The Survey Statistician, 2022, Vol. 85, 10–18.

Ask the Experts

Data Sources for Business Statistics: What has Changed?

Stefan Bender[1] and Joseph W. Sakshaug[2]

[1] Deutsche Bundesbank & University of Mannheim, stefan.bender@bundesbank.de
[2] German Institute for Employment Research & Ludwig Maximilian University of Munich, joe.sakshaug@iab.de

Abstract[1]

The production of business statistics has been experiencing a shift from a primary reliance on single-source statistics based on survey data to a greater reliance on alternative data sources and multisource statistics. Much of this shift has focused on the potential uses of unstructured data sources originating from digitalization processes. This article provides an overview of the current landscape of data sources for business statistics, highlighting some of their advantages/disadvantages, applications, and opportunities and challenges of linking them.

Keywords: administrative data, Big Data, data linkage, data collection, sample surveys

1 Introduction

Traditionally, data on businesses, establishments, and companies have mostly been used to produce single-source statistics based on surveys in which a coherent and pre-defined set of variables is observed. The advantage of this approach is that units, populations, variables, and timing can be explicitly defined by the researcher or statistician. A substantial part of the production efforts come prior to data collection where an explicit data generating process along the Total Survey Error (TSE) framework can be established (Biemer, 2010). In comparison to administrative data and unstructured business data (discussed later in this article), relatively fewer efforts come after data collection where additional activities such as data quality management and post-processing are performed. The differences in the distribution of pre- and post-data collection effort across different types of data sources (surveys, administrative/commercial data, and unstructured business data) are depicted in Figure 1.

In the last years alternative data sources, such as structured data (e.g. administrative records) and unstructured data (e.g. automated data recording) for businesses, establishments, and companies have received increasing attention and are playing a major role in the production of business

[1] Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem. Acknowledgements: The authors would like to thank Frauke Kreuter (LMU Munich) and Naoual El-Ouche, Maurice Fehr, Elena Triebskorn, Susanne Walter (all Bundesbank).

statistics. The use of these data sources is in a sense secondary because these data are typically collected for purposes other than research or producing business statistics. Because researchers and statisticians are not involved in the data generating process, the effort prior to data collection is low or – if there is some involvement – medium. Most efforts come after data collection because the data must be transformed for statistical or research purposes.

In the case of unstructured data, the distribution of effort is opposite to survey data. Because the data are "organic" or "found", relatively little effort is put into the pre-data generation process. But to transform the data from unstructured to structured or to evaluate the data quality of these unknown and (possibly changing) sources requires a lot of time and effort and specific methodologies to produce accurate estimates for the intended target population.

Bringing these alternative data sources together to supplement more traditional data sources offers new possibilities to increase the information richness of the units being observed. But bringing these different data sources together – for example, with record linkage techniques – into one harmonized data source can be a large effort, because in most cases a common identifier is missing and/or the definition of the units of workplaces, establishments, and companies differ in the data sources.

In this article, we provide an overview of the current landscape of data sources for business statistics, highlighting some of their advantages/disadvantages, applications, and opportunities and challenges of linking them.

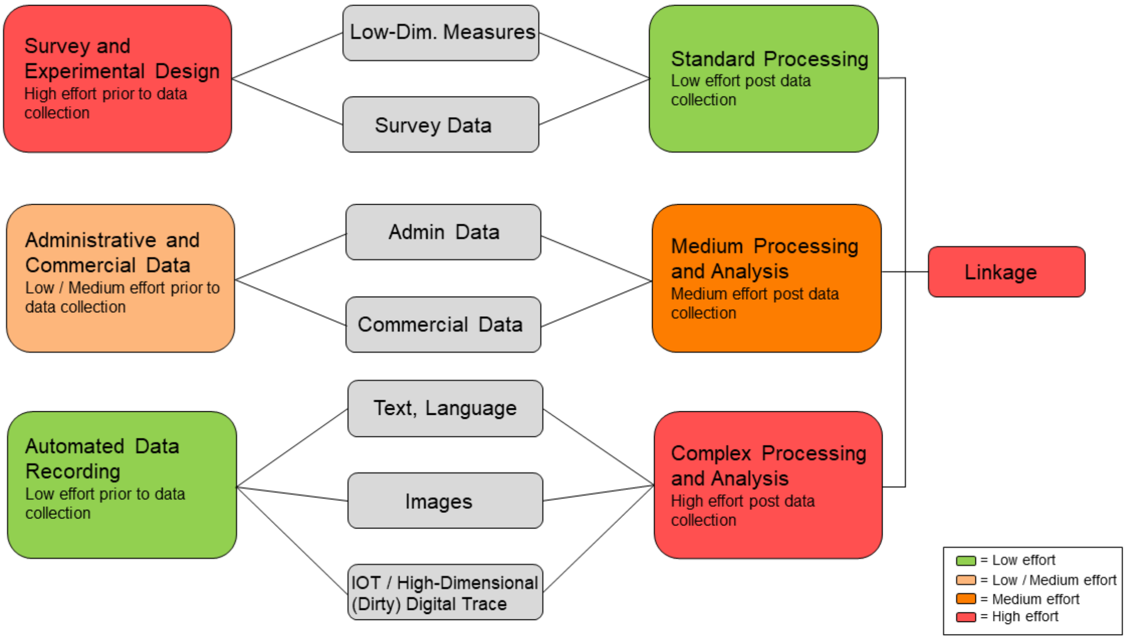

Survey and Experimental Design
High effort prior to data collection
Low-Dim. Measures
Survey Data
Standard Processing
Low effort post data collection
Administrative and Commercial Data
Low / Medium effort prior to data collection
Admin Data
Commercial Data
Medium Processing and Analysis
Medium effort post data collection
Linkage
Automated Data Recording
Low effort prior to data collection
Text, Language
Images
IOT / High-Dimensional (Dirty) Digital Trace
Complex Processing and Analysis
High effort post data collection
🟩 = Low effort
🟧 = Low / Medium effort
🟧 = Medium effort
🟥 = High effort

Figure 1. *Distribution of researcher effort for pre- and post-data collection activities for different data sources (adapted and expanded from Stahl et al., 2021, p.6).*

2 Business Survey Data

Business surveys continue to be a dominant source of structured data used to produce official statistics and evaluate and inform economic policies. These data are structured in the sense that the researcher has control over the design of the sample, questionnaire, and data collection procedures, which leads to a standard rectangular dataset of sampled units and variables available for analysis. The amount of pre-data collection effort is high, but the advantage is that every step of the data generation process is carefully planned and documented and can be deliberately altered to adapt to changing research needs. Further, possible error sources are known in advance according to the TSE framework and the survey can be explicitly designed with those error sources in mind (Biemer, 2010).

Taking a closer look at business surveys, one can immediately see the variations and possibilities that exist for collecting relevant information. In addition to standard cross-sectional and longitudinal designs, there are cross-national business surveys, such as the European Company Survey[2], that allows researchers to perform comparative analyses of business conditions and characteristics. Such data are particularly relevant at the present time as researchers are interested in the impacts of the COVID-19 pandemic on businesses and workplace practices in countries that implemented different containment measures (Jones et al., *in press*). The pandemic has also spurred interest in high-frequency measurements of businesses and how businesses adapt to international crises as they comply with changing regulations. The IAB BeCovid panel survey is one such example of a high-frequency business survey that has collected weekly data from establishments since the early stages of the pandemic (Bellmann et al., 2021).[3]

Although survey data are widely used in the production of business statistics, they are known to be susceptible to errors that can affect their accuracy. For example, nonresponse is a common issue, especially in voluntary business surveys, where response rates have declined over time particularly among larger establishments (König et al*., in press*). Methods to adjust for nonresponse bias, including the use of administrative data (discussed later) and machine learning algorithms, have been the subject of ongoing research (Küfner, Sakshaug, and Zins, 2020). Measurement error and item missing data are also potential issues that affect data quality in business surveys. Given the complex questions asked of businesses and the varying ease with which respondents can access their records or other relevant systems to answer them, there is the potential for misreporting and item nonresponse (Bavdaž, 2010). Although the TSE framework provides an impetus for designing surveys in a way that minimizes the impacts of these error sources, sometimes trade-offs between errors must be made given the survey's budget and research aims.

A key advantage of business surveys is the possibility to embed carefully designed experiments within the data collection. Collecting experimental data is more widespread in household surveys than in business surveys, but recent developments have signaled an increased interest in business survey experimentation (Langeland et al., *in press*). Some experiments are substantive in nature (e.g. vignettes) but also take the form of methodological innovations aimed at reducing survey errors or costs, such as implementing different contact protocols to improve response rates, providing enhanced instructions to complex survey questions in order to reduce item nonresponse, or introducing push-to-web strategies. Sometimes surveys experiment with complete redesigns where multiple changes to the recruitment protocol or questionnaire are implemented simultaneously and compared with the original design on various data quality indicators. However, implementing well-controlled experiments in business surveys can be challenging as production goals are usually prioritized over experimentation, which can lead to unplanned deviations in the implementation of the experiment and possible confounding effects.

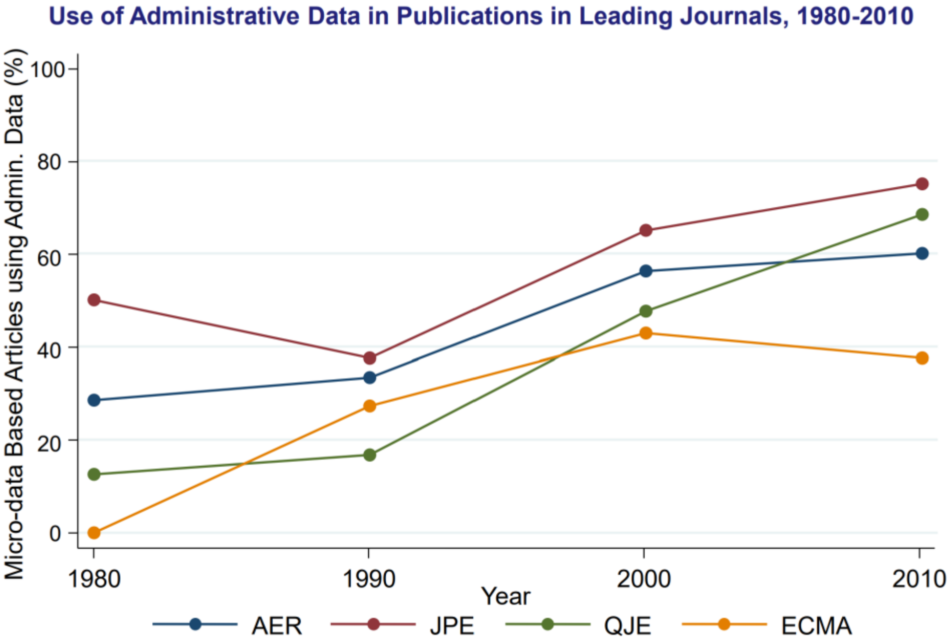
3 Administrative and Commercial Business Data

Administrative data typically refers to data generated or collected by governments or other organizations for purposes other than statistics or research. Sources of administrative business data

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

[3] For the same reason, the Bundesbank has established the Bundesbank Online Panel Firms (Deutsche Bundesbank 2021).

could be business registers or company registrations, records from tax and customs authorities, notifications of social security contributions, reports for fulfilling legal requirements, application forms for loans/credits, and information for subsidies, which were highly relevant during the COVID crisis. Additionally, there is also detailed information from the financial sector available, including investment, trade, financial and capital transactions, financial statements, or insolvency data. For many countries it is possible to bring together administrative data at the employer level with the employee level to have linked employer-employee data. In the field of labor market analysis these linked employer-employee data are one of the main data sources, because they allow researchers to analyze the joint role of worker and firm heterogeneity, both observed and unobserved.



Note: “Administrative” datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.

Figure 2. *Articles using administrative data published in leading economic journals between 1980-2010. Source: Chetty (2012)*

Chetty (2012) showed that the use of administrative data in articles published in leading economic journals has increased in recent decades (see Figure 2)[4]. Administrative data have several advantages that have contributed to their popularity in research. As economist and Nobel Laureate David Card and his co-authors (2010) remarked, “Administrative data offer much larger sample sizes and have far fewer problems with attrition, non-response, and measurement error than traditional survey data sources. Administrative data are therefore critical for cutting-edge empirical research and particularly for credible public policy evaluation.” Administrative data are often comprised of total populations, but they normally have fewer variables than surveys; though, most variables of interest (e.g. dates, services rendered, status changes) are measured very precisely. They often have a longitudinal structure with a detailed time scale, which allows researchers to follow businesses and their workers over an indefinite time span (and without panel attrition). Given their large quantity, administrative data may also provide more granular information than is possible with surveys and without increasing response burden.

Although administrative data are increasingly used, these data also have some drawbacks. Because most administrative data are not collected for statistics or research, the data generation process is outside the control of the researcher. Therefore, the definitions of some variables might not match 100 percent with the theoretical concept under study, or – in the worst case – the relevant variables

[4] Chetty (2012) also shows a concomitant decrease in the use of secondary survey data in published articles for the same time span and same journals.

are missing. Further, there may also be differences in population coverage, unit types, periodicity, and measurement accuracy compared to survey data. A growing literature investigates the quality issues associated with directly using administrative data in business statistics. An overview of these issues is provided by Van Delden and Lewis (*in press*).

Commercial data is another data source used in business statistics. There are several private companies that compile these data from various sources and offer a variety of data products comprising varying types and amounts of information on business units. Bureau van Dijk[5] and Dun & Bradstreet[6] are two of the major international commercial data providers. There is also an increasing number of local and national commercial data providers, which are closing information gaps by providing business data on sustainability, company networks, and other relevant topics. Federal statistical agencies are increasingly purchasing data from commercial providers as these products promise high data quality. However, commercial data can have similar drawbacks as administrative data in terms of definitional differences, coverage deficiencies, missing data, and measurement error. The quality of these data sources is largely understudied in the literature.

4 Unstructured Business Data

Surveys, administrative data, and – to a lesser extent – commercial data are the backbone of statistical research on businesses. These data are considered as structured data because an underlying data generation model is present, which – in the ideal case – organizes the data in a rectangular format and the relationships between the different rows and columns are known. For surveys, the elements of the data generation model map onto the TSE framework and for administrative data they map onto definitions of units and variables based on regulations or laws. Thus, the data are, in a sense, designed or pre-defined, although they may not necessarily map directly onto the definitions and variables used at a statistical agency or by a researcher.

With the rise of new data science techniques, such as natural language processing, web scraping, text mining, machine learning, advanced visualization techniques, and Artificial Intelligence, so-called unstructured data, such as sensor data, satellite images, scanner data, web sites, data communications, etc. are getting more attention by researchers and statisticians. This attention is also driven by the fact that we are surrounded by unstructured data, and some new research topics require unstructured data as one, or the only, data source.

Some researchers have pointed out that most data are unstructured (e.g. King, 2019). The difference between unstructured and structured data are that unstructured data are not based on an explicit data generation model and the data are not pre-defined. Unstructured data for businesses can be, for example, text information from annual reports, newspaper articles about the business itself, internal records, the management or the location(s) of the business, news/discussions/comments in social media (e.g. Twitter feeds or Facebook), speeches of the higher management, protocols from meetings, or financial or trade information from different sources. In addition to these more text-based sources, pictures can play a significant role, for example, photographs of the company, the company's surroundings, and satellite images. Even marketing videos or videos of CEOs' speeches can be sources for analysis. The use of sensing technology and internet data communication in some industries, including smart farming and transportation, also generates massive amounts of sensor data that can be used for analyzing businesses (Wolfert et al., 2017; Punt and Snijkers, 2019).

In most cases unstructured data must be transformed into structured forms in preparation for analysis. Because the information content is not fixed or determined a priori, different techniques are

[5] https://www.bvdinfo.com/

[6] https://www.dnb.com/

used for different purposes. For example, in text analysis, one can think of the following types of analysis: search for relevant content, clustering, classification, sentiment analysis, synonyms, named entity linkage, general extraction, visualization, summarization, and translation. To transform text into structured data, an analysis pipeline with initial processing, adding linguistic features, converting enriched text to a matrix, and the analysis plan should be established (Klochikhin and Boyd-Graber, 2020).

A hot topic application of unstructured business data is the study of climate change. Businesses play an important role in the discussion of climate change, but there is a lack of high-quality and accessible climate-related data at the business level. The lack of data poses a challenge to policymakers, researchers, statisticians, the private sector, and regulators. Although global progress on improving and making climate data available is underway, in the short- and medium-term such data have to be extracted and collected from mostly unstructured data sources (in addition to some commercial data providers). The German Bundesbank with its Sustainable Finance Data Hub[7], as one example, tries to obtain transition risks, which are typically observed at the business level from unstructured data sources. For example, information on greenhouse gas intensities is published in annual reports, dedicated sustainability reports, on company websites, or are estimated. The information can be reported in the form of tables, pictures, or text.

In addition to climate change, there are other examples of using unstructured data for studying businesses. One example is automatic validation of their economic sector. The economic sector of a business is often (self-)reported in different sources, which leads to different economic sector codes for the same establishments due to misreporting and different measurement schema. A natural question is whether unstructured data, namely, visual information about the company's facilities can be used to validate their economic sector. The Bundesbank is planning to combine information from multiple sources, including structured survey and administrative data with geoinformation, satellite images, and street views to get an indication of the necessity of checking the economic sector of a business (Walter, *in press*).

As with survey and administrative data, these new data sources also carry quality considerations. Representativeness, validity and reliability, coverage issues, and changes in frequency of delivery or data generation processes are just a few such considerations. The various ways in which unstructured data can be prepared for analysis and analyzed also presents a risk of multiple (and possibly conflicting) conclusions being drawn from the same data source. Reproducibility of data preparation and analyses is another important consideration as are data availability, access, sharing, and harmonization. Efforts to adapt the TSE framework to the "Big Data" context are currently underway (Amaya, Biemer, and Kinyon, 2020).

5 Linking Multiple Data Sources

Linking multiple business data sources increases the potential to support statistics, evidence-based policy making, and research. The need for linked business data has increased in recent years, especially for tracking multinationals (large business units) or for describing or analyzing the 2008 financial crisis or the consequences of the COVID-19 pandemic, just to name a few examples. There are a multitude of advantages for combining multiple data sources for businesses, including enhancing the richness of substantive information for a given unit, creating population frames with improved coverage, removing survey questions that are covered by alternative data sources, measurement validation and error adjustment, and improving estimation quality.

[7] Information about the Hub can be found in the presentation of Elena Triebskorn: https://www.bis.org/ifc/events/210709_prog/bundesbank.pdf

A prominent example of linking data sources in the business context is the Longitudinal Employer-Household Dynamics program at the US Census Bureau (Abowd, Haltiwanger, and Lane, 2004), a linkage of various survey and administrative datasets that allow researchers to study labor market dynamics within and across firms, the spatial distribution of employment, and various employment statistics. Linking business registers with trade statistics to compile trade flows by business characteristics is another topic of current research that informs policymakers on the role of businesses in the trade of merchandise, services, and foreign direct investment (Snyder and Jansen, 2015). Statistics Canada has been exploring the integration of administrative data and remote sensing data to supplement or replace survey data, reduce response burden, and implement small area estimation techniques to improve the quality of business statistics (Thomassin, 2018; Duval, Laroche, and Landry. *In press*).

Linking data sources is usually straightforward if there is a unique identifier for each entity in the data sources to be linked. One example of a unique identifier for businesses is the Legal Entity Identifier (LEI), which serves as an international business or entity registration number and allows for the tracing of financial transactions to specific companies or organizations. However, if no unique identifier exists, then identifying the same entities from each of the data sources is more challenging. In this case, the researcher must rely on other indicators that partially identify the entities (e.g. company name, address, economic sector, balance sheets) and link entities that have multiple fields in common. In the context of linking large structured and unstructured business data sources, some of the challenges are linking data sources with very few fields in common, data quality issues (e.g. missing data, misspellings, abbreviations), and duplicate entities. Besides possible linkage errors, other data quality issues can arise when producing multisource statistics, including representation errors and measurement errors. Van Delden et al. (*in press*) provide a framework for conceptualizing these error sources in multisource statistics.

6 Conclusions

The data landscape for business statistics has evolved significantly in recent decades from relying on traditional single-source statistics based on surveys, to greater use of alternative data sources such as administrative/commercial data and unstructured data collected from text, images, video, among others, and multisource statistics based on the combination of these data sources. Each data type is unique in its properties and the amount of pre- and post-data collection processing required to prepare the data for analysis. What remains constant throughout this shift is the importance of understanding the underlying data generation process of each data type, so that researchers are aware of the strengths and limitations of the data when generalizing and drawing conclusions from them. While survey data has established quality frameworks for understanding and quantifying the various error sources that can arise during the data generation process, such frameworks for evaluating the quality of administrative data, commercial data, and unstructured data are only recently starting to emerge. Lastly, the collection of administrative, commercial or unstructured data requires data science skills and methodologies for processing and analyzing these data as well as procedures for accessing, documenting, and archiving these data.[8]

References

Abowd, J.M., Haltiwanger, J., and Lane, J. (2004) Integrated Longitudinal Employer-Employee Data for the United States. *American Economic Review*, **94(2)**, 224–229.

Amaya, A., Biemer, P.P., and Kinyon, D. (2020) Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, **8(1)**, 89–119.

[8] For accessing these types of data see, for example, Bender et al. (in press).

Bavdaž, M. (2010) Sources of Measurement Errors in Business Surveys. *Journal of Official Statistics*, **26(1)**, 25-42.

Bellmann, L., Gleiser, P., Kagerl, C., Kleifgen, E., Koch, T., König, C., Leber, U., Pohlan, L., Roth, D., Schierholz, M., Stegmaier, J., Aminian, A. (2021) *The Impact of the Covid-19 Pandemic: Evidence from a New Establishment Survey*. IAB-Forum, 26th February 2021, https://www.iab-forum.de/en/the-impact-of-the-covid-19-pandemic-evidence-from-a-new-establishment-survey/

Bender, S., Blaschke, J., Hirsch, C. (In Press) Statistical Data Production in a Digitised Age: The Need to Establish Successful Workflows for Micro Data Access In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Biemer, P.P. (2010) Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, **74(5)**, 817-848.

Bundesbank (2021) Assessments and Expectations of Firms in the Pandemic: Findings from the Bundesbank Online Panel Firms. *Deutsche Bundesbank Monthly Report*, **April 2021**, 33-56.

Card, D.E., Chetty, R., Feldstein, M., and Saez, E. (2010) *Expanding Access to Administrative Data for Research in the United States*. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas. http://dx.doi.org/10.2139/ssrn.1888586

Chetty, R. (2012) *Time Trends in the Use of Administrative Data for Empirical Research*. http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf

Duval, M-C., Laroche, R., and Landry, S. (In Press) Integrating Alternative and Administrative Data into the Monthly Business Statistics: Some Applications from Statistics Canada. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Jones, J., Ryan, L., Lanyon, A.J., Apostolou, M., Price, T., König, C., Volkert, M., Sakshaug, J.W., Mead, D., Baird, H., Elliott, D., and McLaren, C.H. (In Press) Producing Official Statistics During the COVID-19 Pandemic. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

King, T. (2019) 80 Percent of Your Data Will Be Unstructured in Five Years. *Data Management Solutions Review*, March 28, 2019. https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

Klochikhin, E., and Boyd-Graber, J. (2020) Text Analysis. In: *Big Data and Social Science, 2nd Edition* (eds. I. Foster, R. Ghani, R. Jarmin, F. Kreuter, and J. Lane), Chapman and Hall/CRC, 193-219.

König, C., Sakshaug, J.W., Stegmaier, J., and Kohaut, S. (In Press) Trends in Establishment Survey Nonresponse Rates and Nonresponse Bias: Evidence from the 2001-2017 IAB Establishment Panel. *Journal of Official Statistics*.

Küfner, B., Sakshaug, J.W., and Zins, S. (2020) *Using Administrative Data and Machine Learning to Address Nonresponse Bias in Establishment Surveys*. Presented at the Big Data Meets Survey Science (BigSurv20) Conference, Virtual, November. https://www.bigsurv20.org/conf20/uploads/15/62/Presentation_BigSurv2020_K_fner.pdf

Langeland, J., Ridolfo, H., McCarthy, J., Ott, K., Kilburg, D., CyBulski, K., Krakowiecki, M., Vittoriano, L., Potts, M., Küfner, B., Sakshaug, J.W., and Zins, S. (In Press) Results from Selected

Experiments in Establishment Surveys. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Punt, T., and Snijkers, G. (2019) *Exploring Sensor-Generated Data in Precision Farming: Towards Official Statistics Using Business Sensor Data*. Heerlen: Statistics Netherlands.

Snyder, N., and Jansen, R. (2015) *Linking Business Registers to Trade Statistics*. Presented at the European Establishment Statistics Workshop 2015 (EESW15), Poznan, Poland, September. https://statswiki.unece.org/download/attachments/123143219/2015Slides_S6_4_Snyder%20Jansen-Linking%20Business%20Registers%20with%20Trade%20Statistics-EESW15.pdf

Stahl, F., Bischl, B., Gehrlein, Kreuter, F., and Tochtermann, K. (2021) *BERD@NFDI in a Nutshell*. https://www.berd-nfdi.de/wp-content/uploads/resources/BERD-NFDI-in-a-nutshell.pdf

Thomassin M. (2018) *The Migration of the Canadian Census of Agriculture to an Integrated Business Program Without Contact with Respondents*. Presented at the Fifth International Workshop on Business Data Collection Methodology, Lisbon, September 2018. https://www.ine.pt/scripts/bdcm/doc/ppt/D2_22_01_BDCMLisbon2018_StatisticsCanada_Mathieu%20Thomassin_PPT.pdf

Van Delden, A., and Lewis, D. (In Press) Methodology for the Use of Administrative Data in Business Statistics. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Van Delden, A., Scholtus, S., De Waal, T., and Csorba, I. (In Press) Methods for Estimating the Quality of Multisource Statistics. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Wolfert, S., Ge, L., Verbouw, C., Bogaardt, M-J. (2017) Big Data in Smart Farming – A Review. *Agricultural Systems*, **153**, 69-80.

Walter, S (In Press) *A Picture is Worth a Thousand Definitions: Validating Company Data with Satellite Images and Street View*. Technical Report, Deutsche Bundesbank, Research Data and Service Centre.