(Data) science in financial audits

A data-driven audit where advanced modeling will be the core of the audit: we briefly discuss how cheaper computing power, academic research and open-source software development are important in realizing this future and what we should consider when we build models for the purpose of a financial audit.

Marcel Boersma MSc
is a manager at KPMG and a PhD candidate at the University of Amsterdam at the Computational Science Lab.

boersma.marcel@kpmg.nl

Most technical developments in audit research coincide with research developments in other fields

(DATA) SCIENCE IN FINANCIAL AUDITS

Auditor: "Hi Siri, I am looking at the financial report of Shell. What do you think?" – Siri: "I've used 60.000 public data sources and run 1.5 million simulations. The financial numbers of Shell are very likely to be a fair representation given the current economic circumstances. However, it seems that their margins have increased which cannot be explained by any other available data. I would recommend asking how they've achieved that."

A financial audit invoked through Siri, will this be a crazy statement or the audit of the future? You might be wondering how we can build this future? In this article we will discuss three aspects that could help us realize this future:

1. Economics of computer models: why would modeling make sense from an economic perspective?
2. Academia and open-source development: how can audit innovation leverage advances made in other research areas? More specifically, how could we utilize the faster moving academic and open-source community better?
3. Data and modeling: how can we start building models that can be used in audits?

ECONOMICS

Siri used 60.000 data sources and performed 1.5 million simulations to assess if the financial numbers made sense. Storing that amount of data, designing and running those simulations is costly. Therefore, using simple economics, we must assess whether it is profitable to use computer simulations to assess the financial numbers.

In 1966-67, Benston [Bens66] and Jensen [Jens67], discussed the application of a regression model that could estimate financial positions. The cost of using this model is around 30 USD and it was considered to be an efficient tool ([Bens66]). If we compare these costs with computer models we use nowadays, we see an astonishing drop in costs. Take for example a Tesla Autopilot. To drive your car, it continuously collects sensor data and performs all kinds of computations to keep the Tesla on the road. The costs of the Tesla Autopilot would have been astronomical if the prices would have been at the same level of 1967. An economically viable Tesla Autopilot at that time would have less sensors and less computational power, making it challenging to keep your car on the road. We require a certain minimum amount of data and computational power for the algorithm to be useful.

Similar for the audit models, we need to know if it is possible to reach a level of accuracy that is satisfactory and even then, we need to consider the cost-benefit perspective. Given the recent pricing of computer power and storage, it is likely that running data hungry models is viable from an economic perspective. We could enhance the auditor by enriching the audit workflow with machine learning models where a simple question must be answered: "Is the machine prediction cheaper than human prediction?". In scenarios where this answer is affirmative, we will start building models, often replacing mind numbing tasks for the auditor, so that the auditor can focus on the more challenging tasks at hand (e.g. account valuation). Even when the answer is affirmative, we need to consider the costs of developing such a solution. In the next section we will explain how audit innovation could benefit from academic research and open-source development to significantly speed up the development timeline.

AUDITS AND TECHNOLOGY

Most technical developments in audit research coincide with research developments in other fields. A good example is the application of time series models in audits. Researchers started to develop more advanced econometric analyses like time-series analysis ([Box70]). This research was applied to the audit by Kinney in 1978. And this is not unique, similarly, in the field of computer science and database technologies, we see applied research in audits ([McCa82], [Groo89]) which now can be related to the field of IT Auditing. Researchers in the audit field cleverly try to incorporate technological advances from other fields in audits. This raises the question of which research wave we can catch in order to take audits to the next level.

As a rule of thumb: the more accurate the models are, the costlier they become

In the statement made by Siri, we see that numerous public datasets are cross-checked with advanced simulations in order to determine if the numbers make sense or not. Data and modeling are the key ingredients that enabled Siri to do this analysis. Before we discuss the models, we first discuss how scientific advances can be used in practice. Therefore, we must understand how an idea transfers from a paper to the open-source world. Furthermore, an interesting aspect is that these open-source communities are moving faster and are now more impactful than ever before. To show the connection between the academic world and the open-source world, we discuss a couple of open-source projects below.

In 2003, Ghemawat et al. [Ghem03] published the Google File System paper in the Association for Computing Machinery (ACM), describing a way to store big data on commodity hardware. Doug Cutting implemented this in 2006 ([JIRA06]) as an open-source project which is now known as Hadoop. Hadoop is currently widely used to store big data by many large organizations. Storing a lot of data naturally created the urgency to start performing calculations on this large data set. UC Berkeley's AMPLab created Spark in 2009, which is a distributed processing framework. Spark has been open-sourced in 2010 and had nearly a thousand collaborators in 2015 ([Zaha10], [APAC14]). Another open-source project initiated by Google is Kubernetes, as a way of automated deployment and scaling of applications. Kubernetes, as an open-source project, started in 2014 and is based on fifteen years of experience of Google in scaling applications. In 2018, Kubernetes won the 'Most Impact' Award with nearly 20.000 contributors, almost a million comments in the code management platform ([KUBE18]).

From the above-mentioned examples, it becomes clear that research finds its way into open-source implementations and there is a rapidly growing community supporting and improving these products. By using these open-source products, we can leverage decades of research progress and implementations provided by some of the best engineers the world has to offer. And this is nearly impossible to reproduce by any one company in the world. This is the big wave we can ride to take audits to the next level and benefit from the collaboration of thousands of people around the globe. Although, this mainly tells us that we should use open-source software to speed up the development, it does not tell us how we should develop the models. In the next section we explain why modeling can be used to audit an organization, but more importantly how the designers of these models should therefore be aware of potential pitfalls.

WHY ARE MODELS USEFUL?

A model, which can be anything from a machine-learning model, econometrics model up to a physics model, is a simplified description of reality in mathematical terms. Modeling and simulation is an extremely powerful tool to understand a system and answer questions about that system ([STAN18]). Recent research by MIT [MIT14] provides an excellent illustration of how powerful these concepts are in doing something that seems to be impossible.

Imagine a situation with two rooms separated by soundproof glass. In one room there is a radio and a plant, in the other room, a person. The person wants to listen to the radio through the soundproof glass. Sounds, pun intended, impossible right? With some simple tools and some mathematics, however, we can solve this problem. Recent research [MIT14] shows that with a camera and an object in the other room, e.g. a glass, a bag of potato chips or a plant, is enough to reconstruct the sound in the other room. They used the knowledge that sound travels through air and impacts the objects in the room. They measured this vibrational impact on the plant with a high-speed camera and used this video to reconstruct the sound signal. Various mathematical concepts are used in this scenario to approximate the information of interest. This means that the music sounds similar, but is not exactly the same. In some models we can increase the approximation quality by doing more computations and this nicely connects to the economics of modeling. As a general rule of thumb the more accurate the models are the more costly they become.

Simulation and modeling are used in many areas, ranging from physics, to biology and chemistry, but also in the banking industry to calculate the risk exposure of the investment portfolio. Modeling techniques from agent-based models up to deep neural networks are developed for this purpose. These models are used to make accurate predictions about the system or to simulate the system under various conditions and study the output. In a similar fashion we can consider a company as a system which we want to understand better, and having a model would enable us to ask questions so that we achieve this result. In an audit, we can use cleverly constructed models to recover information of interest in a creative way. So, let's be creative, very creative! In this way we can think of truly new ways of obtaining the information of interest for audits. Before we continue with creative implementations of models in audits, we briefly discuss how these models relate to the current way an auditor operates. This is helpful in assessing which models are helpful and which are not.

The models can be used to retrodict: infer about past events using present information

Auditors study information obtained from the client, which helps them create an understanding of how the organization operates. This understanding is then used to audit the client and assess if their financial statements are a fair representation. Next, we have a look at the objective of models. One of the purposes of computer simulations of a model is to gain a better understanding of the data (information from the client) we already have. The models can be used to retrodict, that is, infer about past events using present information. This creates a deeper understanding about how these events occurred ([STAN18]). The application of models to retrodict is not very different from the objective of the auditor, only the means by which we explain the past year events are completely different. The understanding of the auditor is closely related to the mathematical model, both the understanding of the auditor and the model are used to create a deeper understanding of the subject, namely the financial statements.

However, there is one important element we need to keep in mind, that the objective is to understand. Most companies are sitting on a big pile of data and it is tempting to mine this data to uncover patterns and use these patterns to make predictions. Nevertheless, the prediction itself is not the objective, but understanding is. Professor Peter Sloot refers to this as 'Big Nonsense', and illustrates this with an excellent example ([Sloo16]):

"Astronomers of the Maya civilization and astronomers of the Babylonian civilization were brilliant in predicting astronomical events. For instance, from meticulous observations of the Sun, Moon, Venus and Jupiter they were able to predict the 584-day cycle of Venus or the details of the celestial track of Jupiter with astonishing accuracy. Yet they had no clue about our heliocentric solar system, they believed that the earth was flat, and they were completely ignorant of the real movement of stars and planets while being convinced that the sky was supported by four jaguars, each holding up a corner of the sky."

This example illustrates one of the potential dangers of uncovering patterns from historical data which we now know as Data Science or Big Data ([Sloo16]). The patterns uncovered might trick you into believing something that is not true. Therefore, the models need to be enriched with computational predictive models such that we can falsify or confirm our interpretations ([Sloo16]).

To build a model for the purpose of an audit will be challenging. Nevertheless, if the model is constructed properly then the model can be used to understand the past year activity of an organization. Parts of the audit could be replaced by such models that collect audit evidence. Maybe it is possible to reuse these models between audits in the same industry and calibrate the model to the specific needs of that organization. In any case, these models can analyze countless data sources and perform endless simulations tirelessly which cannot be achieved with manual labor. A hybrid version of the audit workflow where the auditor and the model complement each other could greatly enrich the experience of the auditor. But in order to develop the models, we need to perform research that can falsify or confirm the models and an essential ingredient for this research is the availability of data. Therefore, the clients play a role in the development as well, sharing their data and providing feedback.

CONCLUSION

We started this article with an imaginary situation where we asked Siri to audit the financial statements. Although this scenario is not yet a reality, we do see that technological developments are moving faster than ever before. Leveraging the advances made by the academic and the open-source community can shorten the development timeline dramatically. Furthermore, modeling is in many areas an economically viable business model. Nevertheless, due to the nature of the audit, we must take precaution in what kind of models we want to use and how we want to use them. We do not want to fool ourselves into thinking something that is not true. However, we must realize that an important different step must be taken. Developing this future takes effort from academia, engineers, auditors and, finally, audit clients collaborating intensively. We need the data and feedback of the clients to build these models and falsify or confirm them. The auditee plays a crucial role in developing this future, so we are counting on them as well!

References

[APAC14] APACHE, *The Apache Software Foundation Announces Apache™ Spark™ as a Top-Level Project*, Apache.org, https://blogs.apache.org/foundation/entry/the_apache_software_foundation_announces50, 2014.

[Bens66] G.J. Benston, *Multiple regression analysis of cost behavior*, The Accounting Review, Vol. 41(4), pp. 657-672, 1966.

[Box70] G.E. Box and G.M. Jenkins, *Time Series Analysis Forecasting and Control*, Wisconsin University, Madison Department of Statistics, 1970.

[Ghem03] S. Ghemawat, H. Gobioff and S.T. Leung, *The Google file system*, ACM, Vol. 37, No. 5, pp. 29-43, 2003.

[Groo89] S.M. Groomer and U.S. Murthy, *Continuous auditing of database applications: an embedded audit module approach*, Journal of Information Systems, Vol. 3(2) pp. 53, 1989.

[Jens67] R.E. Jensen, *A Multiple Regression Model for Cost Control - Assumptions and Limitations*, The Accounting Review, Vol. 42(2), pp. 265-273, 1967.

[JIRA06] JIRA, *Initial import of code from Nutch*, Apache.org, https://issues.apache.org/jira/browse/HADOOP-1, 2006. Accessed on: 29-10-2018.

[Kinn78] W.R. Kinney Jr, *ARIMA and regression in analytical review: an empirical test*, Accounting Review, pp. 48-60, 1978.

[KUBE18] Kubernetes, *Kubernetes Wins the 2018 OSCON Most Impact Award*, Kubernetes.io, https://kubernetes.io/blog/2018/07/19/kubernetes-wins-2018-oscon-most-impact-award/, 2018.

[McCa82] W.E. McCarthy, *The REA accounting model: a generalized framework for accounting systems in a shared data environment*, Accounting Review, pp. 554-578, 1982.

[MIT14] Larry Hardesty, *Extracting audio from visual information*, MIT News, http://news.mit.edu/2014/algorithm-recovers-speech-from-vibrations-0804, 2014.

[Sloo16] Peter M.A. Sloot, *Big Nonsense: the end of scientific thinking is near...*, Peter-Sloot.com, http://www.peter-sloot.com/blog/big-nonsense-the-end-of-scientific-thinking-is-near, 2016.

[STAN13] Stanford Simulation, *Computer Simulations in Science*, Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/simulations-science/, 2013. Accessed on: 29-10-2018.

[Zaha10] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker and I. Stoica, *Spark: cluster computing with working sets*, HotCloud, Vol. 10 (10-10), pp. 95, 2010.

About the author

M. Boersma MSc is a manager at KPMG and a PhD candidate at the University of Amsterdam at the Computational Science Lab. In his current role as a manager within the KPMG DANí department and KPMG Global Service Center he focuses on both the research aspect and the delivery of audit innovations.