A Data Analysis Based Framework to Detect Anomalies in Large Data Sets Using Benford's Law

Mustafa Canbolat and D. Donald Kent, Jr.

ABSTRACT

In this paper we propose a data analysis based audit framework where we identify and eliminate clusters of data points that do not have the characteristics of a Benford conforming data set. By looking at the attributes of these data sets, we identify potential audit candidates iteratively with the objective of utilizing the auditing budget in a more efficient way. We analyze a publicly available real data set that contains a list of contracts belong to a public health care organization using the proposed framework. We believe this systemic approach is better than a random selection process to better utilize audit resources.

Mustafa S. Canbolat, PhD, Associate Professor of Management, The College at Brockport, The School of Business Administration and Economics, State University of New York, mcanbola@brockport.edu.

D. Donald Kent, Jr. PhD, CPA, CFE , Associate Professor of Accounting, The College at Brockport, The School of Business

Administration and Economics, State University of New York, dkent@brockport.edu.

Keywords: Benford's Law, audit, auditing

DOI: http://dx.doi.org/10.15239/j.brcacadjb.2017.07.01.ja05

Introduction

According to a Department of Justice's recent media brief (d.o. Justice, December 2013) procurement fraud matters had a peak in fiscal year 2013. The department recovered $3.8 billion from false claims where more than $887 million of this amount was in settlements and judgments based on allegations of false claims and corruption involving government contracts.

It is certain that these kind of fraudulent activities or anomalies are not common, but when they happen it is not easy to detect them and even if they are eventually detected, their negative effect on the perception of these entities in public will be huge.

Health care sector is not immune to this kind of behaviors, either. The National Health Care Anti-Fraud Association (NHCAA) reports that the financial losses due to health care fraud are in the tens of billions of dollars each year.

Therefore it is important to have a data analysis based auditing mechanism in place that detects anomalies and fraud in the past data so that it will create a physiological barrier to the people who have intent to transgress. The difficulty comes with such a system is to determine how to classify the data set in a way that minimizes the cost to identify the maximum number of anomalies.

Data analysis is commonly used in accounting and auditing to detect errors and fraud and to assess performance. One of the most common data analysis methods is financial statement analysis. This method includes, but is not limited to, vertical analysis (common-size financial statements), horizontal analysis (percentage changes) and ratios. Ratio analysis includes performance measures for liquidity, activity, profitability, and coverage (Kieso, Weygant, and Warfield, 2016). The authors note that, while ease of computation is an advantage, limitations of ratios include the use of historical cost, estimated amounts such as depreciation and bad debt expense, and comparability with the same or other entities. Another method of data analysis used by fraud examiners is the Net Worth Method which is part of the investigative methods for conversion of embezzled funds (S. W. Albrecht, Albrecht, Albrecht, and Zimbelman, 2016). The authors show, through a series of computations, that the Net Worth Method computes an amount representing unknown sources of income. Successful use of this method requires skillfulness in obtaining known information from public and private sources, including online sources. Lastly, with the increased power of computers and increased amounts of data, Big Data Analysis is gaining importance. It is changing the way we analyze data by the ability to explore large data sets, assess data lacking neatness and exactness, and focus on correlation over causality. (Mayor-Schonberger and Cukier, 2013).

In this paper, another data analytical method that is widely used in fraud detection and known as Benford's Law is discussed. Benford's Law is a data analysis methodology supported by statistical rigor and can be applied to large data sets using data other can accounting numbers. We propose a data analysis focused audit system where we identify and eliminate clusters of data points that do not have the characteristics of a Benford conforming data set. By looking at the attributes of these data sets, we identify potential audit candidates. We believe this systemic approach is better than random selection process.

Bierstaker, Brody, and Pacini (2006) investigated the accountants' perceptions regarding fraud detection and prevention methods through a survey research and found out that despite the advancements in information technology part of fraud detection such as firewalls, virus and password protection to identify and prevent fraud, most companies lack the analytics part of fraud detection that we discuss in this paper. The accountants surveyed believe that the use of analytics tools would be highly effective but yet admit that they do not have firm resources including software and man power dedicated to fraud analytics.

Background and Literature Review

The phenomenon which is the basis of this paper suggests that the leading significant digits that happen naturally in real life do not appear with the same frequencies but rather they follow a logarithmic pattern where the frequency of occurrence decrease with the increase in the digits. This characteristics of naturally occurring numbers first discovered by mathematician Simon Newcomb (Hill, 1995) in 1881 who realized that the logarithmic tables that scientists use to simplify difficult calculations wear out faster in the first pages than the last pages where the digits appear in an ascending order indicating a higher frequency in the lower digits than the higher digits. The idea was rediscovered by Frank Bedford in 1938. Benford tested the existence of this law by collecting many diverse data sets including numbers on the first pages of newspapers, all the numbers in Reader's Digest, mathematical tables, drainage areas of rivers, population numbers and American League statistics. The collection which contained about 20,000 numbers empirically proved the existence of the law and the idea became the Benford's Law. Newcomb and later Benford independently developed a set of formulas for the probability of digit frequencies as given below.

P(First Digit = d1)= log(1+1/d1), d1{1,2,…,9}

P(Second Digit = d2)= (d1= 1 to 9) (1+1/(d1d2)), d2{0,1,…,9}#

P(First and Second Digits = d1d2)=log(1+1/(d1d2)), d1d2{10,11,…,99}

It is interesting to emphasize that if there is a law of digits, it has to be scale invariant. For example the sizes of lakes can be given in square mile or in square km. Or the invoice amounts can be given in different currencies. So that means if you convert the numbers in a data set that follows the Benford's Law to another unit it should still follow the Benford's Law, and it does. The reason for that is the Benford's Law is scale-invariant (Hill, 1995). Further mathematical details of the theory of Benford's Law can be found in Berger and Hill (2011).

Benford's Law has also been used to discover fraud by uncovering abnormalities in data sets not conforming to this pattern. Nigrini (2012) notes that certain the following data sets do not follow Benford's Law: data sets containing minimum/maximum limits, data used as identification numbers or labels, and data containing more small values than large ones.

Research in accounting and auditing has focused on both applications and problems in using Benford's Law. Applications include use as an "aid in analytical procedures" (Nigrini and Mittermaier, 1997), selection of "more promising" audit samples (da Silva and Carreira, 2013), and fraud detection (Nigrini, 1999; Johnson, 2009 ; Jordon and Clark, 2011; Yang and Wei, 2010). In analyzing state government financial statements, Johnson and Weggenmann (2013) demonstrated that Benford's Law can be used effectively for smaller data sets but noted that smaller data sets may also increase the occurrence of false positives. Grabinski and Paszek (2013) showed that Benford's Law was reliable for analyzing large financial data sets using European publicly listed companies. They noted however a "lesser extent" of reliability to those "representing financial ratios."

Research on Benford's Law has also produced some criticisms. Diekmann and Jann argue that Benford's Law is not a useful tool when discriminating between manipulated and non- manipulated estimates. They also question validity based on the high occurrence of false positives. Research has also noted the potential high results of Type 1 errors (Cleary and Thibodeau, 2005; Rodriguez, 2004); using a Bayesian approach (Geyer and Williamson, 2004). There are also other streams of research aiming for implementing better methodologies for detecting fraud. A basic non-exhaustive classification of these methods are provided in Figure 1.

Kirkos, Spathis, and Manolopoulos (2007) looked at the effectiveness of three different data mining techniques to detect fraud in financial statements. They considered a Bayesian belief network model, neural network model, and decision tree model and found the Bayesian belief network superior to the other two.

Cecchini et al. (2010) collected a large empirical data set containing fraudulent and nonfradulent companies along with their quantitative financial attributes. They implemented a fraud detection model using support vector machines and validated the use of model on a new set of data. The model they used correctly labeled 80% of the fraudulent cases and 90.6% of the non-fraudulent cases indicating that the model has lower Type 1 error rate than Type 2 error rate.

Game theory is also present in fraud detection literature. Cook et al. (1997) is the first paper that considered a game model of auditing using both cooperative and non-cooperative game analysis. Later, Coates et al. (2002) used a modified version of the chicken game (Szilagyi, 2007) to more formally model client-auditor strategies using players as ethical and unethical clients on one side and an auditor on the other with an aim to provide additional insight into ethical and audit effort issues. Recently Anastasopoulos and Anastasopoulos (2012) employed the evolutionary game theory to model the fraud detection problem in auditing. One of the important findings of the model indicates that if the auditor is partially informed about the auditee firm, a more comprehensive audit is necessary to guarantee quality of audit. This means that the additional knowledge such as data collection and analysis is important.

Lately, there is a successful attempt to use the social network theory in fraud detection. Baesen et al. (2015) and Van Vlasselaer et al. (2016) study the knowledge and impact of network information for social security fraud detection. The objective of the paper is to detect a set of companies “that intentionally go bankrupt to avoid contributing their taxes”. The novel approach that the authors identified the shared resources between the companies and used this as a linkage between them.

Modeling approach

As every audit comes with a cost tag attached, it is not always possible to take a look at every item in a large data set. Businesses often conduct audits using a random selection process without conducting a thorough analysis on the data set that is subject to the audit. As the traditional data analysis tools fail to catch anomalies that are related to digit frequencies, some unusual data values stay unrecognized by the system being used.

We propose an elimination and selection based auditing system in place of a random selection process where we look at a data set by calculating statistical error measures for the expected and observed digit values using the first order tests suggested by (Nigrini, 2011, 2012). We mainly consider the significance of the Z-statistic. Figure 2 shows the overall framework of the system proposed.

In order to understand the applicability of the proposed system we analyzed a publicly available real data set of a government health care organization. In this data set we identified the contractor name as the attribute of interest.

Figure 3 is the first two digit distribution for the data set. The figure identifies large spikes at 50 and 99 first two digit combinations. These large spikes suggest that the data may not conform to Benford's law as the deviations are large. We can also see that proportions for some other digit combinations are below and above their respective expected proportions. These additional digit values also contribute to nonconformity of the data to Benford. However we only look at the largest spikes as long as the errors from the other ones stay within the significance level. Table 1 below shows the magnitude of these largest errors.

The results suggest that we first need to take a closer look at the records with the first two digit values of 99. The increase in the frequency of these data values may indicate that the company officials have a tendency to issue an invoice that has a slightly lower value than a value that requires additional authorization which is a common practice.

A closer look at the data set revealed that 32 out of 81 invoices that have the first two digits 99 were issued to a specific contractor. Therefore we take this contractor in our audit pool and remove the records related to this contractor. We call this contractor Company A to conceal its name as we only intent to show how our methodology works.

An interesting fact about Company A is that 20 out of these 32 transactions happened on the same day and they are in the amount of $9,999. A closer look at this company also revealed that there are some other transactions for this company issued on the same day in the amounts of $9,830 and $9,829.

We then remove Company A from the data set and iterate the analysis again this time finding the most occurring FT digits as 50. Notice in Table 2 that some of these 50s from the initial data set are eliminated by the removal of Company A and the largest z value is now dropped to 7.71.

We found that 17 out of 94 of these 94 records are belong to another company, Company B. Adding Company B to the audit pool and removing the data related to Company B completes the second iteration. Table 2 shows the FT digits distributions of the actual and expected proportions after the second iteration. If we continue doing this, the histogram for the actual proportions in the modified data set will become closer to their expected proportions while adding more contractors to the audit pool. It may be a good idea to look at some other attributes such as contractor name, the contract specialist or the completion date for the contract after every iteration and run association rules data mining analysis to identify the relation of these attributes to the significant digits. Note that we only looked at one of the first order tests suggested by Nigrini (2012) in our framework. It may be a good idea to apply some other tests if the removal procedure does not improve the fit.

Conclusion

Benford's Law is a powerful tool to understand if certain digit combinations occur in abnormal proportions in a data set. It differentiates itself from traditional statistical data analysis as forensic accounting look for characteristics of outliers rather than the characteristics of the data set itself as a whole. There are many different types of tests available to conduct deeper analysis on a large data set. With the advancements in data mining research and computer technology, better models and frameworks can be implemented to diagnose and identify "sick" data that also take into consideration the difficulties that "big data" may bring as although Benford's Law works well with larger data sets the statistical significance tests it uses suffer from the excess power issue caused by large sample sizes. We will continue working on this framework to create a structural process that is suitable for very large datasets and also capable of running a different number of tests in a sequential order. Lastly, we want to emphasize that, as with all data analysis related to fraud, this analysis suggest the possible occurrence of anomaly or fraud. Fraud examiners must carefully look at each potential incident and perform a full audit to see if it is an error, discrepancy, or fraud.

References

Albrecht, W. S., Albrecht C.O., Albrecht, C.C., and Zimbelman, M.F. (2016). *Fraud Examination*, 5th edition. Boston, MA: Cengage Learning.

Anastasopoulos, N.P. and Anastasopoulos, M.P., (2012). The evolutionary dynamics of audit. *European Journal of Operational Research*, *216*(2), 469-476. https://doi.org/10.1016/j.ejor.2011.06.003

Baesens, B., Van Vlasselaer, V. and Verbeke, W., (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection.* John Wiley & Sons. https://doi.org/10.1002/9781119146841

Berger, A. and Hill, T.P., (2011). A basic theory of Benford’s Law. *Probability Surveys*, *8*(1), 1-126. https://doi.org/10.1214/11-ps175

Bierstaker, J.L., Brody, R.G. and Pacini, C., 2006. Accountants' perceptions regarding fraud detection and prevention methods. *Managerial Auditing Journal*, *21*(5), 520-535. https://doi.org/10.1108/02686900610667283

Cecchini, M., Aytug, H., Koehler, G.J. and Pathak, P., (2010). Detecting management fraud in public companies. *Management Science*, *56*(7), 1146-1160. https://doi.org/10.1287/mnsc.1100.1174

Cleary, R. and Thibodeau J., (2005). Applying digital analysis using Benford’s Law to detect fraud: The dangers of type I errors. *Auditing: A Journal of Practice & Theory* 24 (1): 77-81. https://doi.org/10.2308/aud.2005.24.1.77

Coates, C.J., Florence, R.E. and Kral, K.L., (2002). Financial statement audits, a game of chicken?. *Journal of Business Ethics*, *41*(1-2), 1-11. https://doi.org/10.1023/A:1021355104022

Cook, J., Hatherly, D., Nadeau, L. and Thomas, L.C., (1997). Does cooperation in auditing matter? A comparison of a non-cooperative and a cooperative game model of auditing. *European Journal of Operational Research*, *103*(3), 470-482. https://doi.org/10.1016/s0377-2217(97)00089-1

da Silva, C. G., & Carreira, P. M. (2013). Selecting audit samples using Benford's Law. *Auditing: A Journal of Practice & Theory*, 32(2), 53-65. https://doi.org/10.2308/ajpt-50340

Diekmann, A. and Jann B. (2010). Benford’s Law and fraud detection: Facts and legends.

German Economic Review 11 (3): 397-401. https://doi.org/10.1111/j.1468-0475.2010.00510.x

D.o. Justice . Available at: http://www.justice.gov/opa/pr/20-13/December/13-civ- 1352.html.

Geyer, C. L. and Williamson, P. P. (2004). Detecting Fraud in data sets using Benford’s Law.

Communications in Statistics: Simulation and Computation® 33 (1): 229-246. https://doi.org/10.1081/sac-120028442

Grabinski, K. and Paszek, Z. (2013). Examining Reliability of Large Financial Datasets Using Benford's Law. *Ekonomske Teme* 51 (3): 515-524.

Johnson, G. C. (2009). Using Benford's Law to determine if selected company characteristics are red flags for earnings management. *Journal of Forensic Studies in Accounting and Business* 1 (2): 39-65.

Hill, T.P., (1995). A statistical derivation of the significant-digit law. *Statistical science*, 354-363. https://doi.org/10.1214/ss/1177009869

Johnson, G. C. and Weggenman, J. (2013). Exploratory research applying Benford's Law to selected balances in the financial statements of state governments. *Academy of Accounting and Financial Studies Journal* 17 (3): 31-44.

Jordon, C. E. and Clark, S. J. (2011). Detecting Cosmetic Earnings Management using Benford's Law. *The CPA Journal* 81 (2): 32-37.

Kieso, D.E., Weygandt, J.J., and Warfield, T.D. (2016). *Intermediate Accounting,* 16th ed. New York, NY: John Wiley & Sons, Inc.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, *32*(4), 995-1003.

Mayor-Schonberger, V. and .Cukier, K. (2013). *Big Data.* First Mariner Books edition 2014. New York, NY: Houghton-Mifflin Harcourt Publishing Company.

NHCAA. Available at: http://www.nhcaa.org/resources/health-care-anti-fraud- resources/the-challenge-of-health-care-fraud.aspx

Nigrini, M.J. and Mittermaier, L.J., (1997). The use of Benford's law as an aid in analytical procedures. *Auditing*, *16*(2), p.52.

Nigrini, M.J. (1999). I've got your number. *Journal of Accountancy* 187 (5): 79-83.

Nigrini, M. J. (2009). Data diagnostics using second-order tests of Benford's Law. *Auditing: A Journal of Practice & Theory* 28 (2): 305-324. https://doi.org/10.2308/aud.2009.28.2.305

Nigrini, M. J. (2011). *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations.* Hoboken, N.J.: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118386798

Nigrini, M.J. (2012). *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection.* Hoboken, N.J.: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119203094

Rodriguez, R. J. (2004). Reducing false alarms in the detection of human influence on data.

Journal of Accounting, Auditing & Finance 19 (2): 141-158.

Szilagyi, M.N., (2007). Agent-based simulation of the n-person chicken game. In *Advances in Dynamic Game Theory* (696-703). Birkhäuser Boston. https://doi.org/10.1007/978-0-8176-4553-3_34

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M. and Baesens, B., (2016). Gotcha! Network-based fraud detection for social security fraud. *Management Science.* https://doi.org/10.1287/mnsc.2016.2489

Yang, S. and L. Wei. (2010). Detecting money laundering using filtering techniques: A multiple-criteria index. *Journal of Economic Policy Reform* 13 (2): 159-178 https://doi.org/10.1080/17487871003700796

CITATION INFORMATION

Canbolat, Mustafa, and D. Donald Kent, Jr. "A Data Analysis Based Framework to Detect Anomalies in Large Data Sets Using Benford's Law." *The BRC Academy Journal of Business* 7, no. 1 (2017): 99–110. http://dx.doi.org/10.15239/j.brcacadjb.2017.07.01.ja05

WEB APPENDIX

A web appendix for this paper is available at:

http://dx.doi.org/10.15239/j.brcacadjb.2017.07.01.wa05