METRON

Publishing model: **Hybrid**

← Back to overview

Search all METRON articles →

Volume 82, Issue 1

April 2024

Special Issue: Survey Methods for Statistical Data Integration and New Data Sources: tools and real data applications for official statistics

Issue Editors: Maria Giovanna Ranalli, Jean-François Beaumont, Gaia Bertarelli, Nathalie Shlomo

9 articles in this issue

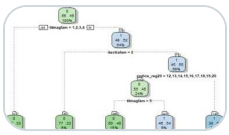
Foreword to the special issue on "Survey Methods for Statistical Data Integration and New Data Sources: tools and real data applications for official statistics"

Editorial | 19 March 2024 | Pages: 1 – 3

Improving the design of the Italian permanent population and housing census: a transition towards a massive use of administrative data

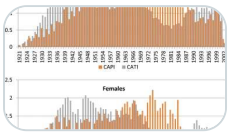
OriginalPaper | 30 November 2023 | Pages: 5 – 17

Adaptive sampling design for the Italian social sample surveys: an application on the population census



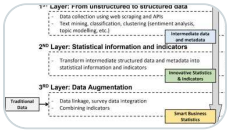
OriginalPaper | 10 January 2024 | Pages: 19 – 35

If the tools to gather information affect data quality: violence against women survey case



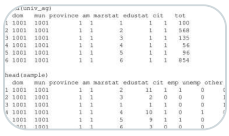
OriginalPaper | 11 March 2024 | Pages: 37 – 70

Augmenting business statistics information by combining traditional data with textual data: a composite indicator approach



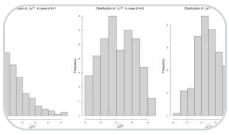
OriginalPaper | Open access | 13 January 2024 | Pages: 71 – 91

mind, A methodology for multivariate small area estimation with multiple random effects



OriginalPaper | 23 November 2023 | Pages: 93 – 107

How the sampling variances affect the linear predictor of the Fay–Herriot model



OriginalPaper | 21 June 2023 | Pages: 109 – 130

Correction to: Foreword to the special issue on "Survey Methods for Statistical Data Integration and New Data Sources"

Correction | 01 February 2024 | Pages: 131 – 131

Publisher Correction: Mind, a methodology for multivariate small area estimation with multiple random effects

Publisher Correction | 22 January 2024 | Pages: 133 – 133

METRON (2024) 82:71–91
https://doi.org/10.1007/s40300-023-00261-4

Augmenting business statistics information by combining traditional data with textual data: a composite indicator approach

Camilla Salvatore[1] · **Silvia Biffignandi**[2] · **Annamaria Bianchi**[3]

Received: 18 November 2022 / Accepted: 7 November 2023 / Published online: 13 January 2024

Abstract
Combining traditional and digital trace data is an emerging trend in statistics. In this respect, new data sources represent the basis for multi-purpose extraction of different statistical indicators, which contribute to augmenting the statistical information, for feeding smart statistics. The production of business statistics can benefit from the use of unstructured data, especially to study novel aspects which are not covered by traditional data sources. This paper proposes a methodological general framework for augmenting information by combining data, both structured and non structured. The statistical challenges of using unstructured data and their integration with traditional data are discussed. The methodological general framework is applied to the construction of smart composite indicators using social media data and their metadata. An empirical exercise illustrates how to apply the methodology in practice.

Keywords Socio-economic indicators · Mazziotta–Pareto index · Sustainable development · Social media · Twitter

1 Introduction

The cost of collecting and processing high quality traditional data, such as surveys, is increasing, and the process of deriving statistical products from this data is demanding and time-consuming [1].

At the same time, the availability of new data has led to an expansion of data collection methods, moving beyond traditional primary data collection to the extraction of statistics from non-traditional sources. These sources, referred to as big, or digital trace/behavioral data, include, among others, social media posts, Google trends and mobile phone data (i.e., location, photos, and other sensor data), and are produced by human online/digital behaviors

✉ Camilla Salvatore
c.salvatore@uu.nl

[1] Department of Methodology and Statistics, Utrecht University, Sjoerd Groenmangebouw, Padualaan 14, 3584 CH Utrecht, The Netherlands

[2] Consultant in Economic Statistics Studies, Bergamo, Italy

[3] Department of Economics, University of Bergamo, Bergamo, Italy

and interactions [2]. Digital trace data are not generated for statistical purposes but can serve as a convenient and timely source of information for understanding and measuring (new) complex socio-economic phenomena [3]. These new data sources provide a basis for the multi-purpose extraction of different statistical indicators, which complement the traditionally available statistical information and feed smart statistics [4–6]. The integration of traditional and digital trace data for producing innovative statistics and indicators is a promising approach. This can enhance the timeliness, providing a finer spatial and temporal resolution, a higher level of detail, new perspectives, and new insights on phenomena, while also reducing the production cost of (official) statistics [7].

Research on indicators constructed using non-traditional sources only, particularly textual data from social media, is prevalent in the literature, especially with reference to social aspects [8–11]. Further, a number of experimental statistics have been developed by National Statistical Institutes (NSIs) using such textual data to study social tensions[1] and consumers' confidence on the economy (see, for example, Daas and Puts [12] and the Istat's Social Mood on Economy Index[2]). However, studies combining traditional and digital trace data-based indicators are scarce.

The use of innovative unstructured data, also combined with traditional data, is relatively underdeveloped in the field of business statistics, despite the potential benefits they can offer. New data sources can be used in a variety of ways, including enhancing the information for a given unit [13]. For example, Statistics Canada used sensor data to augment administrative data and produce more efficient small area estimates for business statistics [14]. Similarly, Statistics Netherlands (CBS) is committed in enhancing business statistics, using web-scraped data from companies' websites in order to detect innovative companies and improve the quality of the appointed NACE codes [15, 16]. The Italian National Statistical Institute (ISTAT) is also committed in developing experimental statistics based on businesses' websites in order to identify their activities or to augment the information collected through the traditional survey on Information and Communication Technologies [17–19]. Also researchers can be interested in experimenting various methodologies combining multiple data sources.

In this paper, we propose a general methodological framework for the construction of smart statistics. The framework is developed following a modular approach for combining the use of digital and unstructured data (and relative metadata) together with traditional data. We consider the field of business statistics and the specific case of constructing composite indicators (CIs) by combining traditional and innovative (e.g. social media or web-based) indicators. An original aspect is that we propose to process metadata[3] in order to build innovative indicators. Processing metadata is an emerging aspect in the analysis of digital trace data and existing experiences rely mainly on checking and improving the quality of the metadata, whereas the computation of indicators based on metadata is a novel contribution.

To the purpose of providing an example to researchers, we develop an illustrative exercise to demonstrate how to implement the proposed method. It serves as a prototype application which shows the steps to be undertaken to build up new, innovative, indicators based both on unstructured and structured data. In our exercise, we consider a commercial database as traditional source for structured data and Twitter as new data source for unstructured data. We

[1] https://www.cbs.nl/en-gb/about-us/innovation/project/social-tensions-indicator-gauging-society.

[2] https://www.istat.it/en/experimental-statistics/experiments-on-big-data.

[3] In order to avoid confusion, we clarify the use of the term metadata in the context of digital unstructured data. It differs from the definition used in statistics, i.e., the information that is required in order to interpret and use statistics. In this context, metadata refers to additional information about the main data of interest. In Twitter, for example, the tweet represents the main data and the date of publication, likes, links, and images are metadata.

focus on the case where data about the same units are available in both sources. However, a similar approach can be adopted at a more aggregate level, namely in the case such individual information is not available.

The reminder of the article is the following. Section 2 discusses the challenges of constructing smart business statistics. Section 3 presents a modular general architecture for the construction of such statistics and the framework to build CIs. Section 4 illustrates the practical exercise on the construction of a prototype indicator. The importance of quality evaluations together with emerging aspects related to the multi-source nature of the integration, is discussed in Sect. 5. Conclusions are in Sect. 6 together with avenues for future research.

2 Challenges of augmenting business statistics with unstructured data

Traditionally, business statistics are derived from survey data, like the European Company Surveys[4], the Business and consumer surveys (BCS)[5] and other surveys carried out by NSIs. In these cases, the data are structured, the data-generating process is under the researchers' control, and errors are allocated along the whole survey process according to the Total Survey Error (TSE) framework [20]. Consequently, surveys are considered as a high-quality data source for business statistics.

Alongside surveys, other popular sources for business statistics are administrative or commercial business data [21]. These are still structured data, quality is checked and improved when necessary. However, these data are not primarily collected for statistical or research purposes. For that reason, they are usually referred to as secondary data. Business registers, documents from local authorities (e.g., tax authority), and law-mandatory reporting are all examples of administrative data. Commercial business data are provided by private companies, for example, Bureau van Dijk,[6] Bloomberg,[7] and Refinitiv.[8]

More recently, the digital transformation has resulted in the emergence of new sources for business and economic statistics [22]. For example, social media posts, annual reports, businesses websites and newspaper articles can be used to study new aspects or gain additional information about companies. In this respect, the production of statistics using traditional data enhanced with new data available from digital sources are referred to as smart statistics [7]. One of the advantages of smart statistics is the ability to augment the information, thereby providing richer insights into the topic of interest. However, there are also several challenges to be considered. Given the wide range of new sources of data, each of them having specific methodological issues, it is not possible to develop a general frame of reference.

In the following, we discuss the main challenges that one encounters when augmenting traditional data with innovative ones, focusing our attention on unstructured textual data. To begin with, it is necessary to extract the data of interest using, for example, web-scraping or Application Programming Interfaces (APIs). Online data are not static. Hence, during data extraction, researchers must be aware of issues pertaining to the changes in data over time, coverage, reliability, and validity of the data, among others. Social media posts, for instance, can be modified or deleted over time, and related metadata can also change (e.g.,

[4] https://www.eurofound.europa.eu/surveys/european-company-surveys.

[5] https://ec.europa.eu/eurostat/web/euro-indicators/business-and-consumer-surveys.

[6] https://www.bvdinfo.com/en-gb/.

[7] https://www.bloomberg.com/.

[8] https://www.refinitiv.com/en/financial-data.

likes, replies, and shares). Therefore, the results may differ based on the timing of retrieval. Similarly, different formulation of the search query in terms of the keyword specified, such as when extracting social media posts or newspaper articles based on firm names or products, can result in the delivery of different data.

Another issue that might arise when one wants to obtain unit level observations is the identification of the correct accounts. For example, when studying the external communication of businesses on social media, not all businesses might be present on social media, or they may have multiple accounts related to specific types of communication (e.g., general communication, promotion and advertisement, business news, clients assistance, recruiting and topic-specific accounts for communicating their socially-responsible behavior). This leads to selection and coverage issues that might affect the quality of the data.

Secondly, unstructured textual data must be transformed into structured data. This can be accomplished in different ways according to the purpose of the analysis. For example, sentiment analysis, topic modeling, and other classification or clustering algorithms can be applied. Moreover, the results might be influenced by the various data cleaning and preprocessing choices [23, 24].

Like survey data, also the analysis of unstructured textual data is susceptible to errors. In this direction, there are efforts being made to adapt the TSE framework to such data, but currently, there is not a general framework in order to account, measure and evaluate errors and data quality [25–27]. Data sources have different characteristics, which require different quality frameworks. The importance of these aspects becomes especially evident when integrating data from different sources, where it is crucial to understand how errors arise, accumulate, and interact during the entire integration process [28]. These are all emerging topics in the literature.

While all these factors should be considered when combining data, our focus here is on proposing a procedure to develop CIs based on the integration of different types of data, structured and unstructured, derived from traditional and non-traditional sources. An overview about quality evaluation is presented in Sect. 5.

3 Methodology

Data integration is becoming increasingly popular as the combination of different sources (e.g., a probability sample surveys with a non-probability one, or a traditional and an innovative—*big data*-source) enables enhanced inference, reduced costs and the measurement of new phenomena or previously unexplored aspects of existing ones [29]. However, when it comes to choosing the right methodology for data integration, a universal approach does not exist [5, 30, 31]. The choice of the methodology depends on various factors, including the research objective (such as finite population or analytic inference, measurement of multidimensional phenomena), availability of variables of interest across sources, similarity or dissimilarity in constructs measured by the two sources, and other relevant considerations. Thus, data integration is statistic and purpose specific.

In Sect. 3.1, we propose a modular general framework for combining traditional and innovative data. The proposed framework is very general and applicable to a variety of scenarios. Next, we address the issue of data integration under the perspective of composite indicators derived from different sources and measuring different aspects of a phenomena. Thus, instead of having a finite population quantity or model parameters to estimate, we consider the task of measuring multi-dimensional phenomena combining indicators from

various sources. Section 3.2 shows how to generate smart business CIs combining structured and unstructured data (e.g. textual data from social media and websites, or other innovative data sources), introducing an adaption of the general framework.

3.1 A modular general framework for the construction of smart business statistics

To produce smart business statistics using unstructured textual data, we develop a modular general framework in three layers. This is an adaption of the modular organization into three layers introduced by Ricciato et al. [7].

In the first layer, the data are collected and transformed into structured data. Such data and related metadata need then to be interpreted by statisticians and serve as input for the second layer. It is worth noticing that the processing of metadata to complement the analysis of unstructured digital data has been examined in a limited number of studies. Indeed, it is an emerging topic and applications relate user/account profiling [32, 33], and geo-spatial applications [34, 35]. As original contribution, we propose to use social media metadata for the construction of CIs as shown in the prototype application (Sect. 4).

In the second block, innovative statistical information is extracted, and indicators are computed. The first and the second layer are augmenting statistical information through the creation of new indicators generated using textual unstructured data.

In the third layer innovative statistics and indicators are used to augment the already available traditional data. Depending on the specific use-case, this can be achieved through methods such as linkage, statistical integration, or by combining indicators. As a result, Smart Business Statistics are produced. Figure 1 summarizes the framework described above.

The modular approach is particularly useful when dealing with new and complex data sources and their integration with traditional ones. Modularity also allows researchers and practitioners to explore other methodological variants (instances) within the same methodological architecture, and possibly propose improvements to specific modules or test sensitivity of the obtained results.

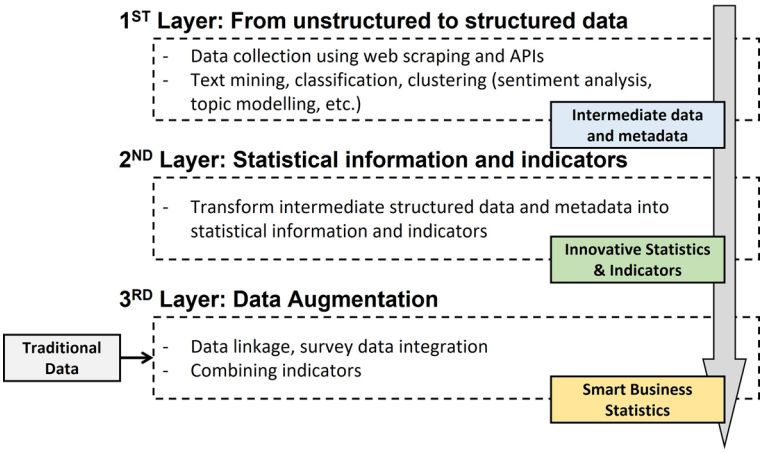


Fig. 1 Modular general framework for producing smart business statistics

3.2 Building smart composite indicators

3.2.1 Background concepts

Before describing the proposed methodology, we shortly remind that when constructing CIs, it is necessary to consider and take decisions on different aspects [36].

First of all, the theoretical framework of the substantial research topic has to be defined. This is crucial for the choice of the data and the variables' definition. It is also important to guide the researcher in the construction process of the CI with respect to methodological decisions related to the normalization of the indicators and the aggregation strategy. Normalization is performed in order to ensure comparability. Based on the variable type (e.g., continuous, categorical, or ordinal) and the aggregation strategy, this can be accomplished in a variety of ways. Common methods are the standardization (z-score), min–max transformation (or re-scaling) or the transformation to index numbers [37].

Aggregation refers to the combination of the individual indicators in order to create a CI. This phase entails considerations on the polarity and the importance of each elementary indicator and the identification of the technique to synthesize the elementary indicators. To properly insert the original indicators into the aggregation procedure, the polarity of indicators should be carefully considered. It refers to the direction of the relationship between the indicator and the phenomenon to be measured. The polarity is positive (negative) if the dimension is positively (negatively) associated to the phenomenon. Further, the selection of the aggregation technique depends on the level of compensability of the individual indicators, which refers to the possibility of balancing a disadvantage on some indicators with a sufficiently large advantage on others. This should be based on theoretical evaluations. In this respect, there are three types of aggregation approaches depending on the degree of compensability: compensatory, partially compensatory, and non-compensatory. For example, full-compensatory aggregation is obtained with the arithmetic mean. In the case of individual indicators from unstructured data, this can be the case of the topic proportions resulting from a topic model. Partially compensatory approaches relate, for example, to the computation of geometric, harmonic, quadratic means, or specific methods like the Mazziotta–Pareto procedure [38]. For example, one could consider the social media dimension related to communication aspects of a certain phenomenon to be partially replaceable with traditional measurements of the same phenomenon. Non-compensatory aggregation is usually performed following multi-criteria approaches.

Aggregation also involves the identification of weights associated to the individual indicators. Weights reflect the relative importance of the indicators to be combined. When no weights are specified, all indicators are implicitly weighed equally. Alternatively, weights can be determined according to subjective and expert evaluations, or statistical methods, such as Principal Component Analysis. However, weights should only be specified when there is a strong theoretical basis for doing so, otherwise a no-weighting strategy should be adopted [39, 40]. Attention should be paid to the implicit importance associated to the original elementary indicators in the case of subsequent aggregations. For a complete overview of CIs construction, please refer to Mazziotta and Pareto [37], OECD [41] and Booysen [39].

When developing CIs, it is important to evaluate the quality of the results taking into consideration the impact of the different methodological decisions that have been made [41]. This topic is further discussed in Sect. 5 also in relation to the multi-source nature of the integration process.

3.2.2 Procedures of the approach

As regards our original contribution, we present a methodology for constructing (a) simple and composite indexes that measure new aspect of phenomena using new data sources and (b) a CI that integrates traditional and non-traditional indexes. We do that by adapting the modular general framework introduced in Sect. 3.1 (see Fig. 1) to this setting.

In empirical studies, it is common to consider several dimensions to represent complex phenomena and to proceed through many levels of aggregation. Our approach is general and offers a flexible solution that can be applied to different cases. In Fig. 2, we present a visual representation of our proposed modular layer approach, emphasizing its practical application by showing an example with two dimensions, one innovative indicator, and three levels of aggregation.

The first layer includes the identification and extraction of the elementary indicators. By way of example and focusing on the innovative data source, assume that, according to the theoretical framework, there are two relevant dimensions that can be measured by the innovative data source, namely D_1 and D_2 and let $I_{D_1,1}, \ldots, I_{D_1,i}, \ldots, I_{D_1,n}$ be the n individual elementary indicators related to dimension D_1 and $I_{D_2,1}, \ldots, I_{D_2,j}, \ldots, I_{D_2,m}$ be m individual elementary indicators related to dimension D_2. Such indicators and dimensions must be identified based on theoretical, empirical, pragmatic, or intuitive considerations [39]. The second layer entails the construction of the innovative index (INN-INDEX). Depending on the specific situation, this can be done as one aggregation or as subsequent aggregation steps. Generally, in the presence of, say, two pillars, the elementary indicators are first combined in order to generate two sub-indicators measuring each dimensions of interest, CI_{D_1} and CI_{D_2} respectively. The approach may be extended to more dimensions depending on the characteristics of the phenomenon and the innovative source being studied. Next, these sub-indicators are further aggregated to create the innovative index (INN-INDEX). This is the second level of aggregation. We assume that the traditional index (TRAD-INDEX) is already available. Otherwise, the same methodology can be applied to obtain a traditional indicator if one does not already exist.

Moving to the third layer, the third level of aggregation relates to the construction of the innovative smart CI (SMART-INDEX). In the second and third levels, attention should be paid to avoid double normalization.

It is important to note that the theoretical framework of the phenomenon being measured plays a crucial role in the construction of the index. All decisions that should be taken at the various step of the three layers and of the CI construction must align with this framework.

Moreover, an advantage of the proposed procedure is that it can be easily adapted to different situations. For example, one might proceed across the whole set of three layers or only compute the CI going through the second and third layer in case the elementary indicators have already been computed.

We illustrate how to apply the proposed methodology through a practical exercise that shows how to construct a prototype CI for measuring Corporate Social Responsibility (CSR) in the next section.

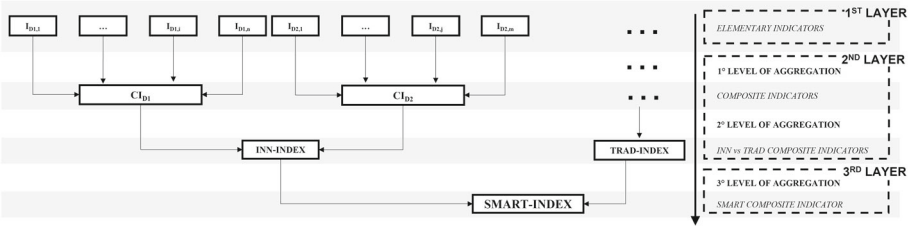


Fig. 2 CI construction strategy on three layers: an example

4 Construction of a prototype

4.1 Context and theoretical framework

This application focuses on the construction of a CI in the field of business statistics and sustainability. Socially responsible behaviors of businesses are linked to the concept of CSR.[9] Given its multi-faceted nature, measuring CSR activities is naturally related to the use of CIs, which allow to summarize complex or multi-dimensional phenomena [43].

The aim of this practical exercise is to measure CSR commitment based on a comprehensive view, including both effective commitment (as traditionally considered) and online communication of CSR-related activities. Online communication is characterised by its content (the topic discussed) and modality (the way it is conducted), leading thus to two pillars. The first one refers to the communication content, i.e., to the text which refers to the communication of CSR activities. The second one refers to communication modality (media richness). This is an important aspect for the communication to be effective and to engage with customers and stakeholders. We expect that the higher the media richness, the more effective the communication will be [44].

Our contribution is to demonstrate how the proposed modular framework can be applied in practice. We show the various steps that should be undertaken for the technical construction of a smart indicator to measure CSR. By providing a step-by-step guide for the technical construction of the indicator, we aim to show how to effectively use social media data in conjunction with (already available) traditional data to create a comprehensive indicator that accounts for various aspects of CSR (augmenting information).

The application should be regarded as an exercise. Hence, we do not provide a comprehensive examination of the CSR theoretical framework nor fully evaluate the meaning of the computed indicators from a substantive point of view. Similarly, we do not delve in discussing the selection of elementary indicators, which is ad-hoc, driven by the information available in the innovative data source, and in the evaluation of indicators' quality (see Sect. 5 for an overview of quality issues).

[9] CSR refers to the implementation of activities aiming at the improvement of firms' reputation and at positively impacting the society [42]. A related aspect, that is becoming more and more important nowadays, is the online communication of CSR activities, which can be investigated thanks to the availability of social media data. Indeed, listening to the online communication is useful to researchers and policy makers in order to monitor the behavior of the business with reference to the implementation of sustainable development and with respect to the Agenda 2030.

4.2 The application of the modular framework

For the sake of illustration, units considered are the firms included in the Dow Jones Industrial Average index, i.e., a stock market index that measures the performance of the 30 largest US listed companies as of the composition in August 2020. The data were collected as part of a previous study and refer to the year 2019 [45]. We retrieved the full list of firms, jointly with the corresponding activity sector from Bloomberg. With respect to sectors classification, Bloomberg adopts the Global Industry Classification Standard (GICS) developed by MSCI and S&P Dow Jones. The final number of firms included in the analysis is reduced to 26 as we only consider the firms for which data are available in both the traditional and digital data source.

For the traditional indicator, we consider the Environmental, Social and Corporate Governance (ESG) database provided by Refinitiv, one of the world's largest providers of financial markets data and infrastructure (commercial data). Data for listed companies refer to their sustainability performance considering various aspects, including emission reductions, social programs, and economic performance. The database collects publicly reported data, checked for quality, and provides a CSR-Strategy Score. This reflects a company's practices to integrate economic (financial), social and environmental dimensions into its day-to-day decision-making process and it ranges between 0 and 100. In the subsequent analyses, the CSR-Strategy Score corresponds to the traditional indicator (TRAD-INDEX), which is therefore already available.

As innovative data source, we consider Twitter, which is one of the main communication channels for companies [44]. Since here the INN-INDEX is based on social media data, it is renamed SM-INDEX. For the construction of the social-media based index (SM-INDEX), we follow the modular methodologies proposed in Sect. 4. This is discussed in detail layer-by-layer in the following subsections. Figure 3 provides a representation of the process described.

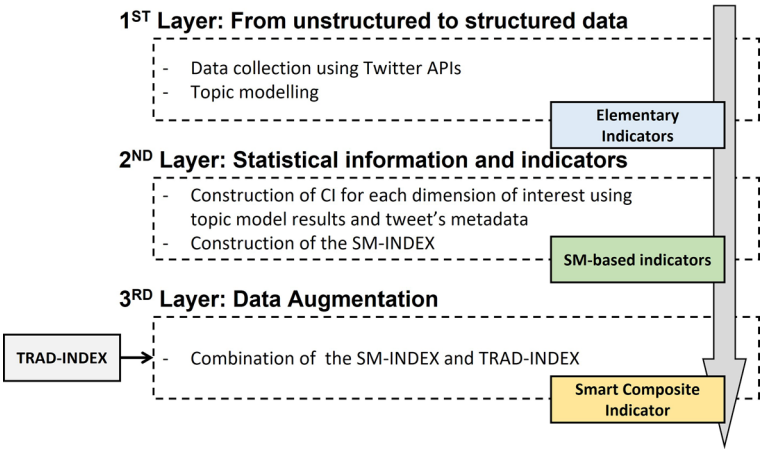


Fig. 3 Modular methodological framework applied to the specific empirical exercise

4.2.1 The first layer: elementary indicators

Following the tasks in the first layer, we identified and retrieved the data from the official Twitter accounts of the companies. Given that companies may have several Twitter accounts, we focused primarily on CSR accounts and, in case these are not available, on the news or multipurpose ones. The objective is to reduce the noise (no-CSR tweets) in the data. We use the same data retrieved by Salvatore et al. [45]. However, we restrict the analysis to the 26 firms for which information is available in both the traditional and the innovative data source. This results in the inclusion of 39 different accounts (5 news, 18 multi-purpose and 16 CSR-related) for a total of 21,919 messages posted in 2019.

The selection of the elementary indicators for each pillar is based on the theoretical framework outlined in Sect. 4.1. To the purposes of identifying elementary indicators for the content pillar, we applied a Structural Topic Model (STM) to discover the topics discussed in the tweets. Next, we grouped those detailed topics to the main CSR dimensions, namely economic, social, environmental and general (or mixed).[10] These essentially correspond to proportion of text devoted to each CSR dimension for each tweet and represent the elementary indicators with respect to the content pillar, namely social (SOCIAL), economic (ECO), environmental (ENV), and general CSR (MIX).

For the modality pillar, we consider tweets' metadata. In this respect, each tweet can contain hashtags (defining the topic of posts and allowing users to associate the tweet with all other tweets using the same identifying hashtags), mentions (engaging with other users), media (e.g., photos), and links (to external web pages). We thus define four elementary indicators for the modality pillar, corresponding to the number of hashtags (#), mentions (@), media (MEDIA), and links (LINK) contained in each tweet, respectively.

These elementary indicators represent the output of the first layer, which is the base for the construction of intermediate CIs in the second layer. Elementary indicators are measured at the tweet level and then aggregated at the firm level (the unit of our analyses).

4.2.2 The second layer: development of the social media-based indicator

The CI for the content dimension is constructed by considering the elementary indicators SOCIAL, ECO, ENV, MIX, corresponding to the proportion of text devoted to each CSR dimension for each tweet. We assume that these proportions are substitutes (compensatory aggregation) with the same importance (no weight). To obtain the CI, we take the sum of these proportions at the tweet level and then aggregate them at the firm level by taking the arithmetic mean (content indicator).

As for the modality pillar, we consider the elementary indicators the presence of hashtags, mentions, media, and links (binary variables), assuming them to be substitutes (compensatory aggregation) with the same importance (no weight). For each tweet, we sum these individual indicators, obtaining a score between 0 and 4. We then aggregate these scores at the firm level by computing the arithmetic mean (modality indicator).

Once the modality and the content indexes are constructed at the firm level, it is necessary to combine them to obtain the SM-INDEX. In this case we propose to apply the Mazziotta–Pareto index (MPI), which is partially compensatory, recognizing that the two dimensions are equally important but partially substitute to gain efficiency in CSR communication. Indeed, a deficiency in the content can be partially compensated by effective communication (and

[10] A short description of the STM and topic model output can be found in Appendixes A and B. Results can be found in greater detail in Salvatore et al. [45].

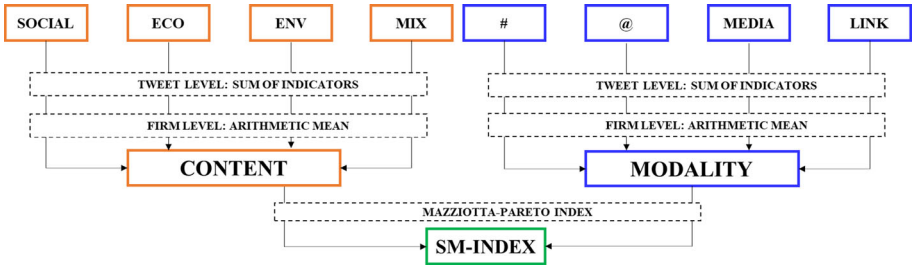


Fig. 4 Composite Indicator aggregation strategy (second layer)

Fig. 5 Composite Indicator aggregation strategy (third layer)

vice versa). The MPI is based on a non-linear function that, starting from the arithmetic mean of the normalized indicators, introduces a penalty for units with unbalanced indicators [38]. Denoting by i the firm and j the pillar (content and modality) and given the data matrix $X = \{x_{ij}\}$, to compute the MPI, we proceed first with standardization as follows

$$z_{ij} = 100 + \frac{x_{ij} - M_{xj}}{S_{xj}} \cdot 10 \quad i = 1, \ldots, 28 \quad j = 1, 2 \tag{1}$$

where M and S refer to the mean and standard deviation of the content and modality indexes. Next, given the positive polarity of the content and modality indicators, we compute the MPI as

$$MPI_i = M_{z_i} - S_{z_i} \cdot cv_{z_i} \quad i = 1, \ldots, 28 \tag{2}$$

where z refers to the standardized data as in (1) and M_{z_i}, S_{z_i}, cv_{z_i} denote the mean, standard deviation, and coefficient of variation of the normalized values for company i, respectively. Figure 4 summarizes the aggregation approach described above.

4.2.3 The third layer: development of an augmented information composite indicator

Considering the SM-INDEX, developed in the second layer, and the TRAD-INDEX, already available, we build a combined innovative smart indicator (SMART-INDEX). The TRAD-INDEX is standardized before the combination, while the SM-INDEX is not, being the aggregation output of previously standardized indicators. For the aggregation of SM-INDEX and TRAD-INDEX, we propose to apply the MPI, considering the positive polarity of the indicators (Fig. 5). Indeed, we assume that the two dimensions are partially compensatory, i.e., efficient communication might partially compensate low effective commitment and high effective commitment might compensate scarce communication.

The SMART-INDEX measures the commitment of companies towards CSR in a more comprehensive way, considering not only the effective commitment (traditional indicator) but also the effort in online CSR communication (social media indicator).

4.3 Empirical exercise: results

Figure 6 shows the values of the social media-based, traditional and combined indicators for each company. The table with detailed result is available in Appendix C. For the TRAD-INDEX the standardized values according to (1) used as input for the Mazziotta–Pareto SMART-INDEX are reported.

The TRAD-INDEX is very similar across all companies, except for Boeing, Chevron, Honeywell, and McDonalds for which it is particularly low and below 100 indicating a low level of effective commitment. A rational behind this similarity is that the index is constructed considering mainly compliance to laws and regulation with respect to CSR reporting that, nowadays, is a common practice for most companies. The SM-INDEX allows to discriminate better the communication about CSR commitment among firms.

The combination of the two indicators provides an innovative measure of CSR commitment and communication effectiveness, giving additional insights to researchers. Table 2 in Appendix C provides the ranking of firms based on the SM-INDEX, TRAD-INDEX, and the SMART-INDEX, respectively. Generally, firms that rank highly on the SM-INDEX place low on the TRAD-INDEX (and vice versa). Companies in the services sector (e.g., Technology and Health Care) have a higher position on the SM-INDEX and a lower position on the TRAD-INDEX. A possible explanation could be that firms in the services sector have a high need for communication via their websites, whereas firms in other sectors do not. This may be because other methods of communicating sustainability are possible when offering a consumer product (such as information on the package).

Due to their equal weighting, the SMART-INDEX provides a middle ground between the two. Nevertheless, researchers may decide to use a different weighting strategy according to their practical and theoretical evaluations [40].

The quality of the resulting innovative CIs (SM-INDEX and SMART-INDEX), can be difficult to asses as there is no benchmark to compare them to. Further analyses, such as uncertainty and sensitivity analyses, can help understand how methodological choices in the construction of the indices affect the results [46]. However, such approaches should be enlarged in order to take into account emerging aspects from novel data sources (such as selection of social media accounts, data pre-processing and analytical methods to transform unstructured data to structured one) and the multi-source nature of the process [47]. An overview about these issues is presented in the following section.

Fig. 6 Social media-based (SM-INDEX), traditional (TRAD-INDEX) and smart indexes

5 Quality considerations

The evaluation of the quality of new socio-economic indicators is of extreme importance to allow their use. This is upmost crucial in the case of smart indicators that use innovative data sources.

With specific reference to CIs, the concept of quality is strictly related to the robustness of the CIs with respect to the decisions made at each analytical step. Traditionally, this refers to normalization methods, weighting approaches, and the evaluation of uncertainty in the weights of sub-indicators [48–50]. The robustness is evaluated based on the ability of the indicator to generate accurate and consistent measures, as well as effectively differentiate between units in terms of their scores or rankings [49]. In this respect, the literature presents various possible procedures for evaluating robustness, mainly uncertainty analysis (UA) and sensitivity analysis (SA). UA focuses on how uncertainty in the input factors propagates through the structure of the CIs and influence its value. SA studies how much each individual source of uncertainty contributes to the output variance. For a general discussion of the procedures, please refer to Saisana et al. [46].

In addition to these traditional quality aspects, when working with unstructured data or non-traditional data sources, new quality considerations arise. For example, results may be affected by data extraction techniques (e.g. selection of social media accounts, scraped web-pages), pre-processing (e.g. data cleaning) and analytical choices (e.g. machine learning methods to extract the information).

Thus, when evaluating smart CIs, there are two key aspects to consider: the quality of the CI itself and the multi-source nature of the integration process. The quality of CIs depends on three factors: the quality of (1) the basic data, (2) the procedures to compute and (3) to disseminate the indicators [51]. According to [51] poor CIs result from inaccurate or non-credible data sources, wrong choices of individual indicators (lack of a theoretical background on the phenomena of interest), inconsistent approaches at the various construction steps (e.g. standardization, aggregation, weighting), lack of robustness analysis, poor description of the indicator construction and incorrect presentation of the results.

Furthermore, when dealing with multi-source statistics, further examinations must be conducted to ensure a comprehensive assessment. In the literature, various frameworks have been proposed for the assessment of quality in multi-source statistics [28, 47, 52–54]. It is evident that when integrating heterogeneous sources, a critical aspect is the assessment of the input and output quality throughout the integration process [52]. To this purpose, Reid et al. [54] propose a three phases approach where quality is evaluated in relation to: (1) the single data source, (2) the integrated data-set and (3) the output.

Given these premises, we propose to adopt a life-cycle perspective that considers quality evaluation across all analyses steps [55] and integrate the quality evaluation of the smart index (based on multi-source data-sets) into the general framework presented in the paper. The above-mentioned aspects (quality and multi-source nature) can be easily allocated into the three layers structure of the modular framework that we propose. The following paragraphs briefly discuss how to incorporate them in each of the three layers. In a future study, we aim to provide a more comprehensive discussion, incorporating a detailed worked example on quality evaluation.

5.1 Preliminary evaluations

Before engaging in the construction of smart composite indicators it is important to define the theoretical framework which defines the multi-dimensional phenomena under investigation. Subsequently, an evaluation of the suitable data sources becomes necessary, considering their characteristics and ability to measure specific aspects of the phenomena. For instance, researchers can take into account dimensions such as relevance, credibility, accessibility, and timeliness as quality criteria to justify the selection of these sources.

5.2 Layer 1: From unstructured to structured data

Following the setting of the paper and focusing on the innovative data source, this step involves evaluating the quality of both the input data (unstructured) and the output (structured-elementary indicators). The definition of quality depends on the specific data sources, as discussed in Sect. 2. For instance, when analyzing Twitter data, it is possible to refer to Salvatore et al. [25]. Generally, aspects related to the data retrieval strategy (e.g., search query, selection of social media accounts or web pages to scrape) as well as the completeness, timeliness, and coverage of the data source should be assessed and well-documented.

When dealing with unstructured textual data, various steps need to be taken to transform it into structured information, which serves as the elementary indicators (output of the first layer). These steps involve data pre-processing (cleaning and dimensionality reduction) and the implementation of machine learning algorithms such as sentiment analysis and topic modeling to extract the relevant information. Every decision made during the pre-processing and analysis phase might have an impact on the value of the resulting elementary indicators. Therefore, it is highly recommended to conduct a sensitivity analysis to assess the stability of the outcomes.

To summarize, in the first layer of analysis, researchers should provide quality indicators or evaluations related to the data source, data selection, data pre-processing, and analyses.

5.3 Layer 2: Statistical information and indicators

The second layer focuses on the construction of the innovative indicators. Traditional UA and SA can be applied to evaluate the robustness of the resulting indicators. However, in addition to classical aspects (standardization, aggregation, weighting, inclusion/exclusion of elementary indicators), incorporating the elements identified in the first layer is crucial (e.g. compare the results for different combination of data retrieval, cleaning, pre-processing and analytical strategies).

Consequently, as part of the quality evaluation, researchers should provide robustness analyses for both sub-indicators and intermediate indicators, considering the tasks performed in both the first and second layers.

5.4 Layer 3: Data augmentation

This step involves the calculation of the final smart composite indicator. As part of the quality evaluation, researchers should provide a comprehensive robustness analysis, considering not only the tasks performed in the third layer but also those carried out in the preceding layers.

By considering the entire process, a comprehensive assessment of the indicator's reliability and robustness can be obtained.

We leave the development of a comprehensive framework for quality evaluation to a future study.

6 Conclusions

The availability of new sources of data, such as social media, provides an excellent opportunity for augmenting business statistics and examining new aspects of phenomena of interest. As a means of augmenting the data, we propose a modular general framework organized in three layers that defines the tasks and the outputs of each block. In this study, we focus on the case of the construction of CIs based on the combination of traditional and digital textual data to derive smart augmented statistical indicators.

The second part of the paper shows, how the proposed methodology can be applied to real data. The specific empirical exercise of measuring CSR proved that traditional and social media-based indicators measure different aspects of the phenomenon, and enriched information is derived through data augmentation. The resulting smart index provides an innovative measure of CSR commitment and communication effectiveness.

This application can serve as a prototype. A similar modular approach and CI methodological framework can be applied to other contexts. As an innovative aspect, we also use Twitter metadata to enhance the information and construct the SM-INDEX. Our paper shows how can be interesting to include them in the construction of a statistical composite indicator. Metadata usage is an emerging topic and more research is required to better understand the opportunities and statistical challenges resulting from their use.

A single digital data source was considered to augment traditional data in this paper. However, the proposed framework allows the consideration of multiple data sources. For example, researchers may supplement traditional data with website information, social media posts, and newspaper articles. Further research will be conducted in this area in the future.

It is worth noticing that the proposed approach relies on the possibility of identifying the units under investigation in each data source. This can be in some way easier for business surveys and very difficult in the case units are individuals. For example, for businesses it is possible to identify their social media accounts or websites. In scenarios where identifying the individual units is not feasible, but aggregated data are available (for example by sector or other characteristics), the modular approach in layers can be adapted and implemented. This direction of research would require specific attention and could be the topic for further investigations.

A key aspect of the modular general framework in layers is its flexibility, enabling researchers to explore various methodological variations, propose enhancements to specific modules, and assess the sensitivity of the results at each stage.

The paper also outlines and discusses the statistical challenges and errors arising throughout the entire production process, from identification of the units of interest in the digital data source to data collection, pre-processing, analysis, and data augmentation. Further, it highlights the importance of evaluating the quality of the innovative indicators. In fact, in addition to traditional quality dimensions and techniques, this necessitates the identification of specific quality dimensions that are relevant to the data source and use case.

We consider the quality of CIs under a wider perspective and we discuss how the proposed modular structure, organized in layers, facilitates its evaluation by allowing for the assessment

of both input and output data/indicators at each layer. This design enables a comprehensive evaluation process throughout the various stages of the analysis. It is noted that traditional robustness analyses do not take into account the multi-source nature of the integration process, and the use of unstructured data as the basis for constructing the indicator. When assessing the quality of the output, it becomes crucial to take into account the multi-source nature of the process as new aspects related to data quality and the impact of analytical choices emerges [47, 52]. The paper provides an overview of these aspects, while an ongoing study will delve into them in more detail, presenting a quality framework for the layers approach. In contrast to existing studies which mainly focus on registers and administrative data, our approach considers innovative sources that provide unstructured data.

Data availability statement The data supporting this research consists of a dataset of tweets obtained through the Academic Twitter API and a set of indicators from the Refinitiv ESG database. Please refer to the Twitter/X APIs and Refinitiv documentation to access and retrieve the data.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article. Camilla Salvatore conducted part of her work on this paper as a Ph.D. student at the Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca.

Appendix

A The structural topic model (STM)

In order to identify the content of unstructured textual data, a common approach is to implement topic modeling (TM). It is an unsupervised learning technique which allows to study the underlying properties of a text in order to discover the topics discussed and get signals from the data. Among the different algorithms to implement TM, we select the STM which was originally designed to analyze open-ended survey questions, and which is becoming increasingly popular due to the possibility of estimating models including document-level metadata and, thus, characterizing the relationship between topics and metadata.

In the following, we briefly introduce the STM algorithm. For more details, please refer to Roberts et al. [56]. Figure 7 represents the model in plate notation. A topic is defined as a mixture over words and a document as a mixture over topics. In STM, document-metadata influences two components of the model, the topical prevalence that is defined as the proportion of the document that is associated to a topic, and the topical content that refers to the usage rate of word in a topic.

For the case study, we consider a previous work where topical prevalence covariates were included and the effect of time and sector on the discussion proportion of topics as part of a larger application-oriented study. As output, the STM model provides the per-word and per-document topic probabilities. We focus on the latter, i.e., we consider the probability of

a document to be generated from a specific topic (also referred as to the proportion of text generated from a topic) as the input to build social media-based indexes. For our analyses, we use R and, in particular, the stm package [57] to estimate the model and the quanteda package [58] to clean and prepare the data.

B Details about topic modeling results

We identified 47 topics, 36 of which related to CSR activities. Table 1 shows an example of the topics for each CSR dimension. More details are available in Salvatore et al. [45].

Table 1 Summary of topic modeling results

CSR dimension	Description of topics
Economic	CEO talks about leadership
	Economic impact of the business
	Announcement of partnerships
Social	Social impacts of innovation and digitalization
	Accessibility and inclusiveness (disability)
	Creating a better world for everyone
	Fighting discriminations
	Preserving the culture of communities
	Sustaining small businesses
	Workplace well-being
Environment	Reducing emissions and pollution
	Clean water
	Marine Conservation
Mixed-general CSR	Sponsorship of events

C Details about composite indicators

See Table 2.

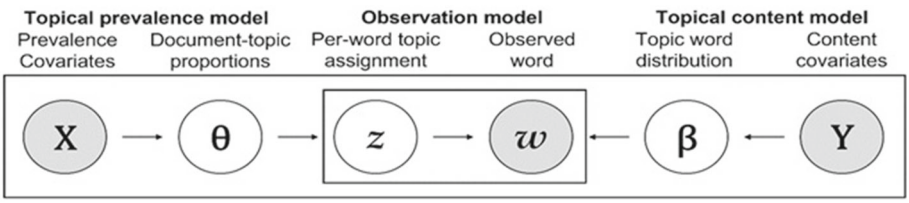


Fig. 7 Structural Topic Model. *Source*: Amended from Roberts et al. [56]

Table 2 Social media-based (SM-INDEX), traditional (TRAD-INDEX) and smart (SMART-INDEX) indexes values with ranking

Firm	SM-INDEX	TRAD-INDEX	SMART-INDEX	Rank. SM-INDEX	Rank. TRAD-INDEX	Rank. SMART-INDEX
J&J	112.11	107.52	109.72	2	2 (=)	1
Cisco	110.91	107.52	109.16	3	2 (=)	2
Amgen	103.80	107.52	105.60	5	2 (=)	3
Coca Cola	101.19	109.52	105.07	9	1 (=)	4
IBM	100.73	107.52	103.91	11	2 (=)	5
Goldman Sachs	112.20	96.96	103.47	1	6 (=)	6
Intel	101.61	104.67	103.11	8	3 (=)	7
Dow	97.40	109.62	102.79	16	1 (=)	8
Verizon	97.79	107.52	102.19	14	2 (=)	9
3M	97.18	107.52	101.82	17	2 (=)	10
Visa	101.77	101.85	101.81	7	4 (=)	11
Procter and Gamble	96.87	107.52	101.64	18	2 (=)	12
Salesforce	103.68	99.64	101.58	6	5	13
Microsoft	96.86	104.67	100.47	19	3 (=)	14
Nike	93.40	107.52	99.47	22	2 (=)	15
Travelers	96.50	101.85	99.03	20	4 (=)	16
Walmart	94.90	101.85	98.13	21	4 (=)	17
UnitedHealth	98.33	96.96	97.64	13	6 (=)	18
American express	88.53	107.52	96.88	25	2 (=)	19
Caterpillar	90.94	96.96	93.76	23	6 (=)	20
JPMorgan chase	89.89	93.86	91.79	24	7 (=)	21
McDonalds	101.02	84.51	91.29	10	8	22
Chevron	99.55	82.20	88.22	12	9	23
Honeywell	104.71	76.63	86.32	4	10 (=)	24
Home depot	79.40	93.86	85.42	26	7 (=)	25
Boeing	97.75	76.63	84.63	15	10 (=)	26

References

1. Luiten, A., Hox, J., de Leeuw, E.: Survey nonresponse trends and fieldwork effort in the 21st century: results of an international study across countries and surveys. J. Off. Stat. **36**(3), 469–487 (2020)
2. Howison, J., Wiggins, A., Crowston, K.: Validity issues in the use of social network analysis with digital trace data. J. Assoc. Inf. Syst. **12**(12), 2 (2011)
3. Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., Usher, A.: Big data in survey research: aapor task force report. Public Opin. Quart. **79**(4), 839–880 (2015)
4. Trappmann, M., Haas, G.-C., Malich, S., Keusch, F., Bähr, S., Kreuter, F., Schwarz, S.: Augmenting survey data with digital trace data: Is there a threat to panel retention? J. Surv. Stat. Methodol. **2022**, 1 (2022)
5. Stier, S., Breuer, J., Siegers, P., Thorson, K.: Integrating survey data and digital trace data: key issues in developing an emerging field. Soc. Sci. Comput. Rev. **38**(5), 503–516 (2020). https://doi.org/10.1177/0894439319843669
6. Struminskaya, B., Lugtig, P., Keusch, F., Höhne, J.K.: Augmenting surveys with data from sensors and apps: opportunities and challenges. Soc. Sci. Comput. Rev. (2020). https://doi.org/10.1177/0894439320979951
7. Ricciato, F., Wirthmann, A., Hahn, M.: Trusted smart statistics: how new data will change official statistics. Data Policy **2**, 1 (2020)
8. Ceron, A., Curini, L., Iacus, S.M.: Politics and Big Data: Nowcasting and Forecasting Elections with Social Media. Routledge, London (2016)
9. Luhmann, M.: Using big data to study subjective well-being. Curr. Opin. Behav. Sci. **18**, 28–33 (2017)
10. Iacus, S.M., Porro, G., Salini, S., Siletti, E.: An Italian composite subjective well-being index: the voice of twitter users from 2012 to 2017. Soc. Indic. Res. **2020**, 1–19 (2020)
11. Rill, S., Reinel, D., Scheidt, J., Zicari, R.V.: Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowl.-Based Syst. **69**, 24–33 (2014)
12. Daas, P.J., Puts, M.J.: Social media sentiment and consumer confidence. Tech. Rep., ECB Statistics Paper (2014)
13. Bender, S., Sakshaug, J.: Data sources for business statistics: What has changed? Surv. Stati. **2021**, 1 (2021)
14. Thomassin, M.: The migration of the Canadian census of agriculture to an integrated business program without contact with respondents. In: 5th International Workshop on Business Data Collection Methodology, Lisbon (2018)
15. Daas, P.J., van der Doef, S.: Using website texts to detect innovative companies. CBS Working Paper No.: 01-21, Tech. Rep. (2021)
16. Roelands, M., van Delden, A., Windmeijer, D.: Classifying businesses by economic activity using web-based text mining. Tech. Rep., CBS discussion paper (2018)
17. Barcaroli, G., Nurra, A., Salamone, S., Scannapieco, M., Scarnò, M., Summa, D.: Internet as data source in the ISTAT survey on ICT in enterprises. Aust. J. Stat. **44**(2), 31–43 (2015)
18. Barcaroli, G., Scannapieco, M., Summa, D.: On the use of internet as a data source for official statistics: a strategy for identifying enterprises on the web. Riv. Ital. Econ. Demogr. Stat. **70**(4), 20–41 (2016)
19. De Fausti, F., Pugliese, F., Zardetto, D.: Towards automated website classification by deep learning. Preprint arXiv:1910.09991 (2019)
20. Biemer, P.P.: Total survey error: design, implementation, and evaluation. Public Opin. Quart. **74**(5), 817–848 (2010)
21. Costanzo, L.: Use of administrative data and use of estimation methods for business statistics in Europe: an overview. In: Admin Data ESSnet Workshop "Using Admin Data-Estimation Approaches" (Vilnius (2011)
22. Bernal, I., Sejersen, T.: Big data for economic statistics. Stats Brief, Issue 28. Tech. Rep, United Nations (2021)
23. Denny, M.J., Spirling, A.: Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit. Anal. **26**(2), 168–189 (2018)
24. Symeonidis, S., Effrosynidis, D., Arampatzis, A.: A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Syst. Appl. **110**, 298–310 (2018)
25. Salvatore, C., Biffignandi, S., Bianchi, A.: Social media and twitter data quality for new social indicators. Soc. Indic. Res. **156**(2), 601–630 (2021)
26. Amaya, A., Biemer, P.P., Kinyon, D.: Total error in a big data world: adapting the TSE framework to big data. J. Surv. Stat. Methodol. **8**(1), 89–119 (2020)
27. Sen, I., Flöck, F., Weller, K., Weiß, B., Wagner, C.: A total error framework for digital traces of human behavior on online platforms. Public Opin. Quart. **85**(S1), 399–422 (2021)

28. De Waal, T., van Delden, A., Scholtus, S.: Quality measures for multisource statistics. Stat. J. IAOS **35**(2), 179–192 (2019)
29. Salvatore, C.: Inference with non-probability samples and survey data integration: a science mapping study. Metron, pp. 1–25 (2023)
30. Rao, J.: On making valid inferences by integrating data from surveys and other sources. Sankhya B **83**(1), 242–272 (2021)
31. Beaumont, J.-F., Rao, J.: Pitfalls of making inferences from non-probability samples: can data integration through probability samples provide remedies? Surv. Stat. **83**, 11–22 (2021)
32. Perez, B., Musolesi, M., Stringhini, G.: You are your metadata: identification and obfuscation of social media users using metadata information. In: Twelfth International AAAI Conference on Web and Social Media (2018)
33. Daas, P.J., Burger, J., Le, Q., ten Bosch, O., Puts, M.: Profiling of twitter users: a big data selectivity study. Tech. Rep., CBS Discussion Paper (2016)
34. Da Mota, V.T., Pickering, C.: Assessing the popularity of urban beaches using metadata from social media images as a rapid tool for coastal management. Ocean Coast. Manag. **203**, 105519 (2021)
35. Rosales Sánchez, C., Craglia, M., Bregt, A.K.: New data sources for social indicators: the case study of contacting politicians by Twitter. Int. J. Digit. Earth **10**(8), 829–845 (2017)
36. Mazziotta, M., Pareto, A.: Methods for constructing composite indices: one for all or all for one. Riv. Ital. Econ. Demogr. Stat. **67**(2), 67–80 (2013)
37. Mazziotta, M., Pareto, A.: Gli indici sintetici. Torino: G. Giappichelli Editore (2020)
38. De Muro, P., Mazziotta, M., Pareto, A.: Composite indices of development and poverty: an application to MDGS. Soc. Indic. Res. **104**(1), 1–18 (2011)
39. Booysen, F.: An overview and evaluation of composite indices of development. Soc. Indic. Res. **59**(2), 115–151 (2002)
40. Mazziotta, M., Pareto, A.: Weighting in composite indices construction: the case of the Mazziotta–Pareto index. Riv. Ital. Econ. Demogr. Stat. **2022**, 1 (2022)
41. OECD: Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing (2008)
42. Carroll, A.B., et al.: The pyramid of corporate social responsibility: toward the moral management of organizational stakeholders. Bus. Horiz. **34**(4), 39–48 (1991)
43. Dahlsrud, A.: How corporate social responsibility is defined: an analysis of 37 definitions. Corp. Soc. Responsib. Environ. Manag. **15**(1), 1–13 (2008)
44. Araujo, T., Kollat, J.: Communicating effectively about CSR on twitter: the power of engaging strategies and storytelling elements. Int. Res. **2018**, 1 (2018)
45. Salvatore, C., Biffignandi, S., Bianchi, A.: Corporate social responsibility activities through Twitter: from topic model analysis to indexes measuring communication characteristics. Soc. Indic. Res. **164**(3), 1217–1248 (2022)
46. Saisana, M., Saltelli, A., Tarantola, S.: Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. J. R. Stat. Soc. Ser. A (Stat. Soc.) **168**(2), 307–323 (2005)
47. Rocci, F., Varriale, R., Luzi, O.: Total process error: an approach for assessing and monitoring the quality of multisource processes. J. Off. Stat. **38**(2), 533–556 (2022)
48. Greco, S., Ishizaka, A., Tasiou, M., Torrisi, G.: On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc. Indic. Res. **141**, 61–94 (2019)
49. Terzi, S., Otoiu, A., Grimaccia, E., Mazziotta, M., Pareto, A.: Open Issues in Composite Indicators: A Starting Point and a Reference on Some State-of-the-Art Issues. Edizioni Roma Tre-Press, Teseo Editore (2021)
50. Freudenberg, M.: Composite indicators of country performance: a critical assessment. Tech. Rep., OECD (2003)
51. Giovannini, E.: Towards a quality framework for composite indicators. OECD (2004)
52. De Waal, T., van Delden, A., Scholtus, S.: Multi-source statistics: basic situations and methods. Int. Stat. Rev. **88**(1), 203–228 (2020)
53. Zhang, L.-C.: Topics of statistical theory for register-based statistics and data integration. Stat. Neerl. **66**(1), 41–63 (2012)
54. Reid, G., Zabala, F., Holmberg, A.: Extending TSE to administrative data: a quality framework and case studies from stats NZ. J. Off. Stat. **33**(2), 477–511 (2017)
55. Groves, R.M., Lyberg, L.: Total survey error: past, present, and future. Public Opin. Quart. **74**(5), 849–879 (2010)
56. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. **111**(515), 988–1003 (2016)

57. Roberts, M.E., Stewart, B.M., Tingley, D.: STM: an R package for structural topic models. J. Stat. Softw. **91**, 1–40 (2019)
58. Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A.: Quanteda: an R package for the quantitative analysis of textual data. J. Open Source Softw. **3**(30), 774 (2018)