POTENTIAL OF BENFORD'S LAW AND MACHINE LEARNING BASED VERIFICATION IN AGRICULTURAL LOGISTICS

STANISLAV LEVIČAR

Vocational College Brežice, Brežice, Slovenia.
E-mail: stanko.levicar@gmail.com

Abstract Food supply chains are becoming increasingly more complex, contributing to emergence of new threats and risks for the involved stakeholders. Additionally, the information technology accelerated development of new and more productive ways of collaboration among organizations (members of supply chains) and helped to optimize their processes. Tighter collaboration among those companies is only possible if sufficient level of trust is established among them, which is often an obstacle that is not easily overcome. Since individual companies (which are part of supply chain) are unable to verify and rely on the data that is provided by third parties, the potential advantages are not fully realized. In this article we try to identify a possibility to remove one important element of this obstacle by using Benford's law as the basis for general-purpose verification tool that is additionally enhanced by statistics based methods of machine learning algorithms that can be implemented in IT supported business operations. The potential usefullness of those methods lies in the fact that they are able to identify the patterns and correlations without explicit users' input.

Keywords: Benford's law; fraud detection; machine learning; supply chain, food.

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-406-4.4
ISBN 978-961-286-406-4

1 Introduction

Since the members of food supply chain are relatively fragmented and on average smaller in size, the synchronization of their activities has turned out to be more and more complex and vulnerable to factors that were previously considered minor relatively to the costs of production and distribution. The IT solutions successfully addressed many of the additional problems that surfaced with such complexity. Tight integration of the IT systems of various companies that form a supply chain has resulted in substantial cost minimizing and time savings, as well as removal of various opportunities for human-based errors and data tampering. The focal point has then shifted towards setting up algorithms for data analysis, and reviewing the resulting output information. More and more of those decisions regarding logistics activities depend on the quality and timeliness of information provided by these information systems. In such environment the concept of 'trust-but-verify' where collaborating parties in the supply chain are generally trusting each other, but at the same time check the information they receive, is being increasingly used for monitoring internal processes as well.

Another important impact that is being felt by SME in agriculture is the changing business as well as natural environment which has multifaceted repercussions for those companies. Globalization and the long-term trend of removal of trade barriers among and inside various economic trade blocs as well as between individual countries has lowered barriers to entry for many foreign producers, and have also caused the environment to be more fluid and unpredictable. Firstly, it is getting ever more difficult for those small producers to assess the risks that are emanating from the environment, and therefore they are unable to make optimal decisions that would maximize their ROI. Secondly, they are usually working in conditions of information asymmetry relatively to larger competitors, as well as other partners and their suppliers (such as insurance providers, larger supply-chains, etc.) (OECD, 2019). Local SMEs in agriculture therefore do not dispose of many good options to absorb the consequences of market volatilities and variabilities within their supply chains.

2 Business model adaptation

Frequently offered solution for the struggling SMEs in agriculture has been for years that they should adapt their business model which usually included finding a niche market and refocusing on produce with higher margins. While such business pivoting surely is a solution for a lot of those companies, it cannot be universal panacea for the majority of businesses in the agricultural sector (Alsos *et al.*, 2011).

Such solutions imply that the existing business models of those SMEs are inherently flawed and are not sustainable, which in many cases is not true. Some of the underlying root causes for diminishing ROI can be eliminated or neutralized – and this way business models of those companies do not need to be substituted, but only enhanced.

The primary component of business model as a concept is usually the value proposition – which is tightly dependent on the price and cost structure of the products and services. The identification of the factors that are influencing them enables companies to better predict the future dynamics in the markets and what measures should they undertake to maximize the probability of achieving the business goals.

3 Reasons for information asymmetry

The problem that many SMEs in agriculture are facing is that they do not have sufficient resources to gain access to data sources as well as know-how needed to process them and transform them into actionable information. One reason is their size which does not permit them to efficiently collect sufficient amount of data, and another is the cost which is usually too high for them to be able to develop their own methods and algorithms which would transform the input data into clear results which would help them make more optimal decisions.

Part of the problem is that each SME has specific business characteristics and features, meaning that it is difficult in advance to prepare IT solutions which would be suitable for large number of SMEs at the same time without modifications which are usually costly and time consuming. Since the environment is constantly changing

those modifications would also require yearly adjustments, thus eliminating important part of the benefits they would supposedly bring.

The information asymmetry is therefore not induced only because of the lack of the quality information sources, but also due to the costs of constant adjustments which would have to be implemented continuously.

4 Data driven decision making

To be able to regain the competitive advantage it is thus not feasible for SMEs to emulate their larger competitors. The decision making process of SMEs should instead be supported by solutions that can produce useful information even if the data sources are limited and raw. Another important aspect of these type of solutions is that it has to be able to automatically adapt without constant intervention and modification from its users.

On the other hand, the process of decision making should still be supported by evidence based on correct data which is collected externally or internally (inside the company). The described problem of limited availability of quality data and the need for constant modifications of the model (without the necessary interventions from the users) can be resolved with the use of solutions that are incorporating statistical or machine learning methods (Finlay, 2018).

Those methods are part of statistics studies that address the before mentioned issue with development of computer algorithms that search for patterns and transform data in usable information. The algorithms of machine learning use advanced statistical methods for analyzing datasets to identify patterns and predict probable outcomes (Lantz, 2015, pp. 3).

The field of machine learning consists of several methods, such as regression algorithms, instance-based algorithms, regularization algorithms, decision tree algorithms, Bayesian algorithms, clustering algorithms, association rule learning algorithms, deep learning algorithms, artificial neural network algorithms, ensemble algorithms, dimensionality reduction algorithms, etc. (Brownlee, 2013). Some of the algorithms are trying to imitate natural processes. Their ability to discover patterns is relatively better in comparison to other types of algorithms, but the complexity of

the rules that it develops "automatically" by learning from real-world experience is often greater than what humans are able to comprehend. This is usually not a problem for most of the fields where such methods can be applied, but there are use cases where the inability to give details of the process and the reasoning that was developed by the algorithm in sufficiently exhaustive manner might prevent the ability of humans to verify it and gauge its reliability.

Nevertheless, this approach still produces the results that are of higher precision in comparison to other techniques – especially if the ROI is taken into account. Examples of machine learning use in agricultural sector are many, especially regarding the production and to an extent regarding market forecast (Razmjooy and Vieira Estrela, 2019), but much less so in the field of business integration within the supply chain, where the correct and robust implementation of the concept of 'trust-but-verify' is the key factor of the success of the supply chain.

5 Benford's law

One way to develop sustainable and cost effective way of 'trust-but-verify' approach is by incorporating the Benford's law based verification into the communication protocols of the agents of the supply chain. The Benford's law is based on the phenomenon of certain significant digits of real numbers probability distribution that was discovered by Simon Newcomb in 1881 and later as well by Frank Benford (1983), by whom the law is named. The main idea of the law is that the digits of certain position in a number in the lists of numbers from many real-life sources of data, appear in a specific proportions. As it follows those proportions are inherent to the numeral system that we use – ie. decimal, but can easily be applied also to any other numeral system. The general formulation that Benford's law presents is that in numeral system with the base b, the lead digit d occurs with the following probability:

$P(d) = \log_b(1 + d^{-1})$, where $d = 1,..., b - 1$

In case of decimal numeral system, that probability would be:

$P(d) = \log_{10}(1 + d^{-1})$, where $d = 1,..., 9$

The above formula is derivated from universal concept, where probability for any significant digit position can be calculated (Hazewinkel 1997). But for most cases the proportions of first digits are useful the most. The leading significant digit is the one which comes first when all preceding zeroes are omitted (regardless of degree), since they do not affect the probabilities of proportions.

The following Figure 1 shows us the Benford's distribution of the leading digits in base 10:

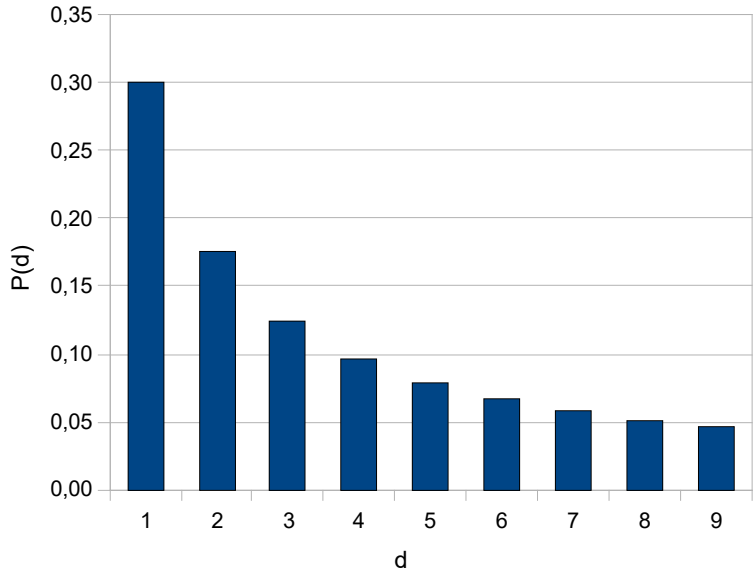

Figure 1: Benford's distribution

In the Table 1 are listed probabilities of first digits according to Benford's law that follows its logarithmic formula:

Table 1: Probability of first digits according to Benford's law formula

First digit	Probability
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

Those expected proportions of frequencies in which usually digits appear, are then compared to obtained data series (from information systems that are in place). The differences that are detected point us to possible error which can be caused by intentional data manipulation, faulty measurement systems or other occurencies that provoke systematic alterations in collected data (Hales *et al.* 2008). Until the law was discovered, scientists believed that the digit probabilities were evenly distributed and that the natural assignments were random (Brown 2005), but with revelation of this Law new way of detecting was introduced that is resistant to more advanced data forging which includes producing values that have many statistical (even derivative) variables similar or equal to accurate data series.

Most famous incorporation of Benford's law is in forensic accounting used by tax collecting agencies in the U.S. (Nigrini 1996) and also for uncovering fraudulent declaration records (Browne 1998). The evidence based on analysis that derives from Benford's law is now considered as valid and legaly admissible in the U.S. There is no reason its use could not be extended as well in other fields and sectors such as agriculture, where certain aspects in this regard were discussed (Hales *et al.* 2009) and were being recognized as viable alternative to other forms of validation procedures, such as statistical sampling (Hales *et al.* 2008), but this area still lacks more concrete proposals of how to efficiently include the law in the information systems that are used for monitoring logistics processes.

6 Requirements for potential use of Benford's law

The Benford's law is valid in situations where set of numbers follows to a logarithmic uniform distribution, which is common to values adherent to many real-world phenomena, like stock prices, birth rates, invoices values, accounting reports, atomic weights of elements, sports statistics (Leemis *et al.* 2000) as well as to certain physical and mathematical constants (Burke and Kincanon 1991). According to Hales *et al.* (2009), the assumptions which have to be met, for the law to be valid, are:

- Numbers must occure naturally, and can not be generated by human intervention.
- The values of the phenomena must not have pre-set limits, breakpoints or other artificial limitations.
- Values must have probability distributions that follow Weibull-like shape.

There are two additional properties of the Benford's law that can be deduced from those assumptions: the first is scale invariance and the second is the validity of the law in multiple probability distributions. Since the law is based on probability proportions of certain digits, and is independent of the numeral base, this means that the units in which the values are measured are not relevant. Another interesting observation regarding the law is that it is still present, even when we mix the values that correspond to previously mentioned assumptions with the ones that are distributed differently (Hill 1995).

7 Machine learning enhancement of Benford's law based verification

Although the Benford's law significantly narrows the testing that is required to identify potentially fraudulent behaviour of the partners in the supply chain, it is often not viable enough if the samples of data are too small, which is often the case with SME in agriculture. This issue can be addressed with the use of the algorithms of machine learning which can have a role as an additional filter in the search for anomalies in the data provided by partners (which may be the result of data tampering). The input of the chosen algorithm would consist of anomalies raised by Benford's law testing, which were subsequently proven to be either false positives or valid fraud attempts. Those algorithms would also take into account the various possible factors that are not processed by Benford's law (like the measurements of

various external and internal variables). After the training phase those algorithms are able to significantly diminish the false positives raised by Benford's law, and therefore enable even smaller businesses in agriculture to efficiently and effectively avoid the costs related to data tampering (or frauds), and would reduce their cost of taking part in the supply chain.

Below is the sample code (for the purposes of demonstration) in programming language Python with installed library *scikit-learn*. The aim of the code is to be able to detect false positives of the Benford's law, and is the following:

```
from sklearn import tree

X = [[input_variable_1, input_variable_2, input_variable_3], […], […]]

Y = ['result', …, …]

clf = tree.DecisionTreeClassifier()

clf = clf.fit(X,Y)

example = [[65, 12, 1, 6, 1]]

prediction = clf.predict(example)

print(prediction)

probability = clf.predict_proba(example)

print(probability)
```

The input variables provide the information about the possible data tampering agent, their characteristics, conditions in which the data was entered in the information system, and the various other external and internal factors that might have an influence on the case. The training phase of the example above also expects to get the information about the proven results of the provided cases. After the training part, it is possible to verify new input data and get the results from the algorithm.

8 Conclusion

For the members of supply chain to be able to integrate more tightly and to gain the competitive advantage it is necessary for them to establish sufficient level of trust among themselves. But even though the relationships and transactions among partners are secured with contracts, the real-time nature of data exchange and just-in-time deliveries require higher level of trust. If a member of supply chain is to rely directly on the data that is provided by its suppliers, he has to have means to verify the validity of those data. But since there are many different sources of data, it would be difficult for them to pre-analyze and define acceptable ranges of the values that are coming from those sources. In this regard the implementation of Benford's law much more directly addresses the problem of data tampering, since it is quite difficult to reproduce the "randomness" of the distribution of individual digits of various values. But since small companies do not have means to collect vast amount of data, which is suitable for Benford's law based verification, this challenge can be significantly amended by applying machine learning algorithms to discern valid anomalies which require additional examination. Additionally, the example shows that even though small and medium sized companies in agriculture do not have the capacity to develop customized solutions, they might nevertheless benefit from machine learning algorithms which can even at a rudimentary level significantly improve the results of the Benford's law based verification and provide clear and provable results that can directly be taken into account during the decision making process. The models used can certainly be expanded and modified, but there are nevertheless many areas where it is not necessary as a first step. More important for SMEs (in the early stages of introducing such methods) is to develop systematic ways of data collection (Kashyap, 2017). One characteristic of agricultural sector is that its performance is influenced by variety of factors, that are often interdependent in many complex relations, especially in supply chains. But the combination of Benford's law and the machine learning algorithms have the potential to decrease

the risks that are emanating from data tampering and effectively lower the costs even for small and medium sized companies in this sector, and thus increase their competitive advantage in the current economic environment.

References

Alsos, G. A., Carter, S., Ljunggren, E., Welter, F. (2011). *The Handbook of Research on Entrepreneurship in Agriculture and Rural Development*. Cheltenham: Edward Elgar Publishing.

Benford, F. (1938). The law of anomalous numbers, *Proceedings of the American Philosophical Society* 78, pp.551–572.

Brown, R. J. C. (2005). Benford's law and the screening of analytical data: the case of pollutant concentrations in ambient air, *The Analyst* 130, pp.1280–1285.

Browne, M. (1998). Following Benford's Law or looking out for number 1, *New York Times* 147 (5), pp.1239–1243.

Brownlee, J. (2013). A Tour of Machine Learning Algorithms. *Machine Learning Mastery*. (Date of publication: 25. 11. 2013.) [WWW] <URL: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/> [Accessed 10 October 2019].

Burke, J., and Kincanon, E. (1991). Benford's law and physical constants: the distribution of initial digits, *American Journal of Physics* 59 (10), pp.952.

Finlay, S. (2018). *Artificial Intelligence and Machine Learning for Business*. Great Britain: Relativistic.

Hales, D. N., Chakravorty, and S.S., Sridharan, V. (2009). Testing Benford's Law for improving supply chain decision-making: A field experiment, *International Journal of Production Economics* 122 (2), pp.606-618.

Hales, D. N., Sridharan, V., Radhakrishnan, A., Chakravorty, S., and Siha, S. (2008). Testing the accuracy of employee-reported data: an inexpen- sive alternative approach to traditional methods, *European Journal of Operational Research* 189 (3), pp.583–593.

Hazewinkel, M. (1997). *Encyclopaedia of mathematics: Supplement*. Norwell, MA: Kluwer Academic Publishers.

Hill, T. P. (1995). A Statistical Derivation of the Significant-Digit Law, *Statistical Science* 10 (4), pp.356-363.

Kashyap, P. (2017). *Machine Learning for Decision Makers*. Bangalore: Apress.

Lantz, B. (2015). *Machine Learning with R*. Birmingham: Packt Publishing.

Leemis, L. M., Schmeiser, B. W., and Evans, D. L. (2000). Survival Distributions Satisfying Benford's Law, *The American Statistician*, 54 (4), pp.236-241.

Newcomb, S. (1881). A note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics* 4 (1), pp.39–40.

Nigrini, M. (1996). A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association* 18 (1), pp.72–91.

OECD (2019). *Digital Opportunities for Better Agricultural Policies*. Paris: OECD Publishing.

Razmjooy, N., Vieira Estrela, V. (2019). *Applications of Image Processing and Soft Computing Systems in Agriculture*. Hershey: IGI Global.