Journal of Accounting Literature 40 (2018) 102–115

Contents lists available at ScienceDirect

Journal of Accounting Literature

journal homepage: www.elsevier.com/locate/acclit

Big data techniques in auditing research and practice: Current trends and future opportunities

Adrian Gepp[a], Martina K. Linnenluecke[b], Terrence J. O'Neill[a], Tom Smith[b,*]

[a] *Bond Business School, Bond University, QLD 4229, Australia*
[b] *Faculty of Business and Economics, Macquarie University, North Ryde, NSW 2109, Australia*

ARTICLE INFO

Keywords:
Auditing
Big data
Data analytics
Statistical techniques

ABSTRACT

This paper analyses the use of big data techniques in auditing, and finds that the practice is not as widespread as it is in other related fields. We first introduce contemporary big data techniques to promote understanding of their potential application. Next, we review existing research on big data in accounting and finance. In addition to auditing, our analysis shows that existing research extends across three other genealogies: financial distress modelling, financial fraud modelling, and stock market prediction and quantitative modelling. Auditing is lagging behind the other research streams in the use of valuable big data techniques. A possible explanation is that auditors are reluctant to use techniques that are far ahead of those adopted by their clients, but we refute this argument. We call for more research and a greater alignment to practice. We also outline future opportunities for auditing in the context of real-time information and in collaborative platforms and peer-to-peer marketplaces.

1. Introduction

This paper analyses the use of big data techniques in auditing, and finds that the practice is not as widespread as it is in other related fields. We first introduce contemporary big data techniques and their origins in the multivariate statistical literature to help unfamiliar auditors understand the techniques. We then review existing research on big data in accounting and finance to ascertain the state of the field. Our analysis shows that – in addition to auditing – existing research on big data in accounting and finance extends across three other genealogies: (1) financial distress modelling, (2) financial fraud modelling, and (3) stock market prediction and quantitative modelling. Compared to the other three research streams, auditing is lagging behind in the use of valuable big data techniques. Anecdotal evidence from audit partners indicates that some leading firms have started to adopt big data techniques in practice; nevertheless, our literature review reveals a general consensus that big data is underutilized in auditing. A possible explanation for this trend is that auditors are reluctant to use techniques and technology that are far ahead of those adopted by their client firms (Alles, 2015). Nonetheless, the lack of progress in implementing big data techniques into auditing practice remains surprising, given that early use of random sampling auditing techniques put auditors well ahead of the practices of their client firms.

This paper contributes to bridging the gap between audit research and practice in the area of big data. We make the important point that big data techniques can be a valuable addition to the audit profession, in particular when rigorous analytical procedures are combined with audit techniques and expert judgement. Other papers have looked at the implications of clients' growing use of big data (Appelbaum, Kogan, & Vasarhelyi, 2017) and the sources of useful big data for auditing (e.g., Vasarhelyi, Kogan, & Tuttle

[*] Corresponding author.
E-mail addresses: adgepp@bond.edu.au (A. Gepp), martina.linnenluecke@mq.edu.au (M.K. Linnenluecke), toneill@bond.edu.au (T.J. O'Neill), tom.smith@mq.edu.au (T. Smith).

https://doi.org/10.1016/j.acclit.2017.05.003
Received 5 May 2017; Received in revised form 5 May 2017; Accepted 25 May 2017
Available online 01 February 2018

(2015); Zhang, Hu, et al., 2015); our work focuses more on valuable opportunities to use contemporary big data techniques in auditing. We contribute to three research questions regarding the use of big data in auditing, raised by Appelbaum et al. (2017) and Vasarhelyi et al. (2015): "What models can be used?", "Which of these methods are the most promising?" and "What will be the algorithms of prioritization?" We provide key information about the main big data techniques to assist researchers and practitioners understand when to apply them. We also call for more research to further align theory and practice in this area; for instance, to better understand the application of big data techniques in auditing and to investigate the actual usage of big data techniques across the auditing profession as a whole.

This paper also integrates research in big data across the fields of accounting and finance. We reveal future opportunities to use big data in auditing by analyzing research conducted in related fields that have been more willing to embrace big data techniques. We offer general suggestions about combining multiple big data models with expert judgement, and we specifically recommend that the audit profession make greater use of contemporary big data models to predict financial distress and detect financial fraud.

The paper proceeds as follows. Section 2 introduces big data techniques, including their origin in the multivariate statistical literature and relates it to the modern mathematical statistics literature. Section 3 offers a systematic literature review of existing research on big data in accounting and finance. This section highlights how auditing substantially differs from the other major research streams. Section 4 identifies novel future research directions for using big data in auditing. Finally, Section 5 concludes the paper with important recommendations for the use of big data in auditing in the 21st century and a call for further research.

2. An introduction to big data techniques

This section presents an overview of big data and big data techniques to promote a greater understanding of their potential application. Auditors that use more advanced techniques need to understand them (Appelbaum et al., 2017). An introduction to big data provides the necessary background to present the main big data techniques available and the key information needed to determine which are appropriate in a given circumstance. Appendix A describes the main big data techniques, summarizes their key features and provides suggested references for readers who want more information.

Big data refers to structured or unstructured data sets that are commonly described according to the four Vs: Volume, Variety, Velocity, and Veracity. Volume refers to data sets that are so large that traditional tools are inadequate. Variety reflects different data formats, such as quantitative, text-based, and mixed forms, as well as images, video, and other formats. Velocity measures the frequency at which new data becomes available, which is increasingly often at a very rapid rate. Finally, the quality and relevance of the data can change dramatically over time, which is described as its veracity. The auditing profession has a large and growing volume of data available to it, of increasing variety and veracity. Textual information obtained online is one new type of data, and we discuss this phenomenon later in the paper. Auditors also face an increasing velocity of data, particularly in the context of real-time information, and this is described in Section 4.

Big data comes in a variety of flavors – "small p, large n", "large p, small n", and "large p, large n", where n refers to the number of responses and p the number of variables measured at each response. These categorizations are important because they can influence which technique is the most suitable. The big data techniques described in Appendix A are suited to different categorizations; for instance, Random Forests[1] is particularly useful for "large p, small n" problems. High-frequency trading generates massive data sets of both high volume and high velocity, creating major challenges for data analysis. Nevertheless, such "small p, large n" problems are perhaps the easiest of the three scenarios and the analytic tools used are, in the main, adaptations of existing statistical techniques. The "large p, small n" scenario is best exemplified by genomics. A single human genome contains about 100 gigabytes of data. Essentially the data is a very long narrow matrix with each column corresponding to an individual and each row corresponding to a gene. The cost of sequencing a genome has now fallen to a point where it is possible for individuals to purchase their own genome. As a consequence, genomics is rapidly transitioning to the "large p, large n" scenario. Climate change research is another example of science at the forefront of the big data "large p, large n" scenario, with multivariate time-series collected from a world-wide grid of sites over very long time frames.

Big data also refers to the techniques and technology used to draw inferences from the variety of flavors of data. These techniques often seek to infer non-linear relationships and causal effects from data which is often very sparse in information. Given the nature of the data, these techniques often have no or very limited distributional assumptions. Computer scientists approach big data from the point of view of uncovering patterns in the complete record – this is often called the algorithmic approach. The patterns are regarded as approximations of the complexity of the data set. By comparison, statisticians are more inclined to treat the data as observations of an underlying process and to extract information and make inferences about the underlying process.

The statistical techniques used in big data necessitate more flexible models, since highly structured traditional regression models are very unlikely to fit big data well. Furthermore, the volume (as well as variety and velocity) of big data is such that it is not feasible to uncover the appropriate structure for models in many cases. The popularity of more flexible approaches dates back to Efron's (1979) introduction of the bootstrap at a time when increasing computer power made such new techniques feasible. The bootstrap is a widely applicable statistical tool that is often used to provide accuracy estimates, such as standard errors that can be used to produce confidence intervals. Regularization is another widely used technique which imposes a complexity penalty that shrinks estimated parameters towards zero to prevent over-fitting or to solve ill-posed problems. Ridge regression, which uses a L2 penalty,[2]

[1] Random Forests for regression-type problems uses bootstrap samples to develop multiple decision trees (usually thousands) and then aggregates them together by averaging. See Appendix A for more information.