Chapter 10

DETECTING FRAUD USING MODIFIED BENFORD ANALYSIS

Christian Winter, Markus Schneider and York Yannikos

Abstract Large enterprises frequently enforce accounting limits to reduce the impact of fraud. As a complement to accounting limits, auditors use Benford analysis to detect traces of undesirable or illegal activities in accounting data. Unfortunately, the two fraud fighting measures often do not work well together. Accounting limits may significantly disturb the digit distribution examined by Benford analysis, leading to high false alarm rates, additional investigations and, ultimately, higher costs. To better handle accounting limits, this paper describes a modified Benford analysis technique where a cut-off log-normal distribution derived from the accounting limits and other properties of the data replaces the distribution used in Benford analysis. Experiments with simulated and real-world data demonstrate that the modified Benford analysis technique significantly reduces false positive errors.

Keywords: Auditing, fraud detection, Benford analysis

1. Introduction

Financial fraud is a major risk for enterprises. Proactive access restrictions and *post facto* forensic accounting procedures are widely employed to protect enterprises from losses. Many practitioners assume that access restrictions do not impact the effectiveness of forensic methods – if they consider the interdependencies at all. However, this is not necessarily true.

Auditors often use Benford analysis [5] to identify irregularities in large data collections. Benford analysis is frequently applied to accounting and tax data to find traces of fraudulent activity [10]. Benford analysis is based on Benford's law [11], which states that the frequencies of leading digits in numbers follow a non-uniform distribution. This Benford distribution is a logarithmic distribution that decays as the digits

increase. When using Benford analysis to check financial data for irregularities, auditors test the data for conformance with Benford's law.

If an enterprise enforces accounting limits for certain employees, for example, a limit of \$5,000, the frequencies of leading digits in the data created by these employees deviate from the Benford distribution. Since this deviation is much larger than that produced by pure chance, Benford analysis of the data would generate more false positive alerts.

This paper respects the implications of access restrictions (e.g., payment and order limits) by using a log-normal reference distribution derived from the data. The resulting modified Benford analysis compares the frequencies of leading digits in the data to the reference distribution. Applying the modified Benford analysis to simulated and real-world data gives rise to lower false positive rates, which, in turn, reduces auditing costs.

2. Benford Analysis

Benford's law states that numbers in real-world data sets are more likely to start with small digits than large digits [1, 9]. Specifically, the Benford distribution determines the probability of encountering a number in which the n most significant digits represent the integer $d^{(n)}$. The probability of the associated random variable $D^{(n)}$ is given by:

$$\Pr(D^{(n)} = d^{(n)}) = \log(d^{(n)} + 1) - \log(d^{(n)}) = \log\left(1 + \tfrac{1}{d^{(n)}}\right) \quad (1)$$

Benford's law has been shown to hold for data in a variety of domains. Nigrini [10] was the first to apply Benford's law to detect tax and accounting fraud.

The Benford analysis methodology compares the distribution of first digits in data to a Benford distribution. Alerts are raised when there is a large deviation from the Benford distribution.

Benford analysis is typically an early step in a forensic audit as it helps locate starting points for deeper analysis and evidentiary search. In order to identify nonconforming data items (i.e., those needing further investigation), it is necessary to quantify the deviation of the data from the reference Benford distribution. This is accomplished using statistical tests or heuristic methods.

A statistical test quantifies the deviation between the data and the reference distribution using a test statistic. The p-value and significance level α are crucial quantities for assessing the selected test statistic. The p-value is the probability that the test statistic is at least as large as currently observed under the assumption that the data is generated according to the reference distribution. A statistical test yields a rejection

if the p-value is small (i.e., the test statistic is large). The threshold for rejection is specified by the significance level α.

An example is the chi-square test, which uses the chi-square statistic to calculate the p-value. Comparison of the p-value with α may result in rejection. A rejection is either a true positive (i.e., fraud is indicated and fraud actually exists) or a false positive (i.e., fraud is indicated, but no fraud actually exists).

Other measures for determining the deviation include the "mean absolute deviation" and the "distortion factor" [10]. The thresholds for rejection are typically chosen in a heuristic manner for Benford analyses that use these measures.

A limitation of Benford analysis is that non-fraudulent data must be sufficiently close to the Benford distribution. Two techniques are available for determining if the data meets this condition: mathematical approaches [2, 4, 13, 14] and rules of thumb [5, 6, 8, 10, 11, 16, 18].

One rule of thumb is that data is likely close to the Benford distribution if it has a wide spread, i.e., it has relevant mass in multiple orders of magnitude. Because accounting data and other financial data usually have a wide spread, we can assume that this rule does not limit the application of Benford analysis in the accounting and financial domains.

Another rule of thumb is that non-fraudulent data must not artificially prefer specific digits in any position. This automatically holds for natural data with a wide spread. However, human-produced numbers (artificial data) such as prices can be based on psychologically-chosen patterns (e.g., prices ending with 99 cents). But such patterns are more common in consumer pricing than in business and accounting environments.

Another rule of thumb is that Benford analysis should not be performed when the data has an enforced maximum and/or minimum [5, 11]. This is problematic because limits are imposed in many accounting environments. When accounting limits exist, it is only possible to apply Benford analysis to the global data, not to data pertaining to single individuals. This is because the global data does not have enforced limits.

3. Handling Accounting Limits

In order to determine how an accounting limit affects the distribution of leading digits, it is necessary to make an assumption about the overall distribution of data. The cut-off point at an accounting limit is just one property of the overall distribution and is, therefore, not sufficient to derive a reference digit distribution.

The first step in handling an accounting limit is to identify a reasonable distribution model for the accounting data without the cut-off. Unfortunately, a normal distribution does not match the Benford distribution. However, the logarithms of the data values can be assumed to have a normal distribution, i.e., the data has a log-normal distribution. A log-normal distribution is specified by the mean μ and standard deviation σ of the associated normal distribution.

Several researchers [6, 13, 16] have considered log-normal distributions in the context of Benford's law. In general, they agree that conformance with the Benford distribution increases as σ increases. The multiplicative central-limit-theorem argument, which is used to explain the validity of Benford's law, also justifies the use of a log-normal data distribution. Bredl, *et al.* [3] have confirmed that financial data can be assumed to have a log-normal distribution.

The next step in handling an accounting limit is to introduce a cut-off to the log-normal distribution corresponding to the limit. The resulting cut-off log-normal distribution may be used in the analysis.

Thus, the "modified Benford analysis" technique involves:

- Identifying a suitable log-normal distribution.
- Cutting-off the log-normal distribution at the accounting limit.
- Deriving a reference digit distribution from the cut-off log-normal distribution.
- Statistically testing the data against the derived distribution.

A suitable log-normal distribution can be identified by estimating the mean and standard deviation parameters from the data. Unfortunately, it is not known *a priori* if the data contains traces of fraud and where these traces are located. Consequently, the identified distribution is affected by fraudulent and non-fraudulent postings. In general, the influence of fraudulent postings on the estimated parameters is marginal and the distortion in the distribution due to these postings is large enough to be detected during testing.

4. Modified Benford Analysis

Two assumptions are made to simplify the determination of the cut-off log-normal distribution. First, the global data is assumed to have no enforced limits. Second, the distribution of data generated by a single employee is assumed to conform to the global distribution except for cut-offs. This may not be true if the employees have different accounting tasks that do not differ only in the accounting limits.

Based on the assumptions, the mean and standard deviation of the global log-normal distribution are estimated as the empirical mean and standard deviation of the logarithms of the global data values. These values are used to create the reference distribution for the overall data and to calculate a cut-off distribution for individual employees with accounting limits.

4.1 Log-Normal Distribution

The desired log-normal distribution is most conveniently obtained by starting with the normal distribution of logarithms, which has the probability density function $\tilde{g}$ and cumulative distribution function $\widetilde{G}$:

$$\tilde{g}(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\Big(\frac{y-\mu}{\sigma}\Big)^2 \right) \tag{2}$$

$$\tilde{G}(y) = \int_{-\infty}^{y} \tilde{g}(t)\mathrm{d}t \tag{3}$$

Note that the functions associated with the uncut distribution have a tilde ($\sim$) above them to distinguish them from the functions associated with the cut-off log-normal distribution.

The distribution is then transformed to the log-normal distribution by calculating the cumulative distribution function $\widetilde{F}$, followed by the probability density function $\tilde{f}$, which is the derivative of $\widetilde{F}$:

$$\widetilde{F}(x) = \widetilde{G}\big(\log(x)\big) \text{ for } x > 0 \tag{4}$$

$$\tilde{f}(x) = \frac{\tilde{g}(\log(x))}{\ln(10) \cdot x} \text{ for } x > 0 \tag{5}$$

4.2 Cut-Off Limits

Introducing a cut-off requires a rescaling of the distribution to obtain a probability mass of 1.0 over the desired range. Given an upper limit $M \leq \infty$ and a lower limit $m \geq 0$, the updated probability density function and cumulative distribution function are given by:

$$f(x) = \begin{cases} \frac{\tilde{f}(x)}{\widetilde{F}(M)-\widetilde{F}(m)} & \text{for } m \leq x \leq M \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$F(x) = \begin{cases} \frac{\widetilde{F}(x)-\widetilde{F}(m)}{\widetilde{F}(M)-\widetilde{F}(m)} & \text{for } m \leq x \leq M \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

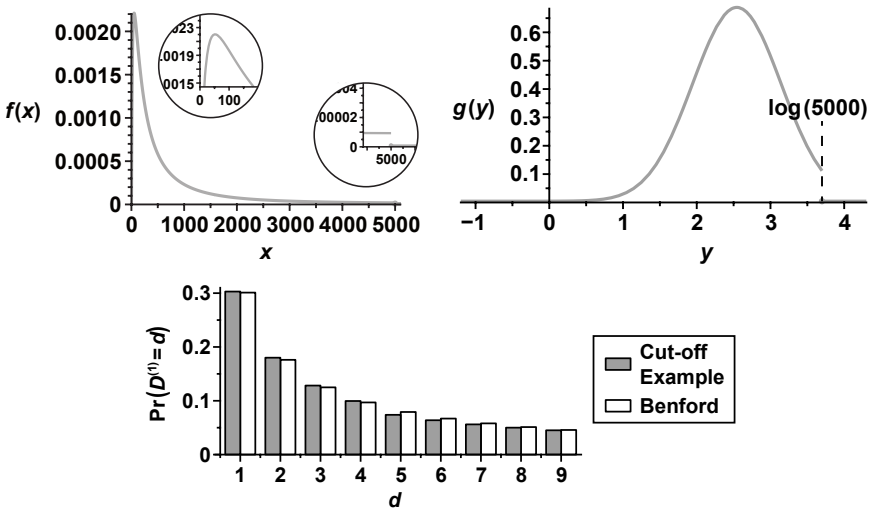

Figure 1. Comparison of cut-off log-normal and Benford distributions.

Similarly, the probability density function g and cumulative distribution function G of the cut-off logarithms are computed using the bounds $m' = \log(m)$ and $M' = \log(M)$.

4.3 Leading Digit Distribution

Computing the distribution of leading digits requires the collection of all numbers $x > 0$ with the same significand $s \in [1; 10)$. These numbers are used to construct the set $\{s \cdot 10^n : n \in \mathbb{Z}\}$. The probability density function θ and cumulative distribution function Θ of the distribution of significands are given by:

$$\theta(s) = \sum_{n \in \mathbb{Z}} f(s \cdot 10^n) \qquad \text{for } s \in [1; 10) \tag{8}$$

$$\Theta(s) = \sum_{n \in \mathbb{Z}} F(s \cdot 10^n) - F(10^n) \qquad \text{for } s \in [1; 10) \tag{9}$$

The computation of the distribution of $D^{(n)}$ uses the distribution of significands. In particular, for $d \in \{1, \ldots, 9\}$, $\Pr(D^{(1)} = d) = \Theta(d+1) - \Theta(d)$.

Our modified Benford analysis technique uses this distribution as the reference distribution in the chi-square test on the leading digits to test for fraud. Figure 1 shows an example with typical accounting parameters specified in U.S. dollars. A cut-off log-normal distribution with

$\mu = \log(350)$, $\sigma = 0.6$ and $M = 5,000$ is compared with the Benford distribution. Although the distribution of first digits differs only slightly from the Benford distribution, the difference could be relevant when analyzing large data samples. Table 1 in the next section shows that Benford analysis yields results of moderate quality for this cut-off log-normal configuration.

4.4 Alternative Setup

If the data only has enforced limits or if the globally-estimated parameters are not suitable for the data generated by an individual employee, then the mean and standard deviation of the global set of logarithms are not suitable parameters. The maximum likelihood method must then be used to obtain suitable parameters. In our case, the maximum likelihood method uses the logarithms of the data values and the density of the cut-off normal distribution to define a likelihood function. An optimization algorithm is employed to determine a local optimum of the likelihood function that yields the parameters of the desired log-normal distribution. Note that this step must deal with cut-offs during the parameter identification step.

5. Results with Synthetic Data

Synthetic accounting data is used to compare the effectiveness of modified Benford analysis versus conventional Benford analysis for two reasons. First, it is difficult to obtain real-world accounting data. Second, it is not possible to control the type and amount of fraud present in real data.

The synthetic data used in the experiments was created by the 3LSPG framework [19]. The simulations produced data corresponding to non-fraudulent and fraudulent employees; the fraudulent employees occasionally made unjustified transactions to accomplices. The fraudsters attempted to conceal their activities by choosing amounts that would be checked less carefully. We assumed that amounts of \$100 or more required secondary approval and, therefore, the fraudsters paid a little less than \$100 (i.e., an amount with 9 as the leading digit) to their accomplices. The frequency of occurrence of fraud was set to 0.01.

Table 1 compares the results obtained using modified Benford analysis (MBA) and conventional Benford analysis (BA) for various distributions. Each analysis used the chi-square test on the first digits with significance $\alpha = 0.05$. The table reports the number of times the tests made rejections over 100 simulations. The rejections correspond to true positive (TP) alerts for fraudsters and false positive (FP) alerts for non-

Table 1. Comparison of modified and conventional Benford analysis.

Distribution Parameters			Sample	BA		MBA	
Limit	μ	σ	**Size**	**TP**	**FP**	**TP**	**FP**
			1,000	11	8	11	8
∞	$\log(1,800)$	0.6	3,000	25	1	25	1
			9,000	84	5	84	5
			1,000	100	99	13	4
5,000	$\log(1,800)$	0.6	3,000	100	100	26	4
			9,000	100	100	91	4
			1,000	14	6	9	5
5,000	$\log(350)$	0.6	3,000	46	15	33	1
			9,000	93	34	80	1

fraudulent employees. The quality of an analysis technique depends on the disparity between the corresponding true and false positive counts.

The conventional Benford analysis results vary according to the limits imposed. The two analysis techniques produce comparable results when an accounting limit is not imposed (limit $= \infty$) because the underlying distribution of data is sufficiently close to the Benford distribution. However, the effectiveness of conventional Benford analysis diminishes when the accounting limit increases the deviation from the Benford distribution. The results show that conventional Benford analysis completely fails for an accounting limit of \$5,000 and $\mu = \log(1,800)$. In the case where $\mu = \log(350)$, conventional Benford analysis distinguishes between fraudulent and non-fraudulent employees. But if one considers the fact that most employees are not fraudsters, the rate of false positives is too high.

The results show that modified Benford analysis performs as well or better than conventional Benford analysis in every instance. The false positive rate from modified Benford analysis is always low, and the rate of detected cases of fraud grows with the sample size because the discriminatory power of statistical tests increases as the sample size increases.

6. Results with U.S. Census Data

The results of the previous section demonstrated that modified Benford analysis is effective regardless of the cut-off log-normal setting. However, while simulated data is guaranteed to match the chosen distribution, real-world data may not fit the log-normal assumption.

This section presents the results obtained with a real-world data set obtained from the U.S. Census Bureau [17]. The data set provides the numbers of inhabitants in U.S. counties according to the 1990 census.

Table 2. U.S. counties with inhabitants within upper and lower limits.

Lower / Upper	0	1K	5K	10K	20K	50K	200K	1M
1K	28							
5K	299	271						
10K	756	728	457					
20K	1,463	1,435	1,164	707				
50K	2,299	2,271	2,000	1,543	836			
200K	2,897	2,869	2,598	2,141	1,434	598		
1M	3,111	3,083	2,812	2,355	1,648	812	214	
∞	3,141	3,113	2,842	2,385	1,678	842	244	30

The advantage of using census data over real-world accounting data is that it can be safely assumed that no fraud exists in the data. Therefore, a Benford analysis technique should result in acceptance; any rejection is a false alert. Indeed, the chi-square test on the first digits yielded $p = 0.063$ – and, thus, no rejection – when using the Benford distribution as reference.

As described earlier, modified Benford analysis requires the computation of the log-normal distribution parameters. The empirical mean and standard deviation of the logarithms of the census data were $\mu = 4.398$ and $\sigma = 0.598$. Using these parameters, the chi-square test in a modified Benford analysis yielded $p = 0.064$. Note that both techniques are applicable to data without cut-offs.

The cut-offs in Table 2 were applied to test the ability of the modified Benford analysis technique to handle cut-offs. The upper and lower cut-off points were used to generate sufficient test cases to compare the accuracy of conventional and modified Benford analysis. The results are presented in Tables 3, 4 and 5. Note that p-values smaller than $2^{-52} \approx$ 2E-16 are set to zero in the tables.

As expected, conventional Benford analysis (Table 3) yields poor results, except for a few cases where the cut-off points introduce minor changes in the distribution. For $\alpha = 0.05$, acceptance occurs in only three cases (bold values).

A quick fix to conventional Benford analysis that respects the limits is implemented by changing the Benford distribution of the first digits to only include the possible digits. The digits that were not possible were assigned probabilities of zero while the probabilities for the possible digits were scaled to sum to one. Table 4 shows that this technique yields a marginal improvement over conventional Benford analysis with four (as opposed to three) acceptance cases.

Table 3. Benford analysis (p-values).

Lower / Upper	0	1K	5K	10K	20K	50K	200K	1M
1K	3E-06							
5K	0	0						
10K	0	0	0					
20K	0	0	0	0				
50K	0	0	2E-15	0	0			
200K	1E-04	9E-05	5E-15	0	0	0		
1M	**0.097**	**0.072**	1E-07	0	0	0	0	
∞	**0.063**	0.041	3E-08	0	0	0	9E-16	1E-04

Table 4. Benford analysis with digit cut-off rule (p-values).

Lower / Upper	0	1K	5K	10K	20K	50K	200K	1M
1K	3E-06							
5K	0	0						
10K	0	0	0.019					
20K	0	0	1E-11	**1.000**				
50K	0	0	2E-15	0.008	0.032			
200K	1E-04	9E-05	5E-15	0	0	3E-09		
1M	**0.097**	**0.072**	1E-07	0	0	0	0.003	
∞	**0.063**	0.041	3E-08	0	0	0	9E-16	1E-04

Table 5. Modified Benford analysis (p-values).

Lower / Upper	0K	1K	5K	10K	20K	50K	200K	1M
1K	**0.575**							
5K	**0.691**	**0.549**						
10K	6E-04	3E-04	**0.510**					
20K	7E-04	9E-04	**0.275**	**1.000**				
50K	8E-04	7E-04	0.001	4E-04	0.037			
200K	0.032	0.025	0.005	2E-05	1E-06	**0.711**		
1M	**0.081**	**0.068**	0.044	0.007	1E-04	**0.302**	7E-06	
∞	**0.064**	**0.054**	0.033	0.005	0.001	**0.588**	1E-07	9E-05

The results in Table 5 show that modified Benford analysis yields much better results – the number of acceptances is thirteen. This result has to be qualified, however, because acceptance occurs in the cases where the cut-offs do not introduce much distortion and where there are

relatively few samples left after the cut-offs are performed. The results show that the log-normal distribution is not ideally suited to the census data. Nevertheless, modified Benford analysis yields significantly better results than conventional Benford analysis for data with cut-offs.

7. Related Work

Several researchers have defined adaptive alternatives to the Benford distribution. One approach [8] addresses the issue of cut-offs by adjusting the digit probabilities in a manner similar to our quick fix. Other approaches [7, 15] employ parametric distributions of digits that are fitted to observed digit distributions by various methods. The latter approaches, however, are not designed to discover irregularities.

Other researchers, e.g., Pietronero, *et al.* [12], start with a suitable distribution model for the data, which they use to derive a reference distribution of digits. They use power laws that are relevant to their domains of application. Note however, that while the approach is similar to the modified Benford analysis technique presented in this paper, it does not address the issue of cut-off points.

8. Conclusions

The modified Benford analysis technique overcomes the limitation of conventional Benford analysis with regard to handling access restrictions. The technique reduces false positive alerts and, thereby, lowers the costs incurred in forensic accounting investigations. The false positive rate is independent of the accounting limits because the modified Benford analysis technique adapts to the limits.

The results obtained with synthetic and real-world data demonstrate that modified Benford analysis yields significant improvements over conventional Benford analysis. Our future research will conduct further assessments of the effectiveness of the modified Benford analysis technique using real-world accounting data and fraud cases. Additionally, it will compare the modified Benford analysis technique with other Benford analysis formulations, and identify improved distribution models that would replace the log-normal model.

Acknowledgements

This research was supported by the Center for Advanced Security Research Darmstadt (CASED), Darmstadt, Germany.

References

[1] F. Benford, The law of anomalous numbers, *Proceedings of the American Philosophical Society*, vol. 78(4), pp. 551–572, 1938.

[2] J. Boyle, An application of Fourier series to the most significant digit problem, *American Mathematical Monthly*, vol. 101(9), pp. 879–886, 1994.

[3] S. Bredl, P. Winker and K. Kotschau, A Statistical Approach to Detect Cheating Interviewers, Discussion Paper 39, Giessen Electronic Bibliotheque, University of Giessen, Giessen, Germany (geb.uni-giessen.de/geb/volltexte/2009/6803), 2008.

[4] L. Dumbgen and C. Leuenberger, Explicit bounds for the approximation error in Benford's law, *Electronic Communications in Probability*, vol. 13, pp. 99–112, 2008.

[5] C. Durtschi, W. Hillison, and C. Pacini, The effective use of Benford's law to assist in detecting fraud in accounting data, *Journal of Forensic Accounting*, vol. V, pp. 17–34, 2004.

[6] N. Gauvrit and J.-P. Delahaye, Scatter and regularity imply Benford's law ... and more, submitted to *Mathematical Social Sciences*, 2009.

[7] W. Hurlimann, Generalizing Benford's law using power laws: Application to integer sequences, *International Journal of Mathematics and Mathematical Sciences*, vol. 2009, id. 970284, pp. 1–10, 2009.

[8] F. Lu and J. Boritz, Detecting fraud in health insurance data: Learning to model incomplete Benford's law distributions, *Proceedings of the Sixteenth European Conference on Machine Learning*, pp. 633–640, 2005.

[9] S. Newcomb, Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics*, vol. 4(1), pp. 39–40, 1881.

[10] M. Nigrini, *Digital Analysis Using Benford's Law*, Global Audit Publications, Vancouver, Canada, 2000.

[11] M. Nigrini and L. Mittermaier, The use of Benford's law as an aid in analytical procedures, *Auditing: A Journal of Practice and Theory*, vol. 16(2), pp. 52–67, 1997.

[12] L. Pietronero, E. Tosatti, V. Tosatti and A. Vespignani, Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf, *Physica A: Statistical Mechanics and its Applications*, vol. 293(1-2), pp. 297–304, 2001.

[13] R. Pinkham, On the distribution of first significant digits, *Annals of Mathematical Statistics*, vol. 32(4), pp. 1223–1230, 1961.

[14] R. Raimi, The first digit problem, *American Mathematical Monthly*, vol. 83(7), pp. 521–538, 1976.

[15] R. Rodriguez, First significant digit patterns from mixtures of uniform distributions, *American Statistician*, vol. 58(1), pp. 64–71, 2004.

[16] P. Scott and M. Fasli, Benford's Law: An Empirical Investigation and a Novel Explanation, Technical Report CSM 349, Department of Computer Science, University of Essex, Colchester, United Kingdom, 2001.

[17] U.S. Census Bureau, Population Estimates – Counties, Washington, DC (www.census.gov/popest/counties), 1990.

[18] C. Watrin, R. Struffert and R. Ullmann, Benford's law: An instrument for selecting tax audit targets? *Review of Managerial Science*, vol. 2(3), pp. 219–237, 2008.

[19] Y. Yannikos, F. Franke, C. Winter and M. Schneider, 3LSPG: Forensic tool evaluation by three layer stochastic process based generation of data, in *Computational Forensics*, H. Sako, K. Franke and S. Saitoh (Eds.), Springer, Berlin, Germany, pp. 200–211, 2011.