N(0, \sigma_t^2)$, that is, normally distributed with mean 0 and variance σ_t^2.

Now that there is a many-one relationship between persons and households, the generic relationships from posts to persons apply equally from posts to households. The households corresponding to the SMI sample $s_{P,t}$ can thus formally be given as

$$s_t = U_t \cap a\,(b(s_{P,t}))\,.$$

Let s_t be of the size n_t. Let the target parameter defined for s_t be given by

$$\theta_{s,t} = \theta(y_{s_t}) = \frac{100}{n_t} \sum_{i \in s_t} y_i\,.$$

In order to replace the CCI by the SMI, it is clear that one would like to have $\theta_t = \xi_t$. However, given the underlying relationship between the social media data posts and the target population, one can only establish an analytic connection between ξ_t and $\theta_{s,t}$, based on the relationship between $(z_j, s_{P,t})$ and (y_i, s_t). It is therefore clear that the principal difficulty for the one-phase approach in this case is the lack of an explicit connection between ξ_t and $\theta_t = \theta(y_{U_t})$, or between $\mathrm{SMI}(z_{s_{P,t}})$ and $\theta(y_{U_t})$. Moreover, it seems that in such situations, external validation will be necessary in order to establish the validity of the analysis results on the basis of social media data, which we consider next.

3.3 Statistical Validation

In the case of the SMI, one does have the possibility of validating its statistical relationship to the CCI, despite the lack of an analytic connection between the two. As can be seen in Figure 1, the two indices display a high correlation with each other over time: the empirical correlation coefficient is 0.88 over the 27 months displayed. However, a high correlation between the two indices alone is not enough. Below, we formulate a test to exemplify a possible venue for statistical validation in similar situations.

As a conceivable scenario in which the SMI can replace the CCI, we set up the null and alternative hypotheses below:

$$H_0 : \theta_t - \xi_t = \mu \quad vs. \quad H_1 : \theta_t - \xi_t \neq \mu ,$$

that is, whether or not the target parameters of the SMI and CCI differ by a constant over time. Or one can apply the procedure below on the log-scale to test if θ_t/ξ_t is a constant.

For our purpose here, we shall make a simplifying assumption that $\tau_t^2 = 0$ and thereby remove the conceptual distinction between SMI as an estimator and its theoretical target ξ_t. In light of the large amount of posts in $s_{P,t}$, the assumption seems plausible. It follows then from (1) and (2) that, under H_0, we have

$$X_t = \mathrm{CCI}_t - \mathrm{SMI}_t = \mu + e_t ,$$

where $e_t \sim N(0, \sigma_t^2)$. Thus, one may compare the total deviation of X_t from its mean $\bar{X} = \sum_{t=1}^{T} X_t$, over the available T time points, to the variances of the CCI: the larger the total deviation exceeds that which is allowed for by the CCI variances, the stronger is the evidence against H_0 compared with H_1.

Formally, let $P = I - \mathbf{1}\mathbf{1}^\top/T$, where I is the $T \times T$ identity matrix and $\mathbf{1}$ is the $T \times 1$ unity vector, and the matrix P is idempotent such that $PP^\top = PP = P$. We have

$$E(PX) = \mathbf{0} \text{ for } X = (X_1, \ldots, X_T)^\top ,$$
$$V(PX) = P\Sigma P \text{ for } \Sigma = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_T^2) .$$

The diagonal matrix Σ corresponds to the assumption that the CCIs are uncorrelated over time. If this is not the case, one may specify the true covariance matrix appropriately, without this affecting the generality of the following development. Now that $\mathbf{1}^\top PX \equiv 0$, one of the component is redundant. Let $X' = (PX)_{(-t)}$ on deleting the t-th component of PX, for any $1 \leq t \leq T$. Let Q be the correspond $(T-1) \times (T-1)$ sub-matrix of $P\Sigma P$, such that X' has the $T-1$-variate normal distribution

$$X' \sim N(\mathbf{0}, Q) .$$

Let $LL^\top = Q$ be the Cholesky decomposition with lower-triangular L, such that

$$L^{-1} Q (L^{-1})^\top = L^{-1} L L^\top (L^{-1})^\top = I_{(T-1)\times(T-1)},$$

and

$$R = L^{-1} X' \sim N(\mathbf{0}, I) .$$

A test statistic for H_0 can thus given as

$$D = R^\top R \sim \chi^2_{T-1} .$$

Owing to confidentiality restrictions, we can only obtain the CCI (from the homepage of Statistics Netherlands), but not the actual values of the SMI, nor the variances of the CCI. The calculations below serve only for the purpose of illustration. Firstly, we eyeball Figure 1 to obtain the approximate values of the SMI, where the empirical correlation coefficient between two series is 0.88 over the 27 months. Next, Figure 2 reproduced from Brakel *et al.* (2017) plots the 95% confidence interval of CCI over 2000–2014, where the coefficient of variation (CV), denoted by $\eta_t = \sigma_t/\mathrm{CCI}_t$, varies approximately between 0.01 and 0.34 over the period relevant to Figure 1. Based on these approximate σ_t^2's, the p-value of the test above is virtually zero, such

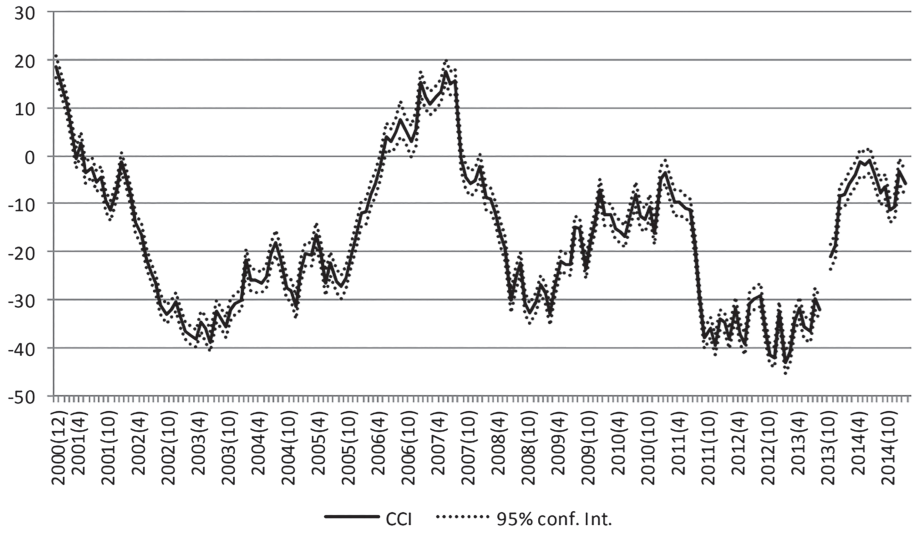

Figure 2. *The consumer confidence index (CCI) series with 95% confidence interval, 2000–2014.*

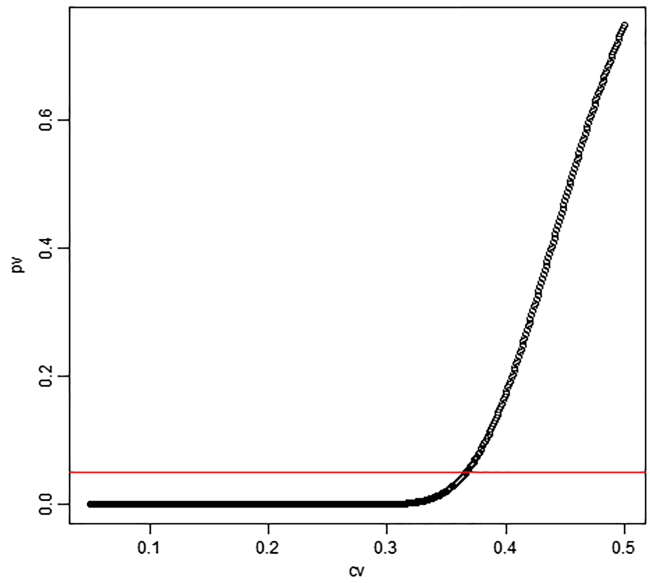

Figure 3. *p-values of test H_0 vs. H_1 for varying coefficients of variation (CVs), level 0.05 mark by horizontal line. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

Color Online, B&W in Print

that H_0 is rejected at the level of 0.05 or much lower. Moreover, for the illustration purpose, here, we stipulate the values of σ_t^2 in relation to the CCI via a constant CV over time, denoted by η, such that $\sigma_t = \eta \mathrm{CCI}_t$. Figure 3 shows the p-value of the test as η varies from 0.05 to 0.5, where the p-value exceeds 0.05 for $\eta > 0.367$. In other words, unless the CV of the CCI is larger than 36.7% for all the 27 months, which is of concern here, the null hypothesis is rejected at the level of 0.05.

3.4 Discussion

Firstly, in the above, we have considered the validity of the SMI, assuming the aim is to replace the CCI with it. Of course, even if the SMI cannot do this directly, there is still the possibility to use it to improve the CCI. Brakel *et al.* (2017) study the two indices over time using a bivariate time series model:

$$\begin{pmatrix} Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} L_t^Y \\ L_t^Z \end{pmatrix} + \begin{pmatrix} S_t^Y \\ 0 \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ 0 \end{pmatrix} + \begin{pmatrix} \upsilon_t^Y \\ \upsilon_t^Z \end{pmatrix},$$

where Z_t is the SMI that is decomposed into trend L_t^Z and an error term υ_t^Z, and Y_t is the CCI that is decomposed into trend L_t^Y, seasonal component S_t^Y, an error term υ_t^Y and $\beta^{11}\delta_t^{11}$ that is an outlier term introduced to accommodate the economic downturn at the corresponding time point. The authors find that using the SMI series as an auxiliary series slightly improves the precision of the model based estimates for the CCI, at a time when the SMI for the current month is available but not the CCI—owing to the longer production lag required for the latter. Notice that such uses of social media data as the auxiliary information for survey sampling do not pose any new theoretical challenges.

Next, disregarding the distinction between $\theta_{s,t} = \theta(y_{s_t})$ and the CCI target $\theta_t = (y_{U_t})$, where one faces a difficulty of representation between s_t and U_t, there is a question whether the SMI (1) appropriately targets the 'intermediary' parameter $\theta_{s,t}$. As remarked by Brakel *et al.* (2017), the CCI survey questions involve the amount of purchases of expensive goods during the last 12 months and the tendency of households to buy expensive goods. It seems relevant to utilise internet search data and actual purchase data of such expensive goods. The implication is that one needs not to rely exclusively on social media data for content extraction but could seek to combine them with other non-survey data. On the one hand, combining data to improve content extraction seems desirable regarding the quality of measurement. On the other hand, doing so is likely to affect the representation dimension of data quality, as previously noticed in Section 2.2. But the quality of representation is worth examining in any case. In the current definition of SMI (1), each post is given the same weight. It is unclear whether this is the most appropriate treatment, because the number of posts per account or user is likely to vary in different subsets of s_t. Indeed, provided a method of differential weighting of the posts in $s_{P,t}$ can be justified with respect to $\theta(y_{s_t})$, targeting $\theta(y_{U_t})$ may no longer be as elusive as it is currently.

Finally, despite our focus in this paper on target parameter θ defined for (y_i, U), it is conceivable that one may be interested in target parameter ξ defined for (z_j, P) directly. In such situations, the quality considerations are analogous to those in the case of targeting θ based on a sample s, for $s \subset U$ and the associated measures $y_s^* = \{y_i^*; i \in s\}$. A basic issue regarding representation is the fact that the sample s_P is not selected from the totality P according to a probability sampling design. Inference from non-probability samples has received much attention. See, for example, Smith (1983), Elliott & Valliant (2017) and Zhang (2019) for inference approaches assuming non-informative selection of the observed sample; see, for example, Rubin (1976) and Pfeffermann *et al.* (1998), of approaches that explicitly adjust for the informative selection mechanism. When it comes to the measurement dimension of data quality, the traditional treatment of measurement errors in surveys (e.g. Biemer *et al.*, 2004) may be less relevant because, as discussed in Section 2.2, content, feature or network extraction from social media data faces quite different challenges and uses quite different techniques than data collection via survey instruments.

4 TWO-PHASE APPROACH

In the two-phase approach, one aims to estimate the target parameter $\theta = \theta(y_U)$ based on a pseudo survey dataset constructed from the sample of social media data to resemble a survey dataset from the target population. Denote by s_{AP} the sample of statistical units in the pseudo survey dataset, and by y_i^* the constructed proxy to y_i for $i \in s_{AP}$.

The quality of the pseudo survey dataset (y_i^*, s_{AP}) with respect to the ideal census data (y_i, U) can be assessed with respect to representation and measurement, under the quality framework of Groves *et al.* (2004) for traditional sample survey data. The key extra concern is the necessary transformation from the initial social media data, which is a process that does not exist for sample survey data. Zhang (2012) outlines a two-phase life-cycle model of statistical data before and during integration, respectively, which includes the transformation from multiple first-phase input datasets to the ones to be integrated at the second phase. The total-error framework of Zhang (2012) is applicable as well to the two-phase approach to statistical analysis based on social media data.

Below, we examine the study of Swier *et al.* (2015), which aims to construct pseudo survey datasets of residence and mobility from geo-located tweets. In particular, this illustrates the generic transformation process under the two-phase approach: from the first-phase data objects (posts) to the second-phase statistical units (persons) in terms of representation, and from values obtained at the first phase (e.g. the geolocation of a post) to the second-phase statistical variable (e.g. location of residence) in terms of measurement. Moreover, we analyse the quality of the resulting pseudo survey dataset according to the total-error framework of Zhang (2012), and we highlight some relevant methodological challenges.

4.1 Case: Residence Location from Tweets

Swier *et al.* (2015) conducted a pilot study at the Office for National Statistics, on the potential of Twitter to provide residence and mobility data for official statistics. The main efforts concerned the construction of relevant pseudo survey datasets, which we summarise below. In addition, some simple analyses were performed, giving indications of the possible target parameters envisaged. We do not explicitly discuss these analyses here.

There are two first-phase input datasets. The first one is collected via the Twitter `Streaming API`, covering the period 11 April to 14 August 2014. The search criteria involve a set of bounding rectangles covering the British Isles, for which a tailor-made application is developed and deployed. The second dataset is purchased from GNIP (a reseller of data, now owned by Twitter), covering the period 1 to 10 April and 15 August to 31 October 2014. Unlike the API data, the GNIP data are filtered by tweets with a 'GB' country code. The tweets from the same period, which cannot be geo-located in either way, are excluded.

Next, the two datasets are merged to create a single dataset, during which a number of tweets are removed. These include the ones that are detected to be generated by bots, or without exact GPS location, or non-GB tweets in the first dataset (mainly those from the Republic of Ireland). In particular, for privacy protection reasons, any tweet from the first dataset is removed, unless it is associated with an account in the purchased GNIP data. All the retained tweets have latitude and longitude (GPS) coordinates.

The process of merging can therefore equally be represented as in the life-cycle model of integrated data (Zhang, 2012), where linkage of separate datasets is carried out via the second-phase units associated each input datasets. In other words, one may first identify the associated Account IDs (second-phase units here) in the API and GNIP datasets, respectively, and then

merge the data for the same Account ID, provided it is present in the GNIP dataset. In this case, one could merge the datasets before transforming the data organised around Tweet ID to Account ID, because the two first-phase datasets share the same identifiable objects (i.e. tweets with Tweet ID).

In this way, at the beginning of the second-phase processing, one obtains a single set of GB-located tweets (81.4 million over 7 months) and the associated accounts. No further second-phase data processing takes place in the representation dimension. For instance, one does not attempt to identify and classify the users behind the observed accounts. Second-phase processing in the measurement is primarily concerned with content extraction of residential location and its classification. This is carried out in the following steps.

- The tweets associated with a given account are *clustered*, using the density-based spatial clustering algorithm with noise (DBSCAN). It groups together points that are closer to each other in terms of spatial density; the cluster formed is regarded valid only if it contains a specified minimum number of points. The points in clusters below the minimum threshold are considered as noise. Of the 81.4 million tweets, 67.4 million are included in one or another cluster that contains three or more tweets. The rest of the clusters with only one or two tweets are classified as 'invalid'.
- Next, each valid cluster is classified as 'residential', 'commercial' or 'others' in terms of address type, using the AddressBase that is the definitive source of address information for Great Britain. To this end, one calculates a weighted centroid of the cluster and finds the closest property to it in the AddressBase. The cluster address type is then classified according to this 'nearest neighbour' property.
- Then, for each account with one or several residential clusters, one of them with the most tweets is classified as the 'dominant' residential cluster.
- Finally, additional classification may be attached to each cluster, such as the administrative geography it belongs to, the number of tweets it contains, the time span of these tweets (short term if less than 31 days vs. long term otherwise).

4.2 Quality Assessment

Before we assess the quality of the pseudo survey dataset (y_i^*, s_A) obtained under the two-phase approach when targeting θ defined for (y_i, U), it is helpful to recapitulate some of the relevant technical issues, even if they do not account for all the sources of errors.

Firstly, some additional API data are actually collected on 10 April and 15 August, which overlap with the GNIP data on these two days. A small number of API tweets are found not be included in the GNIP set, all of which are associated with protected accounts—users may opt to protect their accounts so that their tweets can only be viewed by approved followers. More generally, retrospective changes made by a user to its account or specific tweets may prevent them from being included in the historic point-in-time data available from GNIP, despite these accounts or tweets are accessible via the real-time `Streaming` API. This exemplifies a general cause for discrepancy between Twitter data collected in different ways. Two other examples of general causes are as below.

Filter criteria The filter criteria may not be fully compatible between the APIs and the data brokers. As explained above, in the case here, the geographic filter works differently with the `Streaming API` and GNIP.

Missing data Data from APIs may be missing owing to technical problems, such as moving of IT equipment or broadband router failure.

Next, once the data from the first phase have been merged and transformed, there are generally technical issues with data extraction and processing that are necessary at the second phase. In this case, the DBSCAN clustering of tweets is an unsupervised machine learning technique, for which it is generally difficult to verify the truthfulness of the results. The address type classification is in principle a supervised learning technique. However, it may be resource demanding to obtain a training-validation dataset, by which the classification method can be improved and its accuracy evaluated, similarly for the classification of the dominant residual cluster.

The quality of the dataset (y_i^*, s_A) can be assessed according to the second-phase life-cycle model (Figure 4), along the two dimensions of representation and measurement. The exact nature of the potential errors needs to be related to the envisaged analysis. Below, we consider first representation and then measurement.

In terms of representation, the 'Linked Sets' in Figure 4 is given by $b(s_A)$, which is subjected to coverage errors. Over-coverage is the case if $b(s_A) \setminus U \neq \varnothing$. This is unavoidable here because some of the accounts in $b(s_A)$ are not persons at all and all the bots are not completely removed. Moreover, there may be multiple accounts in s_A that correspond to the same person; such duplicates are another form of over-coverage error. Whether s_A entails under-coverage depends on the assumption. For instance, let the target population U be the adult residents of England. If one assumes that in principle there is an unknown but non-zero probability for everyone in U to have a Twitter account and to have tweeted at least three times from the same location during the 7 months in 2014, then there would be no under-coverage error of $b(s_A)$ for U, but only a non-probability selection issue. However, insofar as these assumptions are untenable, then there would be an under-coverage error in addition.

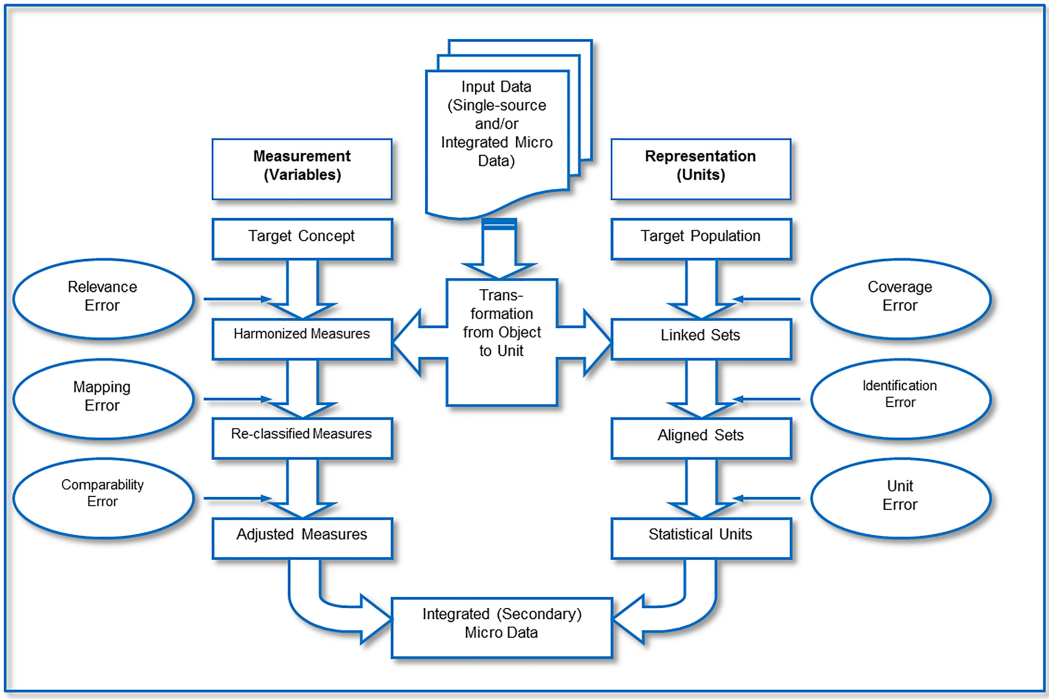

Figure 4. *Phase two life-cycle model of Zhang (2012). [Colour figure can be viewed at* wileyonlinelibrary.com*]*

Color Online, B&W in Print

Next, the identification error may be an issue if domain classification of the target population needs to be based on feature extraction, which is prone to errors, whereas unit error is potentially troublesome if additional statistical units (e.g. household) need to constructed. Neither seems relevant to any of the analyses of Swier *et al.* (2015).

In terms of measurement, an example of 'Harmonized Measures' in Figure 4 is the dominant residential cluster here. Suppose the 'Target Concept' is the de facto place of residence of a person. Relevance error is mostly like the case, unless everyone sends most tweets from her or his de facto place of residence. Or suppose the 'Target Concept' is whether a person is a tourist, and short-term versus long-term classification of the dominant residential cluster is used as a proxy measure of the corresponding person. Again, relevance error is mostly like the case, unless no tourist stays longer than a month and no usual resident stops tweeting after less than a month.

Next, the mapping error is, for example, the case when someone does tweet from her or his de facto place of residence but the clustering-classification algorithm fails to identify it as the dominant residential cluster. This can happen, for example, if the person tweets more when at her or his friend's place, or if the person more often than not switches off GPS location when tweeting at home, or if the person's home is in a dense area and the chosen nearest neighbour property in the AddressBase happens to be a commercial address. Finally, the comparability error could arise if, for example, the classified dominant residential cluster is further adjusted in light of other available measures, although this is not the case in the study of Swier *et al.* (2015).

In summary, the main errors of the pseudo survey dataset (y_i^*, s_A) here are coverage errors in terms of representation, and relevance and mapping errors in terms of measurement.

4.3 Discussion: Statistical Analysis

In the above, we outlined the data processing required under the two-phase approach to social media data, using the study of Swier *et al.* (2015) as the case in point. It is shown that the life-cycle model of (Zhang, 2012) can be applied as a total-error framework for evaluating the quality of the resulting pseudo survey dataset (y_i^*, s_A), where $s_A = a(s_P)$. The study of Swier *et al.* (2015) does not specify any definitive target of analysis. For a discussion of possible statistical analysis of the target parameter θ defined for (y_j, U), let us consider two situations, depending on whether it involves additional datasets or not.

Consider the situation where only the pseudo survey dataset (y_i^*, s_A) is to be used for an analysis targeted at $\theta(y_U)$. The first key issue regarding representation is over-coverage adjustment, from $s' = b(s_A)$ to $s = U \cap b(s_A)$, because $s' \setminus U \neq \varnothing$. This could be based on the mapping either from s' to s or, provided it can be specified, from $t(y_{s'}^*)$ to $t(y_s^*)$, where $t(\cdot)$ denotes the sufficient statistics for θ. Given the over-coverage adjustment, the remaining issues are non-probability representation of s for U, and measurement discrepancy between y_i^* and y_i caused by lack of relevance and imperfect data extraction, similarly to what has been discussed earlier in Section 3.4.

A potentially more promising scenario is to utilise additional datasets, in order to overcome or reduce the deficiency of each dataset on its own. Integration with other Sign-of-Life data can possibly improve the quality of the pseudo survey dataset constructed from social media data. For example, in the case of data for residence and mobility, other Sign-of-Life data on employment, education, utility services and so forth can probably improve the classification of the dominant residential cluster, provided these data are available and can be combined with the tweets data. However, it is also possible that one cannot always overcome the inherent deficiencies of social media data in this way. Making statistics based on multiple sources is a broad challenging topic. It is currently an area of active research and development. See, for example,

De Waal *et al.* (2017) and Di Zio *et al.* (2017) for overviews of related situations and methodological issues. See Zhang (2018) for an overview of estimation methods in the presence of multiple proxy variables.

5 CONCLUDING REMARKS

In the above, we systematically delineated two existing approaches to statistical analysis based on social media data. The fundamental challenge with the one-phase approach in some situations is a lack of analytic connection to the target parameter, which is defined for a different set of units and another associated measure. Nevertheless, external data can in principle be used to verify the statistical validity of this approach. Compared with observational studies based on data subjected to non-probability selection and survey measurement errors, the key extra issues with the two-phase approach revolve around the transformation process from the initial data objects to the statistical units of interest and the algorithmic data extraction required for measurement. In addition, an explicit adjustment for the over-coverage error will be needed in many situations.

For assessment of data quality, we have demonstrated that it is possible to apply relevant total-error frameworks formulated in terms of representation and measurement of generic statistical data. In particular, for both approaches, it seems more promising if one does not simply restrict oneself to the available social media data but seeks to combine them with additional relevant datasets, in order to overcome or reduce the deficiency of each source, despite data integration is by no means a straightforward undertaking in general.

We would like to close with a few remarks. Firstly, in the paper, we have focused on target parameters that are finite-population functions. Such a parameter is often referred to as a descriptive target, in contrast to analytic target parameters that can never be directly observed, regardless of how large the observed number of units and how perfect the obtained measurement may be. For example, the ordinary least squares fit of some specified linear regression coefficients based on a perfect census of the current population is a descriptive target parameter; at the same time, it is an estimate of the theoretical (or super-population) values of these coefficients of the postulated regression model, that is, the analytic target parameter in this case. Our focus on descriptive target parameters helps to simplify the exposition, because the differences between descriptive and analytic inference can be subtle and many but are nevertheless not critical to our aim in this paper. See, for example, Skinner *et al.* (1989), Chambers & Skinner (2003) and Skinner & Wakefield (2017) for introductions to analytic versus descriptive inference based on sample surveys.

Next, there are certainly many similarities to statistical analysis based on administrative data. As we have demonstrated, the total-error framework (Zhang, 2012) for statistical data integration involving administrative sources is applicable as well to the two-phase approach based on social media data. It is worth reiterating the two extra difficulties in comparison. The first one relates to the transformation from the original data objects P to the statistical units U. The same requirement exists equally for administrative data in general. For instance, exams are part of the initial education data objects. However, while the transformation from exams (say, P) to students (say, U) can be carried out unproblematically by the school administration, such straightforward processing is often impossible from social media data objects to the target population of interest. The second extra difficulty concerns data extraction. The available measures in the administrative sources do often suffer from relevance error. Nevertheless, the actual mapping to the 'Re-classified Measures' (Figure 4) seldom requires content or feature extraction

that are necessary for social media data, which, as has been discussed, is generally an additional cause of discrepancy between y_i^* and y_i or between z_j and y_i.

Finally, there seems to be currently an under-explored potential regarding the rich network relationships that can be extracted from social media data. Such network relationships may be difficult to obtain via traditional survey methods, both due to the limitations of the usual survey instruments and the relatively high cognitive and memorial requirements for correct information retrieval by the respondents. In contrast, for network relationships that are directly observable on the social media platform, no subjective information processing will be needed, and the errors associated with such processing are thereby avoided. Making greater use of the network relationships in social media data and developing suitable sampling and analysis methods appear fruitful venues forward, in order to harness the opportunities that have emerged with such big data sources.

References

Berzofsky, M., Mckay, T., Hsieh, Y. & Smith, A. (2018). Probability-based samples on Twitter: methodology and application. *Surg. Pract.*, **11**, 1–12. https://doi.org/10.29115/SP-2018-0033

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. & Sudman, S. (2004). *Measurement Errors in Surveys*. John Wiley & Sons. https://doi.org/10.1002/9781118150382

Blank, G. & Lutz, C. (2017). Representativeness of social media in Great Britain: investigating Facebook, Linkedin, Twitter, Pinterest, Google+, and Instagram. *Am. Behav. Sci.*, **61**, 741–756. https://doi.org/10.1177/0002764217717559

Boyd, D. & Crawford, K. (2012). Critical questions for big data. *Inf. Commun. Soc.*, **15**(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brakel, J., Söhler, E., Daas, P. & Buelens, B. (2017). Social media as a data source for official statistics; the Dutch consumer confidence index. *Surv. Methodol.*, **43**, 183–210. https://doi.org/10.13140/RG.2.2.19294.64326

Bright, J., Margetts, H., Hale, S. & Yasseri, T. (2014). *The Use of Social Media for Research and Analysis: A Feasibility Study Technical Report*. Department for Work and Pensions: United Kingdom.

Chambers, R.L. & Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470867205

Daas, P.J., Burger, J., Le, Q., Ten Bosch, O. & Puts, M.J. (2016). *Profiling of Twitter Users: A Big Data Selectivity Study. Technical Report 201606*. The Hague/Heerlen: Statistics Netherlands.

Daas, P.J. & Puts, M.J. (2014). *Social Media Sentiment and Consumer Confidence. Series No 5; European Central Bank Statistics Paper*. Frankfurt: Germany.

Daas, P.J., Puts, M.J., Buelens, B. & van den Hurk, P.A. (2015). Big data as a source for official statistics. *J. Off. Stat.*, **31**(2), 249–262. https://doi.org/10.1515/jos-2015-0016

Daas, P., Roos, M., Van de Ven, M. & Neroni, J. (2012). Twitter as A Potential Data Source for Statistics, The Hague/Heerlen.

De Waal, T., van Delden, A. & Scholtus, S. (2017). Multi-source Statistics: Basic Situations and Methods, The Hague/Heerlen.

Di Zio, M., Zhang, L.-C. & de Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *Surv. Stat.*, **76**, 17–26.

Elliott, M. & Valliant, R. (2017). Inference for nonprobability samples. *Stat. Sci.*, **32**, 249–264. https://doi.org/10.1214/16-STS598

Gaffney, D. & Puschmann, C. (2013). Data collection on Twitter. In *Twitter and Society*, pp. 55–67. https://doi.org/10.3726/978-1-4539-1170-9

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J. & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Soc. Networks*, **38**, 16–27. https://doi.org/10.1016/j.socnet.2014.01.004

Greenwood, S., Perrin, A. & Duggan, M. (2016). *Social Media Update November 2016*, Pew Research Centre.

Groves, M.R. (2011). Three eras of survey research. *Public Opin. Q.*, **75**, 861–871. https://doi.org/10.1093/poq/nfr057

Groves, M.R., Fowler, J.F. Jr., Couper, M.P., Lepkowski, J.M., Singer, E. & Tourangeau, R. (2004). *Survey Methodology*. John Wiley & Sons.

Halford, S., Weal, M., Tinati, R., Pope, C. & Carr, L. (2017). Understanding the production and circulation of social media data: towards methodological principles and praxis. *New Media Soc.*, **20**(9), 3341–3358. https://doi.org/10.1177/1461444817748953

Hsieh, Y.P. & Murphy, J. (2017). Total Twitter error: decomposing public opinion measurement on Twitter from a total survey error perspective. In *Total Survey Error in Practice: Improving Quality in The Era of Big Data*, pp. 23–46. https://doi.org/10.1002/9781119041702.ch2

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opin. Q.*, **79**(4), 839–880. https://doi.org/10.1093/poq/nfv039

Mellon, J. & Prosser, C. (2016). Twitter and Facebook are not representative of the general population: political attitudes and demographics of social media users. *SSRN Electron. J.*, **4**(3), 1–9. https://doi.org/10.2139/ssrn.2791625

Morstatter, F., Pfeffer, J., Liu, H. & Carley, K.M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. arXiv:1306.5204.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, **2**, 1–135. https://doi.org/10.1561/1500000011

Pfeffermann, D., Krieger, A. & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Stat. Sin.*, **8**, 1087–1114.

Rampazzo, F., Zagheni, E., Weber, I., Testa, M. & Billari, F. (2018). Mater certa est, pater numquam: what can Facebook advertising data tell us about male fertility rates? arXiv:1804.04632.

Rebecq, A. (2018). Extension sampling designs for big networks: application to Twitter. In *Springer Proceedings in Mathematics & Statistics*, pp. 251–270. https://doi.org/10.1007/978-3-319-96941-1_17

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592. https://doi.org/10.2307/2335739

Skinner, C.J., Holt, D. & Smith, T.M.F (1989). *Analysis of Complex Surveys*. Wiley.

Skinner, C.J. & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Stat. Sci.*, **32**, 165–175. https://doi.org/10.1214/17-STS614

Smith, T.M.F. (1983). On the validity of inferences from non-random sample. *J. R. Stat. Soc. Ser. A*, **146**(4), 394–403. https://doi.org/10.2307/2981454

Swier, N., Komarniczky, B. & Clapperton, B. (2015). *Using Geolocated Twitter Traces to Infer Residence and Mobility*, GSS Methodology Series.

Tabassum, S., Pereira, F.S.F., Fernandes, S. & Gama, J. (2018). Social network analysis: an overview. *Wiley Interdiscipl. Revi. Data Mining Knowl. Disc.*, **8**(5), e1256. https://doi.org/10.1002/widm.1256

Wang, Y., Callan, J. & Zheng, B. (2015). Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Trans. Web (TWEB)*, **9**(3), 13. https://doi.org/10.1145/2746366

Yan, W., Wenchao, Y., Sam, L. & Sean, D.Y. (2019). The relationship between social media data and crime rates in the united states. *Social. Med. Soc.*, **5**(1), 1–9. https://doi.org/10.1177/2056305119834585

Yildiz, D., Munson, J., Vitali, A., Tinati, R. & Holland, J.A. (2017). Using Twitter data for demographic research. *Demogr. Res.*, **37**(46), 1477–1514. https://doi.org/10.4054/DemRes.2017.37.46

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Stat. Neerlandica*, **66**(1), 41–63. https://doi.org/10.1111/j.1467-9574.2011.00508.x

Zhang, L.-C. (2018). On the use of proxy variables in combining register and survey data. In *Administrative Records for Survey Methodology*. Wiley Series in Survey Methodology.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Stat. Theor. Relat. Field.* https://doi.org/10.1080/24754269.2019.1666241

[*Received December 2018, Revised April 2020, Accepted July 2020*]