Marzena Farbaniec [*]
Tadeusz Grabiński [**]
Bartłomiej Zabłocki [***]
Wacław Zając [****]

Wykorzystanie prawa Benforda w analizie poprawności danych finansowych na przykładzie informacji o obrocie towarowym

Streszczenie: *Artykuł przestawia przykład analizy poprawności danych finansowych przy zastosowaniu prawa Benforda. Jego celem jest zaproponowanie rekurencyjnej metody umożliwiającej stopniowe zawężanie wielkości zbioru, w którym potencjalnie można podejrzewać występowanie nieprawidłowości. W artykule wykorzystano dane dotyczące obrotu towarów w Polsce w latach 2009 – 2010.*

Słowa kluczowe: prawo Benforda, INTRASTAT, GUS, rozkład cyfr znaczących, analiza danych, rekurencja, audyt finansowy

1. Geneza i istota prawa Benforda

Słynny astronom Simon Newcomb, dyrektor *American Nautical Almanac Office* w Waszyngtonie, w roku 1881 opublikował dwustronicowy artykuł[1], w którym opisał ciekawe zjawisko jakie zaobserwował w bibliotece przeglądając tablice logarytmiczne. Spoglądając na bok księgi dostrzegł, że brzegi pierwszych kartek były najbardziej zabrudzone i stopniowo stawały się coraz czystsze. Z tej obserwacji wywnioskował, że użytkownicy tablic częściej szukali logarytmów liczb zaczynających się na 1 niż na 2, częściej na 2 niż na 3 i tak aż do 9. S. Newcomb podał wzór na prawdopodobieństwo wystąpienia niezerowej cyfry *d* na pierwszej pozycji liczby kilkucyfrowej:

$$P_i = \frac{\log\ i+1\ -\log(i)}{\log\ 10\ -\log(1)} = \log(1+\frac{1}{i})\ ,\ i = 1,2,\dots 9$$

(1)

Zgodne ze wzorem (1) prawdopodobieństwo wystąpienia cyfry 1 wynosi 30,1%, podczas gdy cyfry 9 jedynie 4,6% (por. rys. 1).

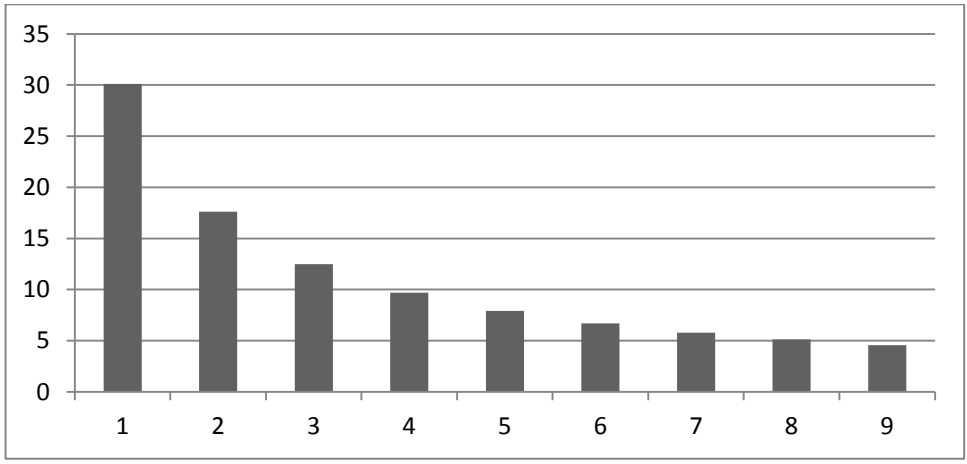
[*] Absolwentka Wydziału Zarządzania Uniwersytetu Ekonomicznego w Krakowie - kierunek Informatyka i Ekonometria, specjalizacja – Modelowanie i prognozowanie procesów gospodarczych. Obecnie pracuje na stanowisku analityka finansowego w firmie Motorola Solutions Polska. Jej zainteresowania naukowe koncentrują się wokół praktycznego wykorzystania rozkładu Benforda.

[**] prof. dr hab., pracownik Katedry Finansów Uniwersytetu Ekonomicznego w Krakowie.

[***] ukończył Wydział Zarządzania Uniwersytetu Ekonomicznego w Krakowie. Obecnie pracuje jako konsultant baz danych w firmie informatycznej we Wrocławiu. Jego zainteresowania naukowe koncentrują się wokół wykorzystania rozkładu Benforda w praktyce, a także psychologii stosunków międzyludzkich, ich relacji i zachowań niewerbalnych.

[****] Absolwent Wydziału Zarządzania Uniwersytetu Ekonomicznego w Krakowie - kierunek Informatyka i Ekonometria, specjalizacja - Zarządzanie informacjami. Student studiów doktoranckich UEK na Wydziale Zarządzania. Administrator bankowych systemów transakcyjnych w firmie ABB Polska. Jego zainteresowania skupiają się wokół praktycznych zastosowań prawa Benforda.

[1] Newcomb S., *Note on the Frequency of Use of the Different Digits in Natural Numbers,* American Journal of Mathematics, Vol. 4, No. 1. (1881), pp. 39-40



Rysunek 1.Rozkład Benforda dla pierwszej znaczącej cyfry w systemie dziesiętnym.

Źródło: opracowanie własne

Artykuł S. Newcomba pozostał niezauważony i przeszedł bez większego echa. Pół wieku później amerykański fizyk, pracownik General Electric – Frank Benford, prawdopodobnie będąc nieświadomym publikacji S. Newcomba, dokonał identycznej obserwacji. Postanowił zgromadzić różnego typu dane pochodzące z tak wielu źródeł jak tylko było to możliwe. Swoje badania F. Benford oparł na 20229 obserwacjach[2] pochodzących ze źródeł dotyczących m.in.: długości rzek, statystyk MLB (Amerykańskiej Ligi Baseballowej), mas atomowych pierwiastków, liczb pojawiających się w artykułach Reader's Digest, danych demograficznych etc. Zauważył również, że niektóre ciągi matematyczne $1/n$ czy $\overline{n}(N \in \mathbb{N})$ wykazują podobną tendencję do podążania za prawem pierwszej cyfry znaczącej[3]. W przeciwieństwie do artykułu S. Newcomb'a praca F. Benforda zdołała zwrócić uwagę czytelników. Tekst S. Newcomb'a został zapomniany a prawo rozkładu pierwszych cyfr zostało nazwane prawem Benforda.

2. Zastosowanie prawa Benforda

Prawo Benforda stosowano lub proponuje się wykorzystywać w różnych obszarach.
W zakresie nauk ekonomicznych do:

- wykrywania fałszywych danych lub niezamierzonych błędów w księgowości (*Digital Analysis - DA),*
- wykrywania oszustw podatkowych[4],
- analizy danych giełdowych (ceny i obrót papierami wartościowymi)[5],
- analizy cen towarów wylicytowanych na aukcjach internetowych[6],

[2]Benford F., *The law of anomalous numbers*, Proceedings of the American Philosophical Society, Vol. 78, No. 4, 1938.
[3]Ryder P., *Multiple origins of the Newcomb - Benford law: rational numbers, exponential growth and random fragmentation*,Staats- und Universitätsbibliothek Bremen, Germany, 2009.
[4] Nigrini M., *A taxpayer compliance application of Bedford,s Law*, Journal of the American Taxation Associates, 18/1996, p. 72-92; M.Nigrini*"Digital Analysis Using Benford's Law",* Global Audit Publications, 2000; Nigrini M., *Digital Analysis Using Benford's Law: Tests and Statistics for Auditors,*The EDP Audit, Control, and Security Newsletter, EDPACS, 28/2001; M. Nigrini,*„I've Got Your Number",* Journal of Accountancy, 187/ 1999, M.Nigrini,*„Adding Value with Digital Analysis",* The Internal Auditor, 56/1999, p. 21–23.
[5] Ley E., *On a pecular distribution of the U.S. stock indicies digits,* American Statistician, 1996.

- analizy długości czasu w jakim klienci czują się związani z obsługującą ich firmą czy bankiem, co jest istotne w procesie projektowania systemów CRM *(Customer Relation Management),*
- oceny poprawności szacunków odszkodowań w firmach ubezpieczeniowych,
- oceny rzetelności wysokości grzywien i kar finansowych orzekanych w procesach sądowych.
- W zakresie ***nauk ilościowych*** do:
- testowania poprawności modeli ekonometrycznych, np. w procesie prognozowania (dane teoretyczne powinny spełniać prawo Benforda w takim samym stopniu jak dane empiryczne na podstawie których model był szacowany),
- optymalizacji obliczeń w rozwiązywaniu problemu transportowego komiwojażera (odległości Euklidesa miedzy różnymi miejscowościami spełniają prawo Benforda),
- weryfikacji poprawności danych statystycznych, w szczególności pochodzących z powszechnych spisów demograficznych.
- W zakresie ***nauk informatyczno-technicznych*** do:
- projektowania architektury pamięci masowych w komputerach[7],
- odróżniania rzeczywistych fotografii od grafiki generowanej przez programy komputerowe (fotografie rzeczywiste powinny mieć wartości pikseli zgodne z rozkładem Benforda).
- analizy rozmiarów plików transferowanych w Internecie oraz czasu ich transferu,
- symulacyjnego badania efektywności *(benchmarking)* algorytmów numerycznych, które dotychczas wykonywano na zbiorach generowanych losowo lub zakładano ich typowe rozkłady (normalny, równomierny)
- W zakresie ***pozostałych nauk*** do oceny:
- skuteczności klinicznej leków,
- prawdziwości danych o emisji toksycznych gazów[8],
- wielkości dotacji dla partii politycznych w wyborach parlamentarnych w kontekście ich zgodności z obowiązującymi przepisami prawnymi[9],
- liczby odnośników na poszczególnych stronach artykułów i książek,
- liczby rannych i zabitych w wypadkach drogowych, kolejowych, samolotowych,
- liczby powtórnych wizyt w sklepach i punktach usługowych,
- liczby zakupionych produktów,
- liczby kopii oprogramowania sprzedanego różnym klientom,
- liczby zdobytych punktów w baseballu przez jednego zawodnika.

3. Charakterystyka proponowanej metody analizy poprawności zbiorów danych

W dostępnej literaturze można znaleźć wiele przykładów zastosowania prawa Benforda do oceny rzetelności zbiorów danych, w tym również finansowych zbiorów danych. Z reguły analizy te mają charakter całościowy i sprowadzały się do próby udzielenia odpowiedzi czy ***cały*** badany zbiór spełnia czy też nie spełnia reguł związanych z prawem Benforda. Brak jest natomiast metody, w której kierując się wskazaniami mierników charakteryzujących stopień zgodności z rozkładem Brenforda, ***stopniowo*** definiować mniejsze podzbiory i analizować rozkłady cyfr znaczących w tych

[6] .Giles D.E "*Benford's Law and naturalyoccuringproces in certain eBay auctions",* Econometric Working Papers EWP 0505, Univ. of Victoria, 2005.
[7] Hill T.P. ,*The first digital phenomenon,* American Scientist, 86/1998.
[8] de Marchi S., J. Hamilton, *Assessing the Accuracy of Self-Reported Data: An Evaluation of the Toxics Release Inventory*, Journal of Risk and Uncertainty, 32/2006.
[9] Tam Cho W.K, Gaines B.J., *Breaking the (Bedford) Law: Statistical fraud detection campaign finances*, The American Statistician, 61/2007, p.218-223

podzbiorach. Miernik dopasowania stanowi kryterium zgodności rozkładu empirycznego z rozkładem Benforda. Mogą nimi być powszechnie stosowane testy statystyczne. Najczęściej wykorzystywanym jest test chi-kwadrat w którym wyznacza się wartość parametru χ^2. Innym testem zgodności rozkładów jest test Kołmogorowa – Smirnowa, niezależny od wielkości zbioru *n*. W wersji oryginalnej tego testu wyznacza się statystykę, przedstawioną wzorem dla KS1. Rzadziej stosuje się test zgodności Kołomogorowa ze statystyką daną wzorem KS2. Wersją zmodyfikowaną testu, zaproponowaną przez Kuipera, jest test KS3. Uwzględnia on fakt cykliczności analizowanych rozkładów. Im bardziej zgodne są porównywane rozkłady częstości, tym mniejsze wartości przyjmują powyższe mierniki.

W zależności od charakteru zmian stopnia zgodności rozkładu cyfr znaczących z prawem Benforda w kolejnych iteracjach analizy możliwe są różne sytuacje:

a. Zmiany mają charakter zróżnicowany, nieregularny,
b. Stopień zgodności jest coraz to większy,
c. Stopień zgodności jest coraz to mniejszy.

Ponadto w procedurze oceny poprawności danych należy uwzględnić (poza kierunkiem zmian) także i poziom mierników zgodności. W tym kontekście można wyróżnić następujące sytuacje:

1. Mierniki zgodności wskazują we wszystkich iteracjach na brak zgodności
2. Mierniki zgodności wskazują we wszystkich iteracjach na zgodność rozkładów cyfr znaczących z prawem Benforda
3. Przy rosnących stopniach zgodności jej poziom zmienia swój charakter z oceny negatywnej (brak zgodności) na pozytywną (rozkład cyfr zgodny z rozkładem Benforda)
4. Przy malejących stopniach zgodności jej poziom zmienia swój charakter z oceny pozytywnej na negatywną.

W tab. 1 zebrano zasady wnioskowania w wyróżnionych 12 sytuacjach. Dwa przypadki oznaczone symbolem [---] są niemożliwe z uwagi na wykluczający się charakter kryteriów klasyfikacji. W dwóch przypadkach zaleca się powstrzymanie od formułowania opinii. W lewym górnym narożniku tabeli zebrane są sytuacje, w których rzetelność zbioru danych należy ocenić pozytywnie, natomiast w prawym dolnym narożniku – negatywne. Oceny pozytywne i negatywne można formułować z różnym stopniem przekonania o ich adekwatności z faktycznym poziomem rzetelności analizowanego zbioru danych.

Tabela 1. Ocena poprawności zbioru danych w zależności od kierunku zmian i poziomu mierników zgodności rozkładu cyfr znaczących z prawem Benforda

Poziom i kierunek zmian mierników zgodności rozkładu cyfr znaczących z prawem Benforda		**Kierunek zmian mierników zgodności**		
		Rosnący	**Zróżnicowany**	**Malejący**
Poziom mierników zgodności	**Wysoka zgodność w każdej iteracji**	Zdecydowanie pozytywna	Pozytywna	Raczej pozytywna
	W kolejnych iteracjach zmiana z braku zgodności na wysoką	Raczej pozytywna	Brak opinii	[--]
	W kolejnych iteracjach zmiana z wysokiej zgodności na jej brak	[--]	Brak opinii	Raczej negatywna
	Brak zgodności w każdej iteracji	Raczej negatywna	Negatywna	Zdecydowanie negatywna

Źródło: opracowanie własne

Odrębny problem uwzględniony w prezentowanych badaniach to kwestia wyboru najbardziej diagnostycznych mierników zgodności rozkładu cyfr znaczących z prawem Benforda. W literaturze można spotkać wiele różnych propozycji w tym zakresie. Jednakże mogą się zdarzyć sytuacje w których poszczególne mierniki zgodności prowadzą do rozbieżnych wniosków. W zaproponowanej w pracy procedurze wykorzystano w tym celu metody taksonometryczne, a konkretnie - diagraficzną metodę Czekanowskiego.

Kolejne etapy w omawianej metodzie można ująć w następujące punkty.

I. Wstępna obróbka zbioru danych

- Pogrupowanie danych wg ich zakresu merytorycznego
- Weryfikacja zbiorów danych – usunięcie duplikatów, wartości zerowych, wartości stałych (np. koszty stałe, występujące wielokrotnie w regularnych odstępach czasu)

II. Ocena jakości (stopnia poprawności) zbiorów

- Wyznaczenie mierników zgodności dla wyjściowych zborów danych
- Klasyfikacja mierników zgodności z wykorzystaniem metody Czekanowskiego
- Wybór najbardziej diagnostycznych mierników zgodności
- Ustalenie zbioru danych o najmniejszym stopniu zgodności rozkładu cyfr znaczących z prawem Benforda

III. Iteracyjna metoda poszukiwania najmniej poprawnego podzbioru danych

- Wyznaczenie najwyższej bezwzględnej wartości wybranego miernika zgodności (test Z) i odpowiadającej mu pierwszej cyfry znaczącej F1
- Wyznaczenie najwyższej bezwzględnej wartości testu Z i odpowiadających mu dwóch pierwszych cyfr znacznych F12 przy założeniu, że analizie podlegają liczby zaczynające się cyfrą ustaloną w poprzednim kroku
- Wyznaczenie najwyższej bezwzględnej wartości testu Z i odpowiadających mu trzech pierwszych cyfr znaczących przy założeniu, że analizie podlegają liczby zaczynające się od dwóch pierwszych cyfr ustalonych w poprzednich dwóch krokach algorytmu
- Powtarzanie wyżej omówionych iteracji dla kolejnych cyfr znaczących[10]

IV. Ustalenie najbardziej podejrzanego podzbioru danych w analizowanym zbiorze

- Podzbiór ten jest określony przez pierwsze cyfry znaczące wskazane w kolejnych krokach algorytmu iteracyjnego

Omówioną metodę analizy wykorzystano do analizy poprawności zbioru danych finansowych odnoszących się do wielkości obrotów towarowych w Polsce w latach 2009-2010, pochodzących z systemu INTRASTAT.

4. Elektroniczny system zgłoszeń INTRASTAT

Jedną z konsekwencji utworzenia Unii Europejskiej było zniesienie granic pomiędzy państwami członkowskimi tworzącymi jednolity rynek. Fakt ten spowodował, że statystyka Unii Europejskiej została pozbawiona źródła danych o wymianie towarowej z innymi krajami UE, ponieważ zaprzestano odpraw celnych, a zatem zniknął dokument celny SAD (*Single Administrative Document*). Konieczne było utworzenie nowego, wspólnego systemu statystyki obrotu towarowego pomiędzy państwami członkowskimi Unii Europejskiej. Obecnie w państwach członkowskich Unii Europejskiej dane o handlu zagranicznym pochodzą z dwóch równolegle funkcjonujących systemów statystyki:

- INTRASTAT - transakcje pomiędzy państwami UE nie wymagające zgłoszenia celnego,
- EKSTRASTAT - obrót towarowy państw członkowskich UE z krajami trzecimi[11]

[10] W przypadku dysponowania zbiorem niewielkich liczb (3-4 cyfry) można rozważyć zamianę ich podstawy i przejść z 10-tnego-systemu zapisu na system np. 8-kowy lub 6-kowy co pozwala zwiększyć liczbę cyfr znaczących.

Wprowadzenie w Polsce systemu INTRASTAT zaowocowało nałożeniem na podmioty prowadzące obrót towarowy z innymi państwami członkowskimi Unii Europejskiej obowiązku przekazywania informacji o zrealizowanych przez nich obrotach, tj. o dokonanym przywozie towarów z terytorium innych państw UE na terytorium Polski i wywozie towarów z terytorium Polski na terytorium innych państw należących do Wspólnoty. Informacje o handlu towarami pomiędzy Polską a pozostałymi krajami Wspólnoty są zbierane przez służby celne bezpośrednio od zobowiązanych podmiotów lub przedstawicieli reprezentujących te podmioty[12].

Tabela 2. Wartości progów statystycznych dla podmiotów realizujących obroty z krajami Unii Europejskiej.

Rok	Kierunek obrotu	Próg podstawowy (zł)	Próg szczegółowy (zł)
2009 - 2010	dla przywozu	1 000 000	33 000 000
	dla wywozu	1 000 000	60 000 000

Źródło - www.stat.gov.pl

Podmioty zobowiązane do przekazywania danych dotyczących ich obrotów towarowych z krajami Unii Europejskiej (por. tab. 2) są zobligowane do dokonywania zgłoszeń na DEKLARACJI INTRASTAT – PRZYWÓZ i DEKLARACJI INTRASTAT – WYWÓZ. Jest to zbiorcza informacja o dokonanych w ciągu danego okresu sprawozdawczego przywozach i wywozach towarów. Okresem sprawozdawczym jest miesiąc kalendarzowy, w trakcie którego dokonano wewnątrzwspólnotowej wymiany towarowej (wywozu lub przywozu towarów). Istnieje możliwość przekazywania informacji obejmującej okres krótszy niż miesiąc w formie deklaracji częściowej, jednak te częściowe informacje muszą łącznie obejmować cały miesięczny okres sprawozdawczy.

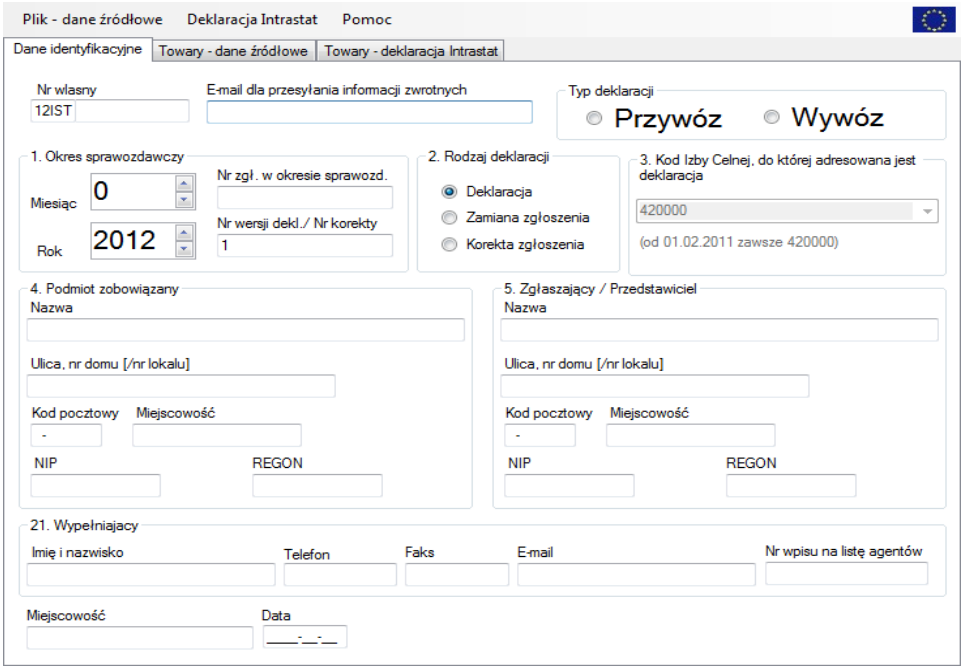
Deklaracje INTRASTAT można składać w formie pisemnej na formularzach lub elektronicznej (w ramach Systemu Obsługi Deklaracji CELINA) - CD-ROM, dyskietka i inne nośniki względnie pliki w formacie XML. Deklaracja INTRASTAT zawiera następujące pola (por. tab. 3 oraz rys. 2).

Tabela 3. Wykaz pól w deklaracji INTRASAT

Pole	11 - Kod kraju wysyłki –dla przywozu albo Kod kraju przeznaczenia – dla wywozu
1 - Okres sprawozdawczy	12 - Kod warunków dostawy
2 - Rodzaj deklaracji	13 - Kod rodzaju transakcji
3 - Kod izby celnej, do której adresowana jest deklaracja INTRASTAT	14 - Kod towaru
4 - Odbiorca - w przypadku przywozu albo Nadawca - w przypadku wywozu	15 - Kod rodzaju transportu
5 - Przedstawiciel - w przypadku zgłoszenia INTRASTAT przez przedstawiciela	16 - Kod kraju pochodzenia - w przypadku zgłoszenia INTRASTAT w przywozie
6 - Łączna wartość fakturowa w PLN	17 - Masa netto (w kg)
7 - Łączna wartość statystyczna w PLN	18 - Ilość w uzupełniającej jednostce miary
8 - Łączna liczba pozycji	19 - Wartość fakturowa w PLN
9 - Numer pozycji	20 - Wartość statystyczna w PLN
10 - Opis towaru	21 – Wypełniający

Źródło: opracowanie własne

[11]Główny Urząd Statystyczny - http://www.stat.gov.pl/gus/intrastat_PLK_HTML.htm, [21.10.2012]
[12]http://www.mf.gov.pl/_files_/sluzba_celna/intrastat/instrukcja_intrastat_v1_1_2011-09-06.pdf, [21.10.2012]



Rysunek 2. Formularz deklaracji INTRASTAT.

Źródło: http://www.mf.gov.pl/index.php?const=2&dzial=416&wysw=4

5. Wstępne wyniki analizy

Analizowane dane dotyczą okresów sprawozdawczych na przestrzeni dwóch lat 2009 oraz 2010 i pochodzą ze systemów INTRASTAT oraz EXTRASTAT. Ujęte są w przekroju poszczególnych podmiotów gospodarczych względnie są zagregowane do poziomu gmin. Informacje zostały ujęte w 8 zbiorach (por. tab. 4).

Tabela 4. Zestawienie zbiorów danych poddanych analizie.

Rok	INTRASTAT		EXTRASTAT	
	Gminy	Podmioty	Gminy	Podmioty
2009	IG09	IP09	EG09	EP09
2010	IG10	IP10	EG10	EP10

Źródło : opracowanie własne

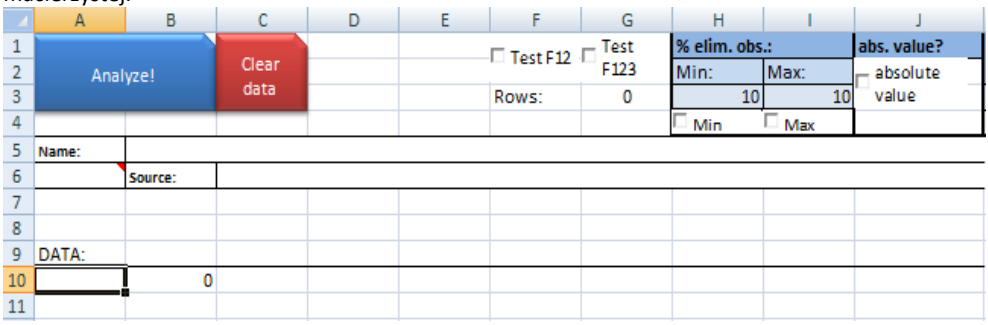
Każdy z tych zbiorów został poddany analizie specjalnie zaprojektowanym narzędziem opartym o arkusz kalkulacyjny MS Excel – Benford Analyzer. Składa się ono z dwóch zakładek:

- wejściowej – wprowadzane są surowe dane oraz parametry określające zakres analizy (rys. 3)
- wynikowej – podawany jest rezultat analizy oraz tabele pomocnicze (rys. 4)

Rezultaty analizy zgodności rozkładów cyfr znaczących dla wszystkich ośmiu zbiorów danych z prawem Benforda ujęto w tabeli, której fragment przytoczono w tab. 5. Wyznaczone mierniki dopasowania zostały poddane analizie taksonometrycznej mającej na celu ich podział na grupy mierników dających zbliżone do siebie wyniki, a następnie wyborze z każdej grupy miernika najlepiej reprezentującego poszczególne grupy mierników. W analizie wykorzystano w tym celu diagraficzną metodę Czekanowskiego (por. rys. 5).

Klasyfikacji poddano 16 parametrów[13] uzyskując ich podział na 4 skupiska przedstawione w tab. 6.

Reprezentantów poszczególnych grup wybrano na podstawie wskaźnika zdefiniowanego jako iloraz średniej arytmetycznej z mierników podobieństwa danego parametru z parametrami ***spoza*** grupy oraz średniej arytmetycznej z mierników podobieństwa danego parametru z parametrami ***wewnątrz*** grupy, w której jest ten parametr zlokalizowany. Wskaźnik ten preferuje więc parametry, które są najmniej podobne (dają najmniej zgodne wyniki ocen) do parametrów z innych grup a równocześnie najbardziej podobne do parametrów z grupy macierzystej.

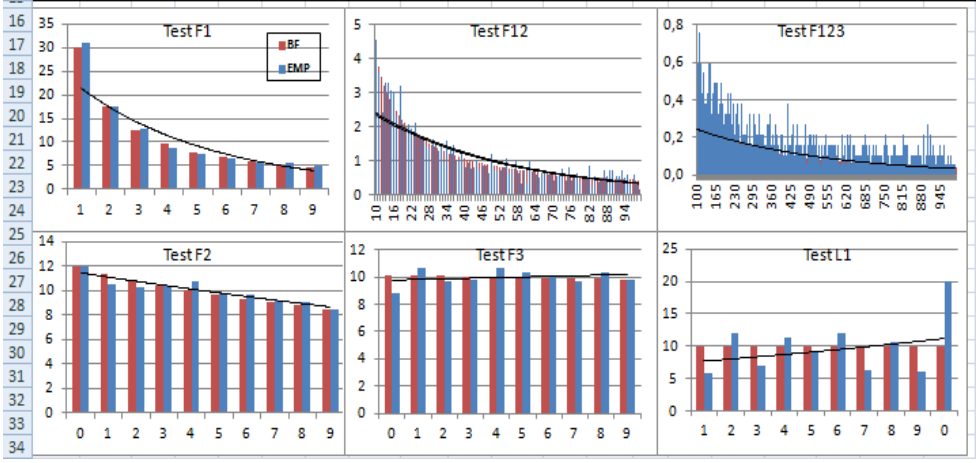


Rysunek 3. Zakładka wejściowa arkusza kalkulacyjnego.

Źródło. opracowanie własne

[13] Omówienie tych parametrów znaleźć można m.in. w pracach: Farbaniec M., Grabiński T., Zabłocki B., Zając W., *Wykorzystanie praw rozkładu cyfr do oceny wiarygodności danych finansowo-księgowych na wybranych przykładach,* X Międzynarodowy Kongres Kontroli Wewnętrznej, Audytu Wewnętrznego, Antykorupcji i Zwalczania Oszustw, Krakowska Akademia im. A.F. Modrzewskiego, Kraków 2011
Farbaniec M., Grabiński T., Zabłocki B., Zając W., *Wyniki wyborów powszechnych w Polsce w latach 2000-2010 w świetle analizy rozkładów cyfr* znaczących, rozdz. XI w monografii W poszukiwaniu skutecznych narzędzi i modeli analizy zjawisk społeczno-gospodarczych ,Wyższa Szkoła Przedsiębiorczości i Marketingu w Chrzanowie, Centrum Szkolenia i Organizacji Systemów Jakości PK, Chrzanów 2012, str. 161-179

Name	0													α=0,05	1,36	1,36	1,75	-	min [-] %
0	Source:	0,00												α=0,01	1,63	1,63	2,00	-	max [+] %
	n	M1	M2	M3	M4	M5	r	p (r)	chi	chi-0,05	p(chi)	k	z - 0,05	z - 0,01	KS1	KS2	KS3	stat.	param.
F1	1841	5,4	0,18	0,54	4,0	76,4	0,998	0,000	5,1	15,5	0,747	9	0	0	0,37	0,53	0,63	max	——
F12	1841	18,8	0,02	0,23	17,1	283,1	0,961	0,000	77,8	112	0,796	90	4	0	0,39	0,56	0,67	min	9 970,0
F123	1841	62,5	0,00	0,08	55,1	951,9	0,734	0,000	851,0	969,9	0,872	900	37	7	0,42	0,59	0,75	avg:	[illegible]
F2	1841	3,2	0,14	0,44	4,3	59,8	0,918	0,000	3,3	16,9	0,950	10	0	0	0,47	0,66	0,49	v-factor	2 229,9
F3	1841	4,5	0,18	0,57	5,7	83,3	-	-	6,0	16,9	0,741	10	0	0	0,43	0,61	0,44	kurt.	1 720,2
L1	1841	32,0	1,30	4,10	41,0	589,0	-	-	308,9	16,9	0,000	10	7	6	3,01	4,26	3,01	skew.	40,9



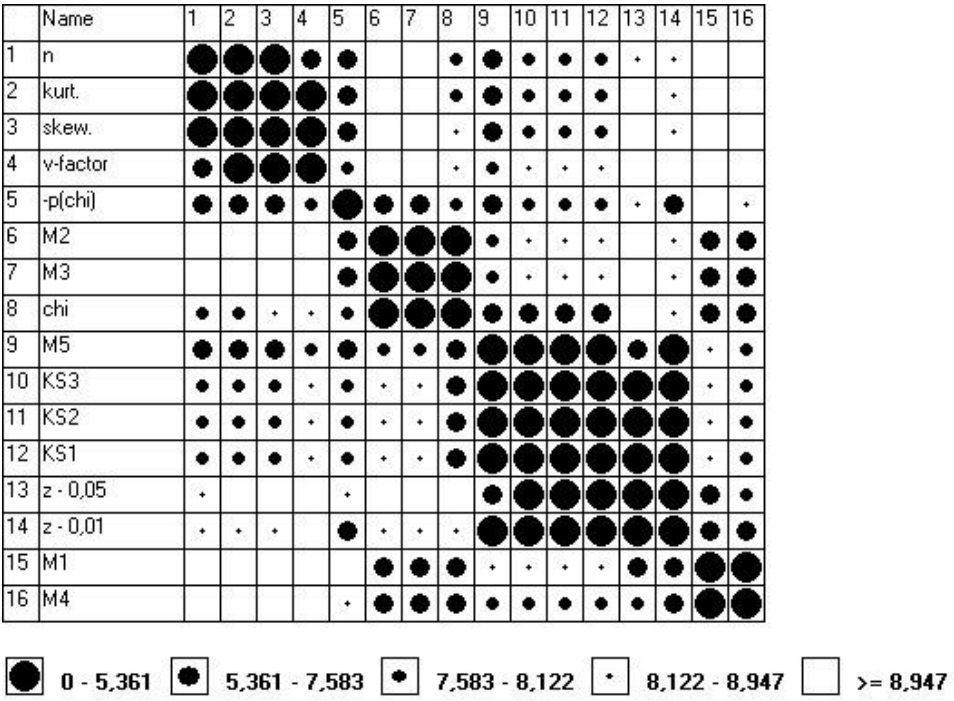
Rysunek 4. Zakładka wyjściowa – rezultaty obliczeń

Źródło: opracowanie własne

Tabela 5. Wartości mierników dopasowania dla analizowanych zbiorów danych

		n	M1	M2	M3	M4	M5	chi	p(chi)	z - 0,05	z - 0,01	KS1	KS2	KS3	v-factor	kurt.	skew.
EG09	F1	2173	6,8	0,23	0,69	5,1	110,3	10,9	0,209	1	0	0,56	0,79	0,74	1 371,4	1 642,8	38,5
	F12	2173	16,9	0,02	0,20	14,7	307,5	73,8	0,878	2	0	0,60	0,85	0,82	1 371,4	1 642,8	38,5
	F123	2173	58,8	0,00	0,07	52,2	1060,0	910,8	0,385	37	17	0,65	0,92	0,83	1 371,4	1 642,8	38,5
	F2	2173	6,2	0,21	0,67	6,7	129,9	10,2	0,332	0	0	0,36	0,51	0,58	1 371,4	1 642,8	38,5
	F3	2173	4,0	0,16	0,50	5,0	86,3	5,4	0,801	0	0	0,44	0,62	0,44	1 371,4	1 642,8	38,5
	L1	2173	29,0	1,22	3,86	38,6	629,6	324,4	0,000	7	6	3,23	4,56	3,23	1 371,4	1 642,8	38,5
IG09	F1	1841	5,4	0,18	0,54	4,0	76,4	5,1	0,747	0	0	0,37	0,53	0,63	2 229,9	1 720,2	40,9
	F12	1841	18,8	0,02	0,23	17,1	283,1	77,8	0,796	4	0	0,39	0,56	0,67	2 229,9	1 720,2	40,9
	F123	1841	62,5	0,00	0,08	55,1	1509,0	851,0	0,872	37	7	9,58	13,55	9,58	2 229,9	1 720,2	40,9
	F2	1841	3,2	0,14	0,44	4,3	59,8	3,3	0,950	0	0	0,47	0,66	0,49	2 229,9	1 720,2	40,9
	F3	1841	4,5	0,18	0,57	5,7	83,3	6,0	0,741	0	0	0,43	0,61	0,44	2 229,9	1 720,2	40,9
	L1	1841	32,0	1,30	4,10	41,0	589,0	308,9	0,000	7	6	3,01	4,26	3,01	2 229,9	1 720,2	40,9
EG10	F1	2165	6,9	0,19	0,57	4,2	96,7	9,3	0,318	0	0	0,47	0,66	0,55	1 313,7	1 532,8	36,8
	F12	2165	18,1	0,02	0,20	14,9	295,7	88,2	0,503	6	1	0,52	0,73	0,81	1 313,7	1 532,8	36,8
	F123	2165	59,4	0,00	0,07	53,1	1712,7	904,7	0,441	46	8	10,30	14,57	10,30	1 313,7	1 532,8	36,8
	F2	2165	4,6	0,19	0,59	5,9	98,7	7,9	0,548	1	0	0,41	0,59	0,57	1 313,7	1 532,8	36,8
	F3	2165	5,1	0,18	0,57	5,7	109,8	7,0	0,641	0	0	0,31	0,44	0,61	1 313,7	1 532,8	36,8
	L1	2165	32,5	1,28	4,06	40,6	703,0	356,5	0,000	9	6	3,23	4,57	3,23	1 313,7	1 532,8	36,8
IG10	F1	1815	4,2	0,20	0,61	4,5	68,6	5,1	0,745	0	0	0,42	0,59	0,47	1 904,1	1 617,1	39,4

Źródło: opracowanie własne



Rysunek 5. Wyniki klasyfikacji mierników dopasowania metodą Czekanowskiego

Źródło: opracowanie własne

Tabela 6. Wyniki klasyfikacji mierników zgodności metodą Czekanowskiego

Grupa A	**Grupa B**
n - liczba danych	chi - test chi kwadrat
kurt - wsp. spłaszczenia	**p(chi) - p-stwo w teście chi kwadrat**
skew - wsp. asymetrii	M2 - miernik podobieństwa rozkładów
v-factor - wsp. zmienności	M3 - miernik podobieństwa rozkładów
Grupa C	**Grupa D**
M5 - miernik podobieństwa rozkładów	M1 - miernik podobieństwa rozkładów
KS1 - stat. Kołmogorowa-Smirnowa	**M4 - miernik podobieństwa rozkładów**
KS1 - stat. Kołmogorowa-Smirnowa	
KS1 - stat. Kołmogorowa-Smirnowa	
Z-0,05 - test z	
Z-0,01 - test z	

Źródło: opracowanie własne

Tabela 7. Wybór najlepszych mierników dopasowania w podziale na grupy

nazwa	śr. - grupa	śr. - poza grupą	iloraz
v-factor	3,00	9,00	3,00
n	2,76	8,47	3,06
skew.	1,74	8,66	4,99
kurt.	1,71	8,58	5,01
-p(chi)	5,26	7,91	1,50
chi	4,51	7,65	1,70
M3	2,96	8,58	2,90
M2	2,95	8,58	2,91
z - 0,01	3,37	8,17	2,42
M5	2,83	7,37	2,61
z - 0,05	3,42	8,98	2,63
KS3	1,75	8,00	4,56
KS1	1,74	7,99	4,59
KS2	1,74	7,99	4,59
M4	0,86	8,13	9,46
M1	0,86	8,56	9,96

Źródło: opracowanie własne

Elementy składowe oraz wartości wskaźnika służącego do ustalenia reprezentantów wyodrębnionych skupisk zebrano w tab. 7. W wyniku omówionej procedury zbiór 16 mierników dopasowania został zredukowany do 4 parametrów:

- Współczynnik zmienności elementów analizowanego zbioru liczb
- Poziom istotności krytycznej wartości testu chi kwadrat
- Statystyka testu Z przy poziomie istotności 0,01
- Miernik podobieństwa M4

Mierniki te zostały wyznaczone dla wszystkich 8 zbiorów danych w odniesieniu do rozkładów następujących cyfr (por. tab. 8)

- F1 –pierwsza F12 – dwie pierwsze F123 – trzy pierwsze
- F2 – druga F3 – trzecia L1 - ostatnia

Tabela 8. Mierniki zgodności rozkładu cyfr z prawem Benforda w analizowanych zbiorach danych

A	B	G	J	L	P
	test	M4	-p(chi)	z - 0,0	v-fact
EG09	F1	5,1	0,209	0	1 371,4
EG09	F12	14,7	0,878	0	1 371,4
EG09	F123	52,2	0,385	17	1 371,4
EG09	F2	6,7	0,332	0	1 371,4
EG09	F3	5,0	0,801	0	1 371,4
EG09	L1	38,6	0,000	6	1 371,4
IG09	F1	4,0	0,747	0	2 229,9
IG09	F12	17,1	0,796	0	2 229,9
IG09	F123	55,1	0,872	7	2 229,9
IG09	F2	4,3	0,950	0	2 229,9
IG09	F3	5,7	0,741	0	2 229,9
IG09	L1	41,0	0,000	6	2 229,9
EG10	F1	4,2	0,318	0	1 313,7
EG10	F12	14,9	0,503	1	1 313,7
EG10	F123	53,1	0,441	8	1 313,7
EG10	F2	5,9	0,548	0	1 313,7
EG10	F3	5,7	0,641	0	1 313,7
EG10	L1	40,6	0,000	6	1 313,7
IG10	F1	4,5	0,745	0	1 904,1
IG10	F12	16,7	0,745	0	1 904,1

Źródło: opracowanie własne

6. Wyniki iteracyjnej metody analizy poprawności zbioru danych

Uzyskane wartości mierników zgodności pozwoliły porangować analizowane zbiory z punktu widzenia stopnia ich zgodności z prawem Benforda. Uwzględniając 4 wybrane parametry zgodności, 6 układów cyfr (testów) oraz 8 zbiorów danych a następnie stosując metodę rangowania zbiorów danych od największego (ranga1) do najmniejszego (ranga 8) stopnia zgodności z regułami wynikającymi z prawa Benforda uzyskano następującą kolejność analizowanych zbiorów (obok identyfikatorów zbiorów przytoczone są sumy rang charakteryzujące syntetyczną poprawność zbioru):

EG10 – 79	EG09 – 88	IP10 – 105	IP09 -107
IG10 – 108	IG09 -113	EP09 – 114	**EP10 - 125**

Dalszej analizie został poddany zbiór EP10 jako najgorzej dopasowany do praw rozkładu cyfr znaczących. Z uwagi na ograniczone ramy opracowania przytacza się wyniki analizy uzyskane przy wykorzystaniu jako kryterium decyzyjnego tylko jednego miernika zgodności – testu Z.

W pierwszym kroku wyznaczono wartości tego testu (Z) dla wybranego zbioru danych (EP10) dla pierwszych cyfr znaczących (F1). Największą wartość statystyki Z w teście F1 zaobserwowano dla pierwszej cyfry równej 4 (por. tab. 9).

Tabela 9. Statystyki Z w teście F1 dla zbioru EP10.

F1	Emp - L	Benf %	Benf - L	Emp %	abs(z)
1	18776	30,10	18844	30,00	0,59
2	11033	17,61	11023	17,63	0,11
3	7832	12,49	7821	12,51	0,14
4	6154	9,69	6066	9,83	1,19
5	4901	7,92	4957	7,83	0,82
6	4178	6,69	4191	6,67	0,20
7	3605	5,80	3630	5,76	0,43
8	3247	5,12	3202	5,19	0,82
9	2871	4,58	2864	4,59	0,13

Źródło: opracowanie własne

W kolejnym kroku uwzględniono w analizie jedynie elementy zbioru rozpoczynające się od cyfry ‘4’. Wynikowe wartości statystyk Z w tym podzbiorze danych przytoczone są w tab. 10. Tym razem najmniejszą zgodność zaobserwowano dla kombinacji cyfr „48”.

Tabela 10. Statystyki Z w teście F12 dla zbioru EP10

F12	Emp - L	Benf %	Benf - L	Emp %	abs(z)
40	684	1,07	671	1,09	0,50
41	656	1,05	655	1,05	0,04
42	639	1,02	640	1,02	0,02
43	632	1,00	625	1,01	0,28
44	625	0,98	611	1,00	0,57
45	625	0,95	597	1,00	1,13
46	593	0,93	585	0,95	0,35
47	559	0,91	572	0,89	0,56
48	609	0,90	560	0,97	2,06
49	531	0,88	549	0,85	0,78

Źródło: opracowanie własne

W trzecim, ostatnim kroku procedury wyodrębniono podzbiór elementów zaczynających się od kombinacji „48” i ponownie wyznaczono wartości statystyk Z (tab. 11). Maksymalną wartość test Z osiąga w tym przypadku dla sekwencji cyfr „480”.

Tabela 11. Statystyki Z w teście F123 dla zbioru EP10

F123	Emp - L	Benf %	Benf - L	Emp %	abs(z)
480	110	0,09	57	0,18	7,11
481	62	0,09	56	0,10	0,74
482	61	0,09	56	0,10	0,63
483	57	0,09	56	0,09	0,11
484	51	0,09	56	0,08	0,68
485	55	0,09	56	0,09	0,13
486	51	0,09	56	0,08	0,65
487	51	0,09	56	0,08	0,63
488	62	0,09	56	0,10	0,86
489	49	0,09	55	0,08	0,87

Źródło: opracowanie własne

W rezultacie wyłoniony został podzbiór elementów, który w pierwszej kolejności powinien być poddany szczegółowej analizie mającej na celu weryfikację poprawności tych danych. Liczba tych elementów w porównaniu z liczebnością zbioru wyjściowego jest niewielka. W omawianym przykładzie zbiór EP10 liczył ponad 60 tys. elementów, natomiast od sekwencji „480" w tym zbiorze zaczynały się tylko 83 elementy.

7. Podsumowanie

Przytoczona iteracyjna procedura identyfikacji fragmentów zbioru danych o najmniejszym stopniu zgodności z prawami rozkładu cyfr znaczących pozwala bardziej skutecznie wykorzystywać narzędzia oceny poprawności zbiorów danych w oparciu o reguły wynikające z prawa Benforda. Podstawową zaletą jest tu ograniczenie pracochłonności związanej ze szczegółową analizą poprawności dokumentów i faktów związanych z konkretnymi danymi w zbiorze. Liczbę takich elementów można ograniczyć do kilkunastu co znacznie ułatwia pracę weryfikatora rzetelności danych.

Ponadto w pracy zaproponowano wykorzystanie taksonometrycznych metod klasyfikacji mających na celu ustalenie stopnia podobieństwa wskazań różnych mierników zgodności porównywanych rozkładów. W rezultacie możliwe jest wybranie mniejszego podzbioru mierników (w krańcowym przypadku tylko jednego), który będzie wykorzystany w trakcie analiz. Takie podejście pozwala przeprowadzać analizy przy znacznie mniejszym nakładzie czasu a równocześnie przy zachowaniu niezbędnej kompleksowości.

Bibliografia

1. Benford F., *The law of anomalous numbers*, Proceedings of the American Philosophical Society, Vol. 78, No. 4, 1938.
2. De Marchi S., J. Hamilton, *Assessing the Accuracy of Self-Reported Data: An Evaluation of the Toxics Release Inventory*, Journal of Risk and Uncertainty, 32/2006.
3. Farbaniec M., Grabiński T., Zabłocki B., Zając W., *Wykorzystanie praw rozkładu cyfr do oceny wiarygodności danych finansowo-księgowych na wybranych przykładach,* X Międzynarodowy Kongres Kontroli Wewnętrznej, Audytu Wewnętrznego, Antykorupcji i Zwalczania Oszustw, Krakowska Akademia im. A.F. Modrzewskiego, Kraków 2011
4. Farbaniec M., Grabiński T., Zabłocki B., Zając W., *Wyniki wyborów powszechnych w Polsce w latach 2000-2010 w świetle analizy rozkładów cyfr* znaczących, rozdz. XI w monografii W poszukiwaniu skutecznych narzędzi i modeli analizy zjawisk społeczno-gospodarczych,

Wyższa Szkoła Przedsiębiorczości i Marketingu w Chrzanowie, Centrum Szkolenia i Organizacji Systemów Jakości PK, Chrzanów 2012, str. 161-179
5. Giles D. E., *Benford's Law and naturalyoccuringproces in certain eBay auctions,* Econometric Working Papers EWP 0505, Univ. of Victoria, 2005.
6. Główny Urząd Statystyczny - http://www.stat.gov.pl/gus/intrastat_PLK_HTML.htm [21.10.2012]
7. Hill T. P. *The first digital phenomenon,* American Scientist, 86/1998.
8. Ley E., *On a pecular distribution of the U.S. stock indicies digits,* American Statistician, 1996.
9. Ministerstwo Finansów http://www.mf.gov.pl/files/sluzba celna/intrastat/instrukcjaintrastat v1 1_2011-09-06.pdf [21.10.2012]
10. Newcomb S., *Note on the Frequency of Use of the Different Digits in Natural Numbers,* American Journal of Mathematics, Vol. 4, No. 1. (1881), pp. 39-40.
11. Nigrini M., *A taxpayer compliance application of Bedford's Law*, Journal of the American Taxation Associates, 18/1996, p. 72-92;
12. Nigrini M., *Adding Value with Digital Analysis,* The Internal Auditor, 56/1999, p. 21–23.
13. Nigrini M., *Digital Analysis Using Benford's Law,* Global Audit Publications, 2000.
14. Nigrini M., *Digital Analysis Using Benford's Law: Tests and Statistics for Auditors,* The EDP Audit, Control, and Security Newsletter, EDPACS, 28/2001;
15. Nigrini M., *I've Got Your Number,* Journal of Accountancy, 187/ 1999
16. Ryder P., *Multiple origins of the Newcomb - Benford law: rational numbers, exponential growth and random fragmentation*, Staats - und Universitätsbibliothek Bremen, Germany, 2009.
17. Tam Cho W.K, Gaines B. J., *Breaking the (Bedford) Law: Statistical fraud detection campaign finance*, The American Statistician, 61/2007, p.218-223.

Use of Benford's Law in the Analysis of Financial Data Correctness on the Example of Information on Trade in Goods

The article presents an example of the financial data correctness analysis using Benford's law. Its objective is to suggest a recursive method that enables gradual narrowing of the set sizes where one can potentially suspect irregularities. This article presents data concerning the trade in goods in Poland in 2009-2010.

Keywords: Benford's law, INTRASTAT, Central Statistical Office, distribution of significant digits, data analysis, recursion, financial audit