

A DATA WAREHOUSE DESIGN FOR THE DETECTION OF FRAUD IN THE SUPPLY CHAIN BY USING THE BENFORD'S LAW

¹Cornelia Kraus and ²Raul Valverde

¹Teradata Corporation, Augsburg, Germany

²Department of Supply Chain Management and Technology Management, Concordia University, Montreal, Canada

Received 2014-04-20; Revised 2014-05-07; Accepted 2014-07-07

ABSTRACT

Large data volumes and the inability to analyse them enables fraudulent activities to go unnoticed in supply chain management processes such as procurement, warehouse management and inventory management. This fraud increases the cost of the supply chain management and a fraud detection mechanism is necessary to reduce the risk of fraud in this business area. This study was carried out in order to develop a data warehouse design that supports forensic analytics by using the Benford's law in order to detect fraud. The approach relies on a generic and re-usable store procedure for data analytics. The data warehouse was tested with two datasets collected from an operational supply chain database from the inventory management and warranty claims processes. The results of the research showed that the supply chain data analyzed obeys to Benford's theory and that parameterized stored procedures with Dynamic SQL provide an excellent tool to analyze data in the supply chain for possible fraud detection. The implications of the results of the study are that the Benford's law can be used to detect fraud in the supply chain with the help of parameterized stored procedures and a data warehouse, this can ease the workload of the fraud analyst in the supply chain function. Although the research only used data from the inventory management and warranty claim processes, the proposed store procedures can be extended to any process in the supply chain making the results generalizable to the supply chain management process.

Keywords: Supply Chain Fraud, Supply Chain Management Systems, Benford's law, Fraud Detection, Data Mining for Fraud, Accounting Information Systems

1. INTRODUCTION

The amounts of data produced each day increase steadily and demand companies to either surrender or rise to the challenges and opportunities that this trend brings along. Large data volumes and the inability to analyse or even process them, enables fraudulent tendencies to go unnoticed, even when new tracking devices or automatic scans are introduced and are determined to prevent those tendencies, but also add to the amount of data. Business areas like Supply chains become more complex as the world is getting smaller, trading and shipment is done throughout the globe and the number of involved parties, systems and transactions is rising. Design and Implementation methodologies for systems to process that data must be aligned to keep pace and systems must offer new concepts to not only process the volumes of data, but also in terms of velocity and variety. Generic approaches that enable analysts to apply algorithms fast without re-implementing them but to

concentrate on the definition of the analytical dataset and the interpretation of the results are necessary for effective and successful data warehouse design and implementation. Solutions to address these challenges are more or less non-existent, data warehousing mostly focuses on reporting issues and a stable data basis, while forensic disciplines discuss algorithms and pretend the "perfect" data loaded, stored and ready for use. The goal of the project is to develop a data warehouse design to support analytics, reporting and data mining and forensic analytics by using the Benford's law within the area of supply chain. This is to detect fraud or to point out data anomalies that are worth further examination by subject matter experts.

1.1. Supply Chain Fraud and Data Mining

Risk consulting company Kroll indicates that "the same information technologies that help shippers manage global supply chains could make them more vulnerable to supply-chain fraud" (Hoffman, 2008). The detailed fraud report of 2008 points out that especially the increasing

Corresponding Author: Raul Valverde, Concordia University

demand for natural resources often forced to act fast, exploring new sites, setting them up for production and find supplier and logistic companies to provide energy, staff and the like. As these sites are often in remote areas, the number of available suppliers are quite limited and this might lead responsible managers into situations where “flexibility” and the need to “make things happen” cause non-legal or at least questionable actions. Dependency on one supplier, conflict of interest or bribery may be some of those actions taken (Kroll Advisory, 2010).

Hesketh (2010) however points out that within the supply chain data is not uniformed, mostly not administrated properly and can be highly inaccurate or incomplete, which makes the whole process of the movement of goods in-transparent to both buyer and seller. Not describing the shipped goods creates a whole list of risks for all parties involved, like to the carrier if for example the weight is not properly defined, what seems to be common practice in the container ship industry (which is hereby the only sector of that industry) and “is due to many shippers deliberately under-declaring containers’ weights in order to minimize import taxes calculated on cargo weight.” This lack of information, data collection leads to a lack of knowledge, reduces the options for improvement and even enables criminals or shady operators to gain advantage of these deficits, causing damage of 20 billion USD annually.

The demand for data analytics in fraud prevention is high. Mayne (2010) points out and gives the example of enterprises that are using forensic methods to predict supply chain fluctuations, often driven by the need to improve their business, follow compliance rules or concerns regarding data breaches. Data analytics has a history of successful application in fraud detection; Massa and Valverde (2014) were able to implement a fraud detection system based of anomaly detection data analysis techniques to predict computer intrusion attacks in e-commerce web applications. Mishra *et al.* (2013) used data mining in the area of credit card transactional fraud for the Indian market. Forensics has become more and more the responsibility of IT departments, as they administer the data, maybe on multiple servers spread across the world, with virtual storage (“cloud solutions”) involved, making forensic analysis not an easy task.

According to Hu (2009), the computer forensics process mainly consists of obtaining, identifying and analysing evidence. The author points out that one problem with forensic evidence is that already the collection might alter the original state of the evidence (such as last-accessed timestamp for a file). This seems highly relevant when it comes to intrusion detection and other more network or system breach related issues, but for legacy supply chain management systems, tax-compliance and accounting in general, most transactions are still to be preserved on paper or at least in the form they were transmitted.

Regarding the security of e-procurement solutions, security violations continue to rise, in number as well as frequency note Stephens and Valverde (2013). The more

information technology is the basis to drive business value, the more important and critical transaction data becomes.

Data Mining is defined by Nirkhi *et al.* (2012) as “part of the interdisciplinary field of knowledge discovery in databases”, which made its first steps in the 1980s and grew rapidly within the next decade and made its appearance within the area of digital forensics to gain insights from large volumes of data.

Hand *et al.* (2001), cited by Nirkhi *et al.* (2012) describe the formal methodology of data mining by the following steps:

- Determine type and structure of the datasets
- Decide how best to quantify the data
- Choose proper algorithmic processes for a scoring function
- Decide what data management principles are needed to implement the algorithms efficiently

Additionally Nirkhi *et al.* (2012) point out how important visualization is for the digital investigators, as it enables them to determine vital information “rapidly and efficiently”.

Identifying data mining algorithms or methods to be used in the area of digital forensics in general and fraud detection in particular was either approached in general terms often by mere descriptions of the need of doing so or very detailed mathematical articles, demonstrating the statistical correctness of a method or algorithm. Both extremes were not included into the research for this project.

1.2. Benford’s Law Theory

Benford’s law may be categorized as a descriptive data mining method, as it discriminates data, but also as predictive, as it identifies characteristics of datasets that may help to predict future schemas.

There is a large amount of research on Benford’s law, Durtschi *et al.* (2004) counted more than 150 articles over the period of 50 years, approaching it from many aspects, like mathematical integrity to use in the economics. As most of the mathematical theory is beyond the focus of this project, only articles discussing Benford in general or with relation to fraud detection were regarded.

These expected frequencies of digits in lists of numbers were described by Benford (1938) and verified but also controversially discussed in numerous papers until today. The use of the method is widespread, Geyer and Marti (2012) applied it to volcanological data and indicate their datasets follow Benford’s law and that their “analysis shows that excessive data round-off, data errors, or anomalies may be detected when comparing the data with Benford’s law expected frequencies”.

On the other end of the discipline scale, Nigrini and Mittermaier (1997) introduce in their study Benford’s Law as a method for accounting auditing, where analytical procedures are required and successfully apply

the law to a sample of data from an oil company. However, Özera and Babacanb (2013) found that the off-balance sheet account disclosures of a Turkish bank does not comply to Benford’s law but admit that if “a data set does not comply with Benford Distribution is enough to be suspicious about that data set”.

Nigrini (1999) listed multiple applications for Benford’s law and digital analysis, some of which give a hint as to where the analysis might be helpful within the supply chain process:

- Estimations in the general ledger
- The relative size of inventory unit prices among locations
- Duplicate payments
- Processing inefficiencies due to high quantity/low dollar transactions
- Customer refunds

Benford found out that numbers with low first digits occurred more often, after he observed that in a book with logarithmic **Table 1** the pages with low first digits (1,2) were more worn out than the ones for high first digits (8,9). He then derived the expected frequencies of the digits:

$$P(d = d_1) = \log_{10}(1 + 1/d_1) \text{ for } d \in \{1, 2, \dots, 9\}$$

where, d is a number $\{1, 2, \dots, 9\}$ and P is the probability (Nigrini, 2012).

Table 1. Expected Frequencies for 1st to 4th digits (Source: Nigrini, 1996)

Digit	1st (%)	2 nd (%)	3rd (%)	4th (%)	5th or greater (%)
0		11.97	10.18	10.02	10.00
1	30.10	11.39	10.14	10.01	10.00
2	17.61	10.88	10.10	10.01	10.00
3	12.49	10.43	10.06	10.01	10.00
4	9.69	10.03	10.02	10.00	10.00
5	7.92	9.67	9.98	10.00	10.00
6	6.69	9.34	9.94	9.99	10.00
7	5.80	9.04	9.90	9.99	10.00
8	5.12	8.76	9.86	9.99	10.00
9	4.58	8.50	9.83	9.98	10.00
	100.00	100.00	100.00	100.00	100.00

The first-two digit test is regarded as a more focussed test than the first digit test by Nigrini (2012) and “is there to detect abnormal duplications of digits and possible biases in the data:

$$P(d = d_1, d_2) = \log_{10}(1 + 1/[10 * d_1 + d_2]) \text{ for } d \in \{10, 11, \dots, 99\}$$

2. MATERIALS AND METHODS

The design approach for the data warehouse that was followed is described below:

- Discover the entities required for the data model for the data warehouse by examining the typical processes and events in supply chain management
- Interview two forensic consultants at Teradata Germany with experience in supply chain in order to identify a set of rules that will be use to identify attributes for the different entities of the data model for the data warehouse. This part of research will be focused on the warranty claims and inventory management processes. Although only two processes were selected for this part of the study, the same procedure can be followed for all the different processes of the supply chain
- Design a stored procedure that will dynamically create views for Benford Analysis
- Identify suitable datasets to test data warehouse and store procedure
- Generates a set of views for interpretation and visualization of the data

3. RESULTS

Based on the analysis of Hesketh (2010) that provides an excellent overview of the processes, events and involved parties within the supply chain management, the following subject areas were identified as the main source of entities for our data warehouse design:

- Party: All Individuals, companies, suppliers, buyers, consignors and the organization of the company itself
- Geography
- Locations like plants, distribution centers, ports, routes, GPS coordinates and street addresses
- Time
- Item: The parts produced, ordered, preassembled
- Item instance: Some items might be uniquely identified, like engines
- Procurement: Purchase order and associated Sales orders
- Inventory: Inventory state (snapshots), In and out Events, lot-sizes
- Shipment: The pre-order of containers, vessel space, the Advanced Shipment Notice, number of parts to be shipped with estimated time of arrival
- Issue Management (like Warranty and Returns) damaged part returns, the warranty process and terms

The major entities of every subject area were identified and an entity relationship data model for the data warehouse created (**Fig. 1**).

Two forensic consultants with Teradata Germany with experience in supply chain were interviewed in order to identify a set of rules that can be used to detect fraud in warranty claims and inventory management that

is the main scope of this research. These rules will be used to identify attributes for the different entities of the data model for the data warehouse. The rules identified are summarized below:

Inventory: Unpacking a container that was delivered to a production plant implies transactions to the inventory of the plant. Boxes within the container might be damaged along with the contained goods. By comparing the content with the Purchase order and bill of lading deviations might become obvious. Claims to consignor, packaging facility or supplier might be created on this basis. Fraud can be concealed by the taking of a small amount of parts from a number of boxes, then blaming the supplier for shortages or declaring items as damaged and then blaming the consignor or packaging facility.

Returns and Warranty Claims: Claims may occur at certain stages of the supply chain or affect it end-to-end. For example, a car manufacturer uses spark plugs from a supplier to build its cars. The car is shipped to the dealer who sells it to a customer. Three months later the customer claims to the dealer that his car does not ignite. Inspection shows that the spark plugs are malfunctioning so the dealer will exchange the plugs but open a warranty claim to the manufacturer of the car. The car manufacturer then will raise another warranty claim to the manufacturer of the plug. Not all failed parts are sent back to the original supplier for inspection (regarding to the costs and amounts) but payments are granted either on a subset that was inspected or on good-will.

Receiving goods: Goods might be stolen or damaged on a frequent basis, maybe depending on the numbers of parts delivered or price of the parts.

3.1. Store Procedure Design

A stored procedure was designed to dynamically create four types views, based on the input-parameters handed over with the procedure call. The design is depicted in **Fig. 2**. The store procedure's parameters are described in the **Table 2** below.

The store procedure has four views. The VV_BENFORD_STAT view contains statistical information about the dataset chosen. VV_BENFORD_DET view provides detailed information of the Bendford analysis and helps to determine the first and first two digits of a set of numbers. The VV_BENFORD_FIRST_RES counts how often the first digit appeared in the dataset while the VV_BENFORD_TWO_RES counts how often the first two digits appeared in the dataset. The results of the views provides a Z score that measures how statistically different the actual result is from the expected result and therefore attempts to help answer the question of how confident one can be about the observed relationship.

3.2. Data Collection

Two datasets were collected from projects developed by Teradata Germany. Both datasets are car manufacturer supply chain/quality data. As the data was not allowed to leave the company premises, all testing was done on-site and on anonymised production data. Therefore the developed objects were deployed to a sandbox on the production system and executed.

The first dataset (**Fig. 3**) reflects two years of warranty claims from dealers towards the car manufacturer regarding defective parts and consequential costs.

The second data set (**Fig. 4**) is much smaller dataset reflects one year of inventory data, filtered only for transactions that have a transaction type = 'LOST', which might include thefts, damaged or lost items.

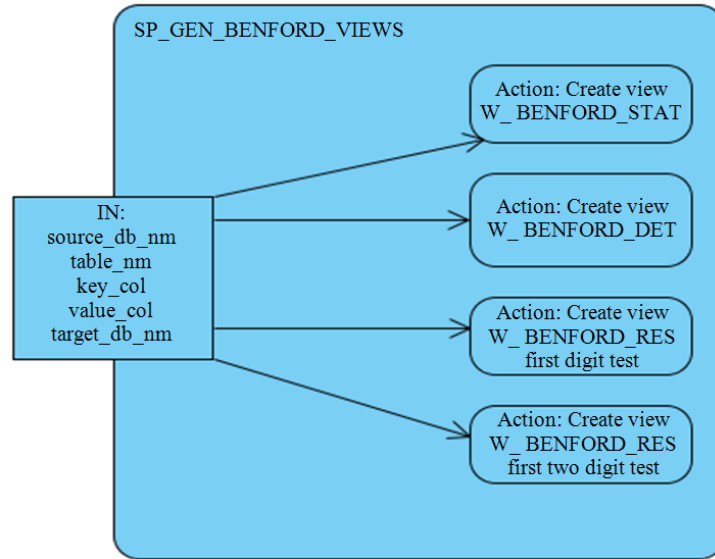


Fig. 2. Activity diagram for stored procedure SP_GEN_BENFORD_VIEWS

record_cnt	value_zero_cnt	value_avg	value_no_zero_avg	value_max_a	value_min_am	
1	2191848	828557	72,7045623875378	116,891661178721	19308,8500	0,0000

Fig. 3. Statistical information of dataset 1

record_cnt	value_zero_cnt	value_avg	value_no_zero_avg	value_max_amt	value_min_amt	
1	3994	394	710,5888	788,3588	86439,5400	0,0000

Fig. 4. Statistical information of dataset 2

```

Query (14.0 VM)
01 REPLACE PROCEDURE benford.SP_GEN_BENFORD_VIEWS
02 (
03     IN source_db_nm CHAR(30),
04     IN table_nm CHAR(30),
05     IN key_col CHAR(30),
06     IN value_col CHAR(30),
07     IN result_db_nm CHAR(30)
08 )
09
10 BEGIN
11
12 CALL DBC.SYSEXECSQL('REPLACE VIEW ' || trim(both from : result_db_nm) || '.W_BENFORD_STAT
13 AS
14 SELECT
15     COUNT(*) AS record_cnt,
16     SUM(CASE WHEN ' || : value_col || ' = 0 THEN 1 ELSE 0 END) AS value_zero_cnt,
17     AVG(' || : value_col || ') AS value_avg,
18     AVG(CASE WHEN ' || : value_col || ' = 0 THEN NULL ELSE ' || : value_col || ' END)
19     AS value_no_zero_avg,
20     MAX(' || : value_col || ') AS value_max_amt,
21     MIN(' || : value_col || ') AS value_min_amt
22 FROM ' || TRIM(BOTH FROM :source_db_nm) || '.' || TRIM(BOTH FROM :table_nm) || ' ';
23
24

```

Fig. 5. Call of dynamic SQL within a stored procedure


```
Query (14.0 VM)
01 CALL benford.SP_GEN_BENFORD_VIEWS ('benford', 'T_SAMPLE_SET', 'sample_id', 'sample_amt', 'benford');
```

Fig. 6. Stored procedure call

```
Answerset 1
REPLACE VIEW benford.VV_BENFORD_STAT
AS
SELECT
    COUNT(*) AS record_cnt,
    SUM(CASE WHEN sample_amt = 0 THEN 1 ELSE 0 END) AS value_zero_cnt,
    AVG(sample_amt) AS value_avg,
    AVG(CASE WHEN sample_amt = 0 THEN NULL ELSE sample_amt END) AS value_no_zero_avg,
    MAX(sample_amt) AS value_max_amt,
    MIN(sample_amt) AS value_min_amt
FROM benford.T_SAMPLE_SET;
```

Fig. 7. Result of dynamic SQL after procedure call

	Name	Type	#
	SP_GEN_BENFORD_VIEWS	Procedure	
	T_BENFORD	Table	
	T_SAMPLE_SET	Table	
	VV_BENFORD_DET	View	
	VV_BENFORD_FIRST_RES	View	
	VV_BENFORD_FIRST_TWO_RES	View	
	VV_BENFORD_STAT	View	

Fig. 8. All used and generated database objects

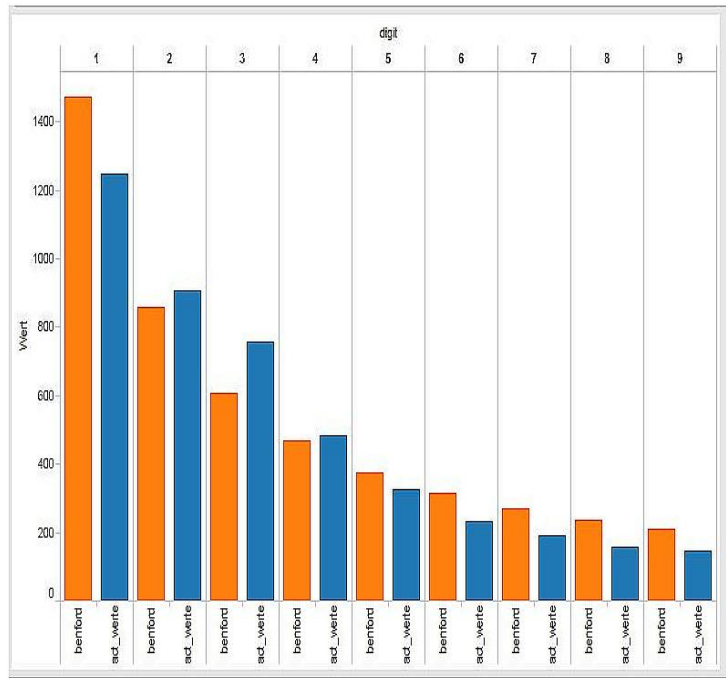


Fig. 9. Dataset 1 frequencies (blue) compared to Benford's expected frequencies (orange)

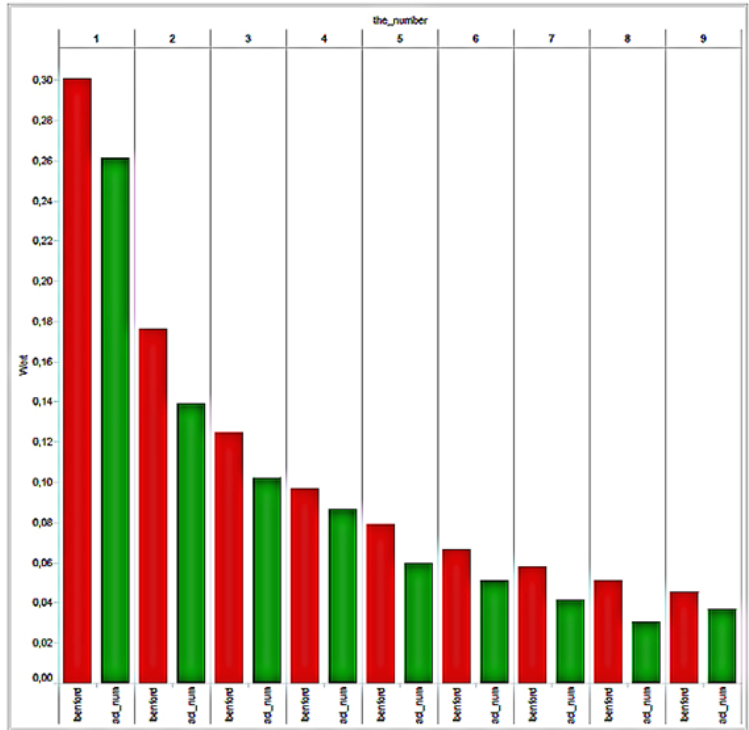


Fig. 10. Dataset 2 frequencies (green) compared to Benford's expected frequencies (red)

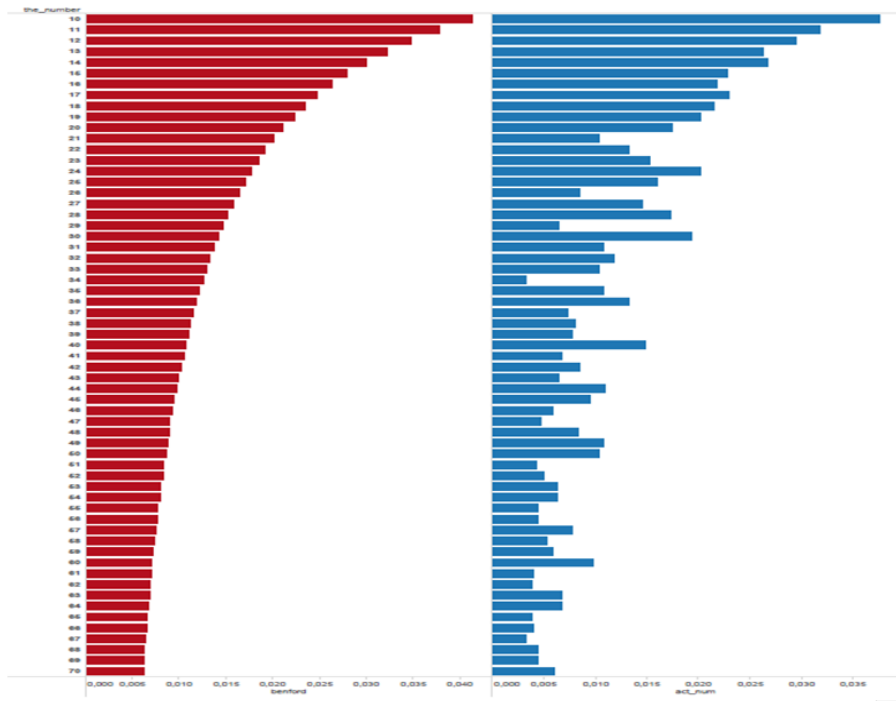


Fig. 11. Dataset 2- first two digit test (Benford in red, actual values in blue)

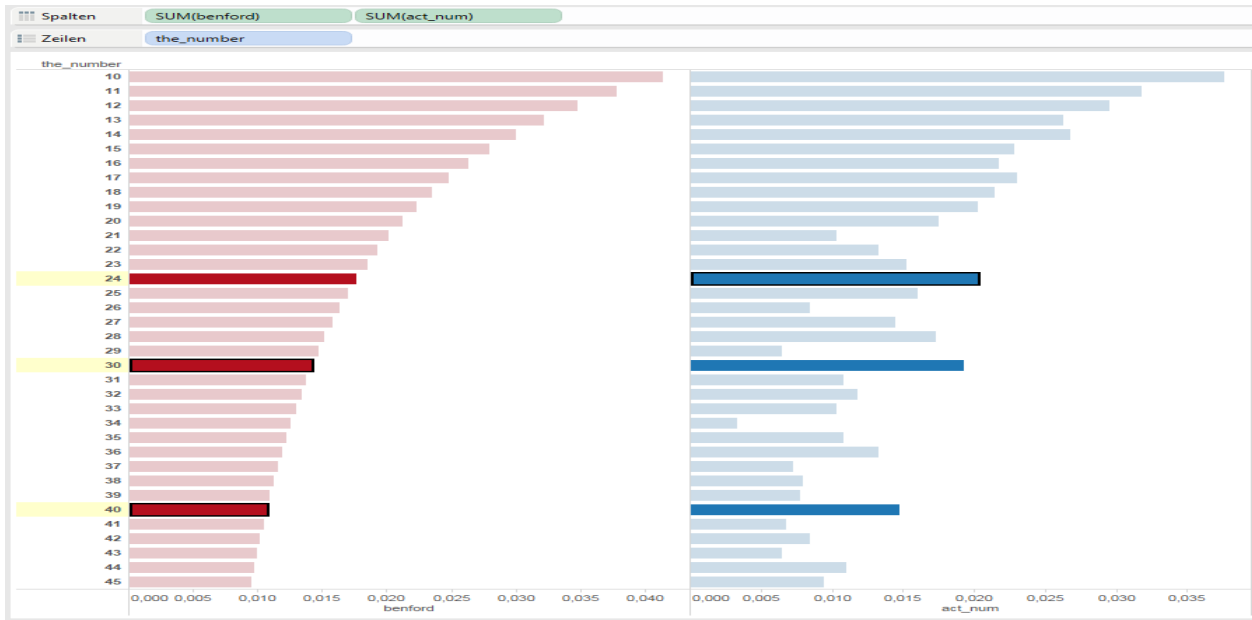


Fig. 12. Dataset 2- some obvious deviations (range 10 to 45 for first two digits)

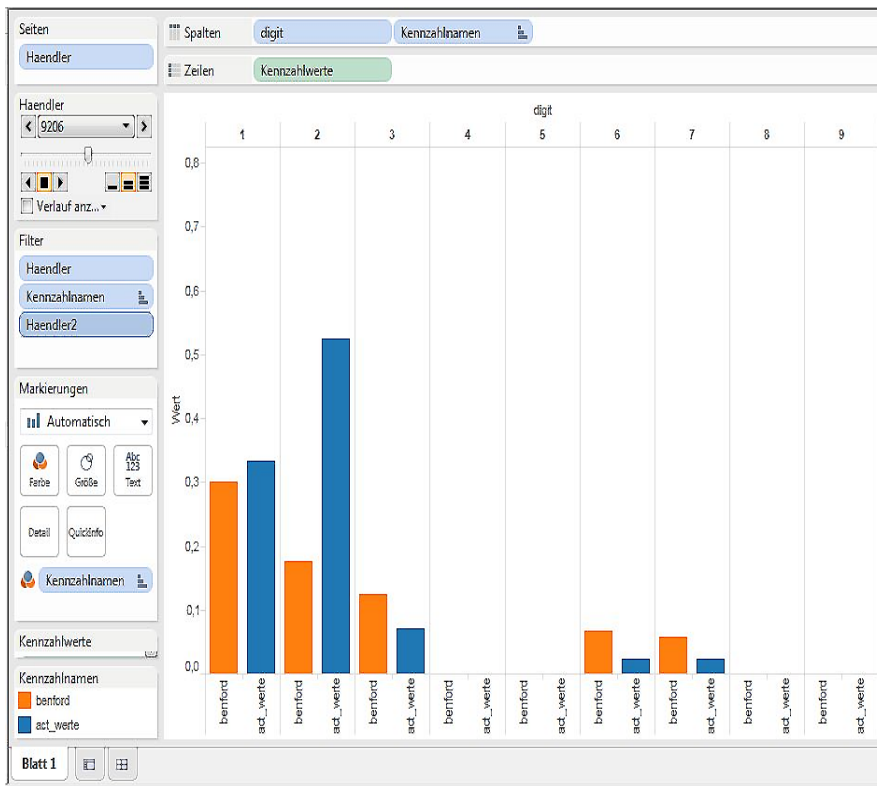


Fig. 13. Tableau Screen showing one deviation from Benford

3.3 Benford Implementation

The Benford probabilities were calculated and stored in a lookup table with SQL-Statements one at a time for the numbers from 1 to 9 and then from 10 to 99 so that the first digit test as well as the first-two digit test can be performed. In case also the second, third or fourth digit test needs to be performed, the used pattern can be extended easily. As described in section 3.1, the stored procedure is intended to create 4 views, which is where the Dynamic SQL is used. For the first view (Statistical information about the chosen dataset) the call for generating the view within the stored procedure is depicted in **Fig. 5**.

The stored procedure is called handing over the IN-Parameters as depicted in **Fig. 6**.

The IN-Parameters are replaced upon execution within the SYSEXEC SQL call, creating a valid SQL statement which is executed directly and results in the creation of the view as depicted in **Fig. 7**.

Once the stored procedure is successfully executed with a first set of data (in T_SAMPLE_SET), all used and created database objects are shown below in **Fig. 8**.

Third or fourth digit tests might be implemented following the same pattern.

3.4 Visualization

Tableau (www.tableausoftware.com) was used for visualisation. The views provided by the stored procedure were the basis for reporting and via the VV_BENFORD_DET view any connection to related information could be achieved. Following these guidelines, three visualisation-sets were created and used to display the three datasets available.

Figure 11 Dataset 2-first two digit test (Benford in red, actual values in blue)

The Visualisation in **Fig. 9** depicts dataset 1. Benford's Values for first digits are in orange, the result set values in blue. As the pattern of the actual values align with the Benford-values, this dataset obeys to Benford's Law, however some values step out marginally, like for the digits 1, 2 and 3. The Visualisation in **Fig. 10** depicts dataset 2. Benford Values for first digits are in red, the result set values in green. As the pattern of the actual values align with the Benford-values, this dataset obeys close to Benford's Law, even more than Dataset 1. For a more detailed look into large datasets, two options were chosen for dataset 2 to find peculiarities. This can be regarded as the next step after identifying peculiarities in **Fig. 12**.

Example for groupings applied next:

- A grouping based on vendors

- Only vendors having a deviation from Benford's law > 0.15 for at least one digit

While reviewing the result, vendors with low numbers of claims had high deviations from Benford. Therefore a filter was applied that only vendors with more than 20 claims in the period were considered.

Figure 13 shows the result of applying those to filters for one particular vendor (#9206).

4. DISCUSSION

For Benford's first digit test, both datasets show no outstanding peculiarities and follow the pattern of Benford's law. According to Nigrini (1996), if the complete dataset follows Benford, then also subsets (not grouped by digits, of course) should also comply.

In **Fig. 11**, the first two digit test is depicted with Tableau. Here, a slightly different chart type was chosen, as the number of digits was too high to assign them to the x-axis and still get a chart to fit on the screen without scrolling.

Overall, the dataset obeys also to Benford's law (first two digit test), as the pattern matches for most of the first two digits to Benford's numbers. The visualisation is much more granular and more details need to be reviewed to find peculiarities.

The analyst would use this chart to identify deviations and "drill" down for some to review the source of the deviation. In **Fig. 12** three of the values that deviate from Benford's value are highlighted.

Overall, the dataset obeys also to Benford's law (first two digit test), as the pattern matches for most of the first two digits to Benford's numbers. The visualisation is much more granular and more details need to be reviewed to find peculiarities.

The analyst would use this chart to identify deviations and "drill" down for some to review the source of the deviation. In **Fig. 12** three of the values that deviate from Benford's value are highlighted.

The advantage of a drill-down option or further groupings & filters is obvious when looking at the many deviations that might be worth to investigate further, as not only the highlighted values but even more might be worth to investigate.

The chart in **Fig. 13**, shows that especially the first digit 2 was highly frequented in his claim amounts, much more than others and having a high deviation from Benford for this digit. This observation makes an analysis on the lowest level necessary, the actual records underneath. A drill down to the detail data for dataset 1 lists all the amounts claimed in **Fig. 13** by vendor #9206.

- Durtschi, C., W. Hillison and C. Pacini, 2004. The effective use of Benford's law to assist in detecting fraud in accounting data. *J. Forensic Account.*, 5: 17-34.
- Geyer, A. and J. Marti, 2012, Applying Benford's law to volcanology. *Geology*, 40: 327-330. DOI: 10.1130/G32787.1
- Hand, D., H. Mannila and P. Smyth, 2001. *Principles of Data Mining*. 1st Edn., MIT Press, Cambridge, ISBN-10: 026208290X, pp: 546.
- Hesketh, D., 2010. Weaknesses in the supply chain: Who packed the box. *World Customs J.*, 4: 3-20.
- Hoffman, W., 2008. Fighting supply-chain fraud. *Traffic World*.
- Hu, D., 2009. Exploratory study on computer forensics technology. Proceedings of the 2nd International Conference on Computer and Electrical Engineering, Dec. 28-30, IEEE Xplore Press, Dubai, pp: 608-611. DOI: 10.1109/ICCEE.2009.258
- Kroll Advisory, 2012. Global fraud report. Kroll Consulting, USA, Fall.
- Massa, D. and R. Valverde, 2014. A fraud detection system based on anomaly intrusion detection systems for e-commerce applications. *Comput. Inform. Sci.*, 7: 117-140. DOI: 10.5539/cis.v7n2p117
- Mayne, M., 2010. Elementary, my dear Watson? For IT Security Professionals, SC Magazine.
- Mishra, J.S., S. Panda and A.K. Mishra, 2013. A novel approach for credit card fraud detection targeting the Indian market. *Int. J. Comput. Sci.*, 10: 172-179.
- Nigrini, 1996. A taxpayer compliance application of Benford's Law. *J. Am. Taxat. Assoc.*, 18: 72-91
- Nigrini, M. J., Mittermaier, L. J. 1997. The use of Benford's law as an aid in analytical procedures. *Auditing*, 16: 52-67.
- Nigrini, M.J. 1999. 'I've Got Your Number'. *J. Accountancy*, 187: 79-83.
- Nigrini, M., 2012. *Benford's Law: Applications for Forensic Accounting, Auditing and Fraud Detection*. 1st Edn., John Wiley and Sons, Hoboken, New Jersey, ISBN-10: 1118152859, pp: 330.
- Nirkhi, S.M., R.V. Dharaskar and V.M. Thakre, 2012. Data mining: A prospective approach for digital forensics. *Int. J. Data Min. Knowl. Manage. Process*, 2: 41-48.
- Özera, G. and B. Babacanb, 2013. Benford's Law and Digital Analysis: Application on Turkish Banking Sector. *Bus. Econom. Res. J.*, 4: 29-41.
- Stephens, J. and R. Valverde, 2013. Security of e-procurement transactions in supply chain reengineering. *Comput. Inform. Sci.*, 6: 1-20. DOI: 10.5539/cis.v6n3p1
- Tödter, K.H., 2009. Benford's law as an indicator of fraud in economics. *German Econom. Rev.*, 10: 339-351. DOI: 10.1111/j.1468-0475.2009.00475.x