Benford's law and the detection of frauds in international trade

Andrea Cerioli[1], Lucio Barabesi[2], Andrea Cerasa[3], Mario Menegatti[1] and Domenico Perrotta[3]

[1]Department of Economics and Management, University of Parma, Parma, Italy
(e-mail: andrea.cerioli@unipr.it, mario.menegatti@unipr.it)

[2]Department of Economics and Statistics, University of Siena, Siena, Italy
(e-mail: lucio.barabesi@unisi.it)

[3]European Commission, Joint Research Centre, Ispra, Italy
(e-mail: andrea.cerasa@ec.europa.eu, domenico.perrotta@ec.europa.eu)

Benford's law for fraud detection. Foundations, methods and applications.
Stresa, July 10–12th, 2019

Most of this work is based on:

Barabesi L., Cerasa A., Cerioli A. and Perrotta D. (2018). Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud. *Journal of Business & Economic Statistics*, **36**, 346–358.

Cerioli A., Barabesi L., Cerasa A., Menegatti M. and Perrotta D. (2019). Newcomb-Benford law and the detection of frauds in international trade. *PNAS*, **116**, 106–115.

Still in progress ...

Benford's law (BL) for international trade

- **No need to introduce BL to this audience ...**
- **Limit Theorem for the significant-digit distribution**: if distributions are selected at random and random samples are taken from each of these distributions, the significant-digit frequencies of the combined sample converge to BL
- This limit Theorem has intuitive relationship with **the data generation process in international trade**: each trader t performs n_t (random) trades on m_t (random) products
- Relationship between BL and data generation international trade is reinforced by results from **Economic Theory** (see our PNAS paper)
- **Anti-fraud analysis**: deviations from BL (when expected) may be taken as possible instances of fraud $\Rightarrow$ **Fraudsters may be biased toward simpler distributions for digits** (Uniform, Dirac, etc.)

Typical **statistical tools** for detecting frauds in international trade data (AMT Project):

- **outlier detection**: frauds are uncommon with respect to the bulk of the transactions (but masking may occur)
- **robust regression**: estimation of the fair price for each product (price is the slope in the quantity-value scatter plot)
- **robust clustering**: in practice, there is heterogeneity in the quality of the same type of product and also in the conditions within the EU market

These methods identify **potential frauds as outlying transactions**.

The BL approach shifts the focus from individual transactions to individual traders: analyze all the transactions of each trader $\Rightarrow$ potential fraudsters stand out as outlying traders wrt BL

Our statistical goals

1. **Develop suitable methods for detecting anomalous deviations from BL (identify potential fraudsters)**: **goodness-of-fit tests with low rate of false detections and good power**

2. **Evaluate the properties of these tests in the framework of international trade data**: **simulation study and calibration of the tests when the conditions for BL are not met**

Statistical modeling of fraud I

T traders in the market. For $t = 1, \ldots, T$ and each $k \in \mathbb{Z}^+$:

$$\begin{aligned} \pi_{\mathbf{k}}^{(\mathbf{t})}(\mathbf{d_1}, \ldots, \mathbf{d_k}) &= P(D_1(X^{(t)}) = d_1, \ldots, D_k(X^{(t)}) = d_k) \\ &= (1 - \tau_t)\Psi_k^{(t)}(d_1, \ldots, d_k) + \tau_t \Upsilon_k^{(t)}(d_1, \ldots, d_k) \end{aligned}$$

where

- $\Psi_k^{(t)}(d_1, \ldots, d_k)$ is the probability of observing $\{D_1(X^{(t)}) = d_1, \ldots, D_k(X^{(t)}) = d_k\}$ in the **absence of fraud**
- $\Upsilon_k^{(t)}(d_1, \ldots, d_k)$ is the probability of the same event for a **manipulated transaction**
- $0 \le \tau_t \le 1$ is the **probability of fraud** for trader t.

Statistical modeling of fraud II

A tractable version of the contamination model assumes that $\pi_k^{(t)}(d_1, \ldots, d_k)$ depends on the trader only through

m_t: **number of products** traded by trader t
n_t: **number of transactions** made by trader t

Therefore

$$\pi_{\mathbf{k}}^{(\mathbf{t})}(\mathbf{d_1}, \ldots, \mathbf{d_k}) = (1 - \tau_t)\Psi_k^{(m_t, n_t)}(d_1, \ldots, d_k) + \tau_t \Upsilon_k^{(t)}(d_1, \ldots, d_k)$$

Under these models, trader t is a **potential fraudster** if the **null hypothesis**

$$H_0^{(t)} : \tau_t = 0$$

is rejected, in favor of the alternative

$$H_1^{(t)} : \tau_t > 0$$

based on n_t independent copies of $X^{(t)}$, say $X_1^{(t)}, \ldots, X_{n_t}^{(t)}$.

Statistical modeling of fraud III

In the coarser analysis of **aggregated accounting data** (traders, companies, taxpayers, etc.) the contamination model is

$$\pi_{\mathbf{k}}(\mathbf{d_1}, \ldots, \mathbf{d_k}) = (1-\tau)\Psi_k(d_1, \ldots, d_k) + \tau\Upsilon_k(d_1, \ldots, d_k)$$

and the sample is pooled across all the units: **deviation from BL does not allow to identify the potential fraudsters.**

In the analysis of **individual transactions** for a given product (standard customs approach), the contamination model for X (transaction value) is

$$F(X) = (1-\tau)F_0(X) + \tau G(X)$$

with $F_0(X)$ the distribution of "genuine" values: the sample $X_1, \ldots, X_n$ is pooled across all the traders and **no information about the "serial" behavior of each trader is included in the outlier detection process.**

Test statistics

- For simplicity, we mainly use the first digit D_1 and the **chi-squared statistic** (but other choices are possible!):

$$T_{\{1\}} = \sum_{d_1=1}^{9} \frac{(N_1(d_1) - n\pi_1(d_1))^2}{n\pi_1(d_1)}$$

has a χ^2_8 distribution in large samples. The χ^2_8 approximation may still be good in moderate and small samples: **no parameter to be estimated!**

- Test on **joint digit distributions** (e.g., $T_{\{1,2\}}, T_{\{1,2,3\}}, \ldots$). The **quality of asymptotic approximations rapidly deteriorates**: 90 cells with two digits; 900 cells with three digits; etc.

Our (first) proposal

- We start by keeping the number of false alarms to the prescribed level (e.g., 1%) in the simultaneous (i.e., multiple-digit) BL test: **reject in case of strong evidence against BL** $\Rightarrow$ **recall that limiting the proportion of False Positives is a crucial issue in routine analysis of trade data**.
- If the simultaneous BL hypothesis is rejected: we want to know **which digit is responsible for rejection**.
- Lower level tests are based on **conditional distributions**: **we control the Type-I error rate for all of them** $\Rightarrow$ small number of false signals also when testing conformance of individual digits to the BL (much smaller than using the marginal distributions of $D_1, D_2, \ldots$).

Computation of (conditional) *p*-values is based on an **efficient Monte Carlo procedure**: **our (conditional) tests are exact in finite samples** and do not rely on the asymptotic χ^2 approximation.

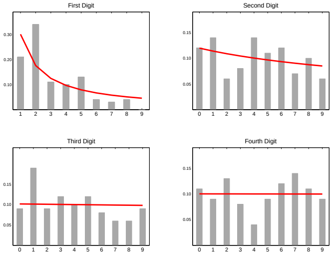
Anti-fraud analysis in international trade

- **Single Administrative Document** (SAD): information on the goods commodity code, the movement of the goods, the customs procedure and the traded quantities and values
- As in most anti-fraud applications we focus on the **statistical value** reported in each SAD ⇒ analyze the digits of statistical values
- Misdeclaring the statistical value likely implies fraud ⇒ **Undervaluation**: pay less duties or excises, evade import restrictions such as anti-dumping measures; **Overvaluation**: money laundering, higher export refunds or duty compensations, evade internal taxes

Analysis of a few interesting traders for illustrative purposes

First example – Trader A

- **Taken from SAD import records collected in 2011 by the Customs Office of MS1**
- $n = 100$ transactions, from 38 to 131,213 euros
- $m = 23$ traded products



$T_{\{1\}} = 28.71$: **testing the first digit of the import values for Trader A suggests non-conformance to BL** (**exact 0.99-quantile**: 20.44)

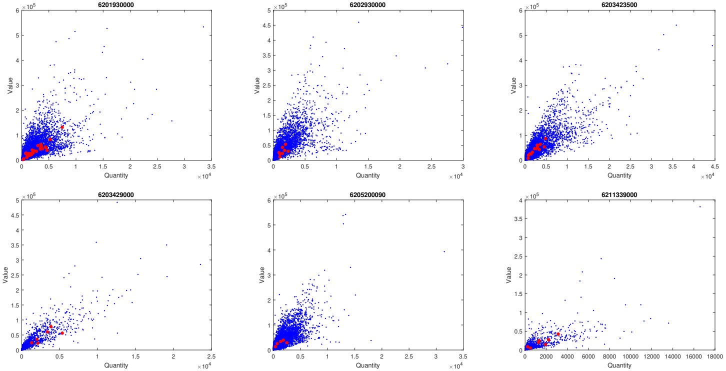
Trader A – Our multiple-stage analysis

$\{l_1, \ldots, l_m\}$	$T_{\{l_1,\ldots,l_m\}}$	Marg.	Cond.
{1,2,3,4}	8010.37	10422.0	10422.0
{1,2,3}	837.68	1083.7	1223.9
{1,2}	89.50	128.8	155.8
{1}	28.71	**20.4**	31.8
{2}	11.54	21.7	22.7
{3}	12.29	21.6	22.2
{4}	7.41	21.7	21.7

- Start by testing a **simultaneous 4-digit BL hypothesis** based on the exact **trader-specific** joint distribution of D_1, D_2, D_3, D_4
- **No simplification of the global BL null can be rejected**
- Rejection of $H_0^{\{1\}}$ based on the marginal distribution of $T_{\{1\}}$ provides weak evidence of data fabrication: **it might be an instance of false discovery**
- This conclusion is supported by further investigation of the transaction data for Trader A: **behavior of Trader A within the market**

Trader A – Further evidence

Quantity – value scatter plots for the six main products for Trader A (in red)

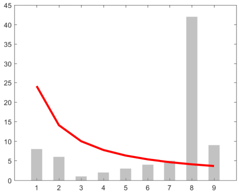


Most of transactions seem to be well in line with the market: **no evidence of "serial fraud"**

Second example – Trader B

- $n = 80$ transactions, from 86.75 to 9,346 euros
- $m = 8$ traded products
- **Taken from SAD import records collected in 2012–2015 by the Customs Office of MS2** (focus on traders that operate on a set of sensitive products) $\Rightarrow$ **Benchmark example: detected (and convicted) serial fraudster**

First digit distribution



$T_{\{1\}} = 388.82$

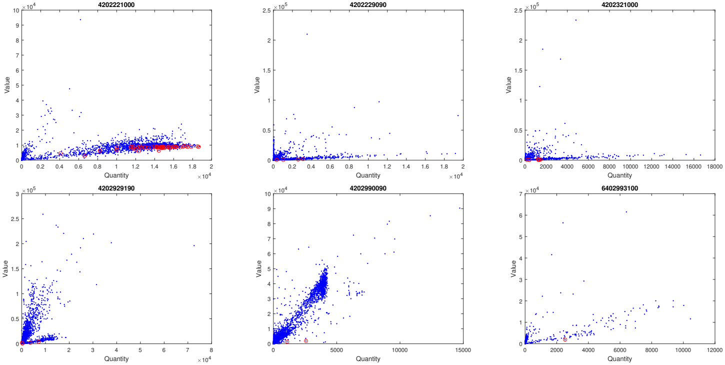
Trader B – Our analysis (with two stages)

$\{l_1, ..., l_m\}$	$T_{\{l_1,\ldots,l_m\}}$	Marg.	Cond.
{1,2}	787.57	**129.9**	**129.9**
{1}	388.82	**20.4**	**32.6**
{2}	14.79	21.7	28.9

- For simplicity: start with two digits
- The exact quantiles are obtained for the specific value of n for this trader
- **The statistical conclusion is straightforward!**
- We again look at the transaction data

Trader B – Further evidence

Quantity – value scatter plots for the six main products for Trader B (in red)



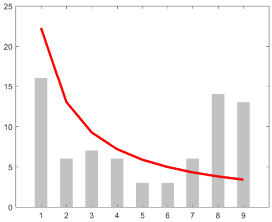
The scatter plots point to **systematic underpricing**

The evidence provided by BL analysis is even stronger than that given by the scatter plots: in the case of “serial” fraudsters we borrow strength from all transactions

Third example – Trader C

- $n = 74$ transactions, from 92.16 to 11,570 euros
- $m = 6$ traded products
- **As Trader B: data from the Customs Office of MS2 $\Rightarrow$ another benchmark example: detected (and convicted) serial fraudster**

First digit distribution



$T_{\{1\}} = 64.01$

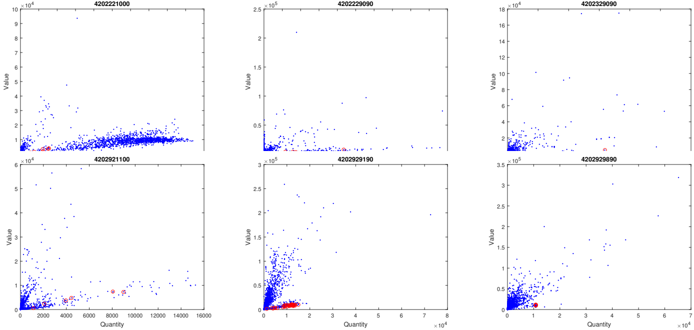
Trader C – Our analysis (with two stages)

$\{l_1, ..., l_m\}$	$T_{\{l_1,...,l_m\}}$	Marg.	Cond.
{1,2}	177.18	**130.4**	**130.4**
{1}	64.01	**20.5**	**32.6**
{2}	5.07	21.7	29.5

- For simplicity: start with two digits
- The exact quantiles are obtained for the specific value of n for this trader
- **Strong evidence against BL for the first digit, even after conditioning on rejection of the two-digit hypothesis** (substantial difference wrt Trader A)
- We again look at the transaction data

Trader C – Further evidence

Quantity – value scatter plots for the six products for Trader C (in red)



As for Trader B:

- the scatter plots point to systematic underpricing (especially for some products)
- the evidence provided by BL analysis is even stronger than that given by the scatter plots

Some Monte Carlo evidence with trade data

Quality of the BL approx. for "regular" transactions in real markets

Simulated trading behaviour by **picking price and quantity at random from Italian customs declarations in 2014**. Our "ideal" market: **refer to the talk by Andrea Cerasa for details**

- $\mathfrak{P}$ = set of 5,447 products with at least 50 transactions (corresponding to 6,265,198 trades, $\approx$ 99% of the total market)
- 10,000 simulated traders; each trader makes n transactions on m different products ($m \leq n$)
- Simulated transactions:
  - Select randomly m elements $p_1, \ldots, p_m$ from the set $\mathfrak{P}$ with probability proportional to the number of trades of each product
  - Select randomly m values $n_1, \ldots, n_m$ such that $n_j > 0 \quad \forall j$ and $\sum_1^m n_j = n$
  - From the transactions of each selected product p_j, extract randomly n_j exchanged quantities and n_j unitary prices. The element-by-element product of these vectors generates an n_j-vector of values
  - Iterate the previous step for $j = 1, \ldots, m$: n-vector of values free from manipulations (no fraudsters)
- many configurations with $n = 50, \ldots, 500$ and $m = 1, \ldots, n$

Conclusions about MC evidence – No fraud

We have computed **empirical test sizes** (nominal size: $\alpha = 0.01$) for:

- standard χ^2 test on first digit of simulated transaction values ($T_{\{1\}}$)
- exact Two-Stage test (TS)

Summary of results:

- **Good agreement between empirical and nominal sizes for the χ^2 test if m is close to n** ($m \geq 0.2n$): the Benford distribution for D_1 fits well to simulated "honest" transactions even when n is small
- The two-stage test is conservative (it adjusts for multiplicity of tests on digits): **the False Positives rate is $\ll 1\%$**
- **The Benford approximation worsens considerably when the ratio m/n decreases: large number of False Positives**
- **Effect of n for fixed m**: not enough variability in the traded products to ensure validity of BL $\Rightarrow$ **If m is small (and fixed), increasing n is not the solution**

Our (second) proposal: Corrected χ^2 tests

Compute (for the first digit)

$$T^*_{\{1\}} = \psi(m; n) T_{\{1\}}$$

where

$$\psi(m; n) = \frac{\chi^2_{8,1-\alpha}}{t_{\{1\},1-\alpha}(m; n)}$$

$\chi^2_{8,1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_8 and $t_{\{1\},1-\alpha}(m; n)$ is the $(1-\alpha)$-quantile of the distribution of $T_{\{1\}}$ for the given values of n and m

We estimate $\psi(m; n)$ by Monte Carlo simulation: new run with 10,000 independent traders for each m and n

Improved correction: $T^{**}_{\{1\}} \Rightarrow \psi(m; n)$ is estimated by MC simulation on the **same set of products** traded by the subject under investigation

With both $T^*_{\{1\}}$ and $T^{**}_{\{1\}}$ in our simulations we obtain good control of Type-I error rates even with $m = 1$!

Power under contamination

Simulated trading behaviour from Italian customs declarations (see before) with some **contaminated transactions** (frauds)

Different contamination (fraud) schemes:

1. Replace the first-two digits with random integers from the Uniform distribution (most difficult to detect)
2. Replace the first-two digits with the same pair of random numbers chosen uniformly among $51, \ldots, 59$: accumulation on $D_1(X) = 5$ (see Trader B)
3. Replace the first-two digits with the same pair of random numbers chosen uniformly among $1, \ldots, 99$: accumulation on randomly chosen digit
4. Model-based contamination: digits come from the Generalized Benford distribution

Different contamination rates (for each fraudster): $0.2 \times n, 0.5 \times n, 0.8 \times n$

Different proportions of fraudsters (in the whole market): 5%, 10%, 20%

MC evidence – Power vs False Positive Rate I

Uniform contamination. Contamination rate: $0.8 * n$; 5% of fraudsters

Least favourable contamination – Realistic scenario?

Nominal test sizes: $\alpha = 0.01$. First entry: **Power**; Second entry: **False Positive Rate**

		n		
		50	100	200
$m = 0.2n$	$T_{\{1\}}$	0.586 0.360	0.926 0.265	1.000 0.199
	$T^*_{\{1\}}$	0.514 0.292	0.894 0.189	0.998 0.121
	TS	0.110 **0.052**	0.526 **0.037**	0.950 **0.014**
$m = n$	$T_{\{1\}}$	0.586 0.302	0.938 0.184	1.000 0.151
	$T^*_{\{1\}}$	0.574 0.295	0.932 0.177	1.000 0.165
	TS	0.154 **0.013**	0.574 **0.003**	0.964 **0.002**

- **Two-stage procedure: reasonable power properties and very low False Positive rates** (lower than one-digit $T_{\{1\}}BH$)
- $T_{\{1\}}$ and $T^*_{\{1\}}$ have similar (high) power, but FPR is still relatively large
- Power increases with n and with contamination rate (results available)

MC evidence – Power vs False Positive Rate II

Alternative contamination schemes:

- **Power = 100% for all tests**
- The FPR advantage of our Two-Stage procedure is even larger when $m \geq 0.2n$: **FPR ≤ 0.01 in most settings**

When $m \ll n$:

- **The power of $T^{**}_{\{1\}}$ is larger than the power of $T^{*}_{\{1\}}$ and is very close to that of the uncorrected and (very) liberal chi-squared test**
- **In our simulations, Test $T^{**}_{\{1\}}$ has the correct size (0.01) in the absence of fraud and power close to 1 under most contamination schemes!**
- Possible (but not crucial) disadvantage of $T^{**}_{\{1\}}$: **computationally intensive** $\Rightarrow$ specific correction factor for each m-ple of products

Conclusions – What have we learned?

- **If strict control of the FPR is a major issue: Two-stage or Multi-stage procedure**

- **If control of the FPR is relaxed: Adjusted Chi-squared tests**

Conclusions – What's next?

Methodology:

- Corrections for the Multi-stage test when $m \ll n$
- Refinement of the Monte Carlo correction factor to incorporate **dependence** in each trader behavior: “autocorrelation”, market constraints, recurring trade operation, etc.

International trade:

- More extensive benchmark: **talk by Andrea Cerasa – Robert Stadler**
- **Combine the signals** provided by alternative tests and by other anti-fraud statistical tools: **talk by Marco Riani and FSDA Toolbox**

Software:

- **Benford module in SAS developed by Francesca Torti: talk joint with Caroline Gasparro**
- **WebAriadne**: Web-based service for routine use by EU Customs and OLAF, developed by Emmanuele Sordini, Massimiliano Gusmini and the JRC technical staff