# Univerzita Karlova v Praze

## Filozofická fakulta

**Ústav informačních studií a knihovnictví**

Obor: Informační studia a knihovnictví


**Diplomová práce**

# Aplikace Benfordova zákona ve scientometrii

Application of Benford's law in scientometrics


David Jiří Šlosar


Praha 2020

Vedoucí práce: prof. RNDr. Jiří Ivánek, CSc.

**Prohlášení:**

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 29. července 2020 ……..

Podpis studenta

## Poděkování:

Chtěl bych poděkovat prof. RNDr. Jiří Ivánkovi, CSc. za jeho intenzivní podporu při psaní této práce a také za seznámení s Benfordovým zákonem, čímž výrazně ovlivnil mé výzkumné zaměření. Dále patří poděkování Knihovně Akademie věd České republiky, která poskytla zdroje, bez kterých by realizace této práce byla velmi náročná a také nadřízeným, kteří mi umožnili studium při zaměstnání.

## Klíčová slova

Benfordův zákon, scientometrie, citační data, bibliometrie,

## Abstrakt

Tato diplomová práce je zaměřena na zjištění míry sledování Benfordova zákona v citačních datech. Detailně jsou popsána data i jejich získání. Nejrozsáhlejší analýza byla provedena nad datasetem 8,6 milionu záznamů vědeckých výstupů z databáze Web of Science, za pětileté období, s výběrem tří nejpočetnějších a nejcitovanějších typů dokumentů. Pro zjištění míry sledování Benfordova zákona byla použita popisná statistika MAD (Mean Absolute Deviation). Také byla zjištěna míra sledování Benfordova zákona u dvou datasetů, produkce veřejných vysokých škol České republiky a u Akademie věd České republiky a to za stejných podmínek, jako u dalších analýz.

## Keywords

Benford law, scientometrics, citation data, bibliometrics

## Abstract

This diploma thesis is focused on determining the degree of presence of Benford‘s law in citation data. The data and their acquisition are described in detail. The most extensive analysis was performed on a dataset of 8.6 million records of scientific outputs from the Web of Science database, over a five-year period, with a selection of the three most numerous and most cited types of documents. Descriptive MAD (Mean Absolute Deviation) statistic were used to determine the degree of presence of Benford's law. The degree of presence of Benford's law was also determined for two datasets, the production of public universities in the Czech Republic and the Academy of Sciences of the Czech Republic under the same conditions as in other analyses.

Obsah

# Úvod

Tato diplomová práce se zabývá přítomností Benfordova zákona v citačních datech. Citačními daty se v této práci rozumí prosté počty citací u jednotlivých záznamů. Staženy jsou všechny dostupné záznamy z Web of Science Core Collection za roky 2014 až 2018, všechny obory WoS Categories a typy dokumentů Article, Proceedings paper a Review. Jako hlavní popisná statistika slouží MAD (Mean Absolute Deviation).

Výzkumné otázky:

1. Do jaké míry sledují citační data záznamů z databáze Web of Science za roky 2014 a 2018, typů dokumentů Article, Proceedings paper a Review, všech oborů WoS Categories Benfordův zákon?
2. Jaké jsou rozdíly v míře sledování Benfordova zákona u jednotlivých kombinací let, typů dokumentů a oborů?
3. Do jaké míry sledují citační data veřejných vysokých škol České republiky a Akademie věd České republiky Benfordův zákon?

V oblasti scientometrie bylo na toto téma publikováno jen malé množství článků. Analýza nad tak velkým datovým souborem, jaký používá tato práce, nebyla nalezena.

V teoretické části je představen Benfordův zákon a jeho teoretické aspekty, které jsou pro tuto práci relevantní. Následuje stručné porovnání s třemi hlavními bibliometrickými zákony. Praktická část obsahuje popis výběru zdroje citačních dat, jeho vlastností, vlastností získaných citačních dat, způsob jejich získání, zpracování, kategorizování a vyhodnocení. Přítomny jsou taktéž četné tabulky a grafy ilustrující získané výsledky.

Realizací této práce jsou poskytnuty podklady pro vytvoření ucelené představy o přítomnosti Benfordova zákona v citačních datech z databáze Web of Science. V praktické části je provedeno zejména šest analýz na globálních datech databáze Web of Science. Analýzy pracují s různě agregovanými daty, tak aby ilustrovaly aspekty sledování Benfordova zákona v citačních datech pro různé kategorie. Je tak zohledněn vliv roků vydání, typů dokumentů a oborů. Dále jsou analyzována agregována citační data za veřejné vysoké školy České republiky a Akademii věd České republiky.

Tato práce obsahuje 109 057 znaků.

# 1. Benfordův zákon

## 1.1 Vznik

Byl to Simon Newcomb, který roku 1881 jako první publikoval článek o Benfordově zákonu, nikoliv Frank Benford, který publikoval své podání až roku 1938. Na Newcomba jakožto objevitele se zapomnělo a dokonce nebyl ani v Benfordově článku citován. Benford totiž nevěděl o existenci Newcombova článku a zákon v podstatě znovuobjevil. Dle počtu citací na původní články (Benford 1703 citací, Newcomb 901[1] citací) a názvu tohoto statistického zákona je pravděpodobné, že další citace budou udělovány spíše Benfordovi. Nutno poznamenat, že Benford daný zákon rozpracoval mnohem podrobněji a šířeji. Také zákon nazval „Law of Anomalous Numbers“, což se ale nestalo normou. (Newcomb, 1881 a Benford, 1938)

Newcomb a shodně i Benford objevili tento statistický zákon tím, že zaregistrovali míru opotřebení na logaritmických tabulkách (také na tabulkách odmocnin a trigonometrických). První strany těchto tabulek byly výrazně opotřebované užíváním, zatímco poslední strany byly téměř jako nové. Začátky logaritmických tabulek totiž obsahují čísla začínající jedničkou a s postupem knihy dále se čísla zvyšují, až na posledních stranách jsou čísla začínající devítkou. Oba vědci shodně usoudili, že uživatelé z různých vědních oborů, kteří s těmito tabulkami pracovali, také operovali častěji s čísly začínajícími na jedničku. (Newcomb, 1881 a Benford 1938)

Newcomb začal úvahou o pravděpodobnostním zastoupení prvních signifikantních[2] číslic v přirozeně vzniklých datasetech čísel. Dále postupoval odvozováním s logaritmy a antilogaritmy[3]. Nakonec stanovil tabulku pravděpodobnosti výskytu číslic na první signifikantní pozici, s tím, že určil i pravděpodobnosti výskytů signifikantních číslic na dalších pozicích. (Newcomb, 1881)

Oproti tomu Benford přímo sledoval různé datasety a jejich chování. Z pozorování stanovil premisu, že datasety by měly pro každé číslo obsahovat alespoň 4 číslice (například 1125, 3,124 či 0,006987), pocházet z různých na sobě nezávislých zdrojů, pocházet z nestrukturovaných zdrojů (lépe sledovaly Benfordův zákon čísla z titulních stran novin než strukturovaných

[1] Počty citací byly přejaty z Google Scholar (GS) 12.2 2020. Autor si je vědom jistých omezení v přesnosti počtů citací od GS, neobjevil však jinou citační databázi, kde by byly oba články indexovány.

[2] Signifikantní číslice je u následujících čísel vždy trojka: 3,14159; 3587; 0,0316. Vychází se z matematického zápisu, kdy stejná čísla lze vyjádřit jako: 3,14159; $3{,}587 * 10^{3}$; $3{,}16 * 10^{-2}$. Při tomto zápisu je již první signifikantní číslice jasnější, čili první zleva. Za první signifikantní číslici tedy není považována nula.

[3] Pokud je dekadický logaritmus 1000 roven 3, tak dekadický antilogaritmus 3 je roven 1000. Tento pojem byl široce používán v éře ručních výpočtů a za pomoci tabulek.

matematických tabulek), být dostatečně velké (nestanovil však jak moc velké datasety, nejmenší jím užitý dataset obsahoval 91 záznamů), nemít striktně omezený rozsah čísel a být příliš silně ovlivněné konstrukcí datasetu. Dále uvedl vzorec výpočtu pravděpodobnosti první signifikantní číslice spolu s letmým odvozením. (Benford 1938)

## 1.2 Definice

Intuitivně by se dalo očekávat, že v libovolném vzorku dat je pravděpodobnost, že číslo bude začínat konkrétní číslicí, přibližně stejná pro všechny číslice. Existuje velké množství datasetů, které tuto vlastnost nemají a jednou množinou takových datasetů jsou takové, které se blíží Benfordovu zákonu.

Benfordův zákon je statistický zákon, který popisuje, s jakou pravděpodobností se bude určitá číslice vyskytovat na prvním signifikantním místě čísla. Dataset, ve kterém je tento zákon sledován, však musí splňovat jisté předpoklady:

1. Dataset nesmí mít žádné radikální omezení rozsahu hodnot. Například dataset výšek lidí nebo hodnot IQ v populaci dospělých jedinců prakticky nemůže obsahovat čísla začínající trojkou.

2. Konstrukce datasetu nesmí být silně zatížena lidskou vůlí. Pokud se jedná o cílené generování pseudonáhodným generátorem, kde je záměrným výsledkem set neopakujících se či rovnoměrně rozložených čísel, tak zde se žádný Benfordův zákon neobjeví. Stejně tak ceny produktů v katalozích nabídek ve většině obchodů mají za účelem dosažení psychologického efektu a lepších tržeb modifikované hodnoty cen. Produkty jsou často prodávány v ceně typu 9,90 Kč místo 10 Kč. Statistické rozložení první číslice v cenách jednotlivého zboží tedy tíhne k číslici devět.

Benfordův zákon však platí například v případě tržeb zákazníků. Ti totiž nenakupují/neplatí po jednotlivých položkách, ale platí za nákup jakožto celek. Zde se objevuje efekt takzvané náhodné lineární kombinace (Random Linear Combination). Zatížení cen jednotlivých produktů psychologickým efektem je z hlediska Benfordova zákona téměř anulováno díky třem faktorů náhody a nejistoty: počtem produktů zakoupených zákazníkem, které produkty si zákazník koupí a kolik kusů od každého produktu si zákazník koupí. Tyto kombinace opět vedou k logaritmickému rozložení první signifikantní číslice v datasetu. (Kossovsky, 2015, s. 49)

3. Rozsah čísel v datasetu musí být dostatečně velký. Kossovsky tvrdí, že datasety, pro které platí rozdíl maximální a minimální hodnoty $F_{diff} > 3$, podle vzorce $F_{diff} = \log(max) - \log(min)$, tak jsou dostatečně robustní pro spolehlivé porovnání relativních četností prvních signifikantních číslic vůči pravděpodobnostem Benfordova zákona (Kossovsky, 2015). Dále však doporučuje místo použití minimální hodnoty v datasetu a hodnoty maximální vložit do vzorce hodnotu 90. percentilu a 10. percentilu.

4. Dataset by měl obsahovat dostatečně velký počet čísel. V příliš malých datasetech totiž není možné, vlivem nízkého počtu hodnot, přiblížit se relativní četností první signifikantní číslice k pravděpodobnostem dle Benfordova zákona. Pokud by byl uvažován dataset o například 30 prvcích, který by se měl četnostmi výskytů prvních signifikantních číslic blížit pokud možno co nejvíce blížit Benfordově zákonu, tak by vypadal takto: (1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8, 9). Pro číslici jedna je rozdíl v relativní četnosti v datech vůči pravděpodobnosti výskytu dle Benfordova zákona velmi malý. Je naměřena jednička na prvním místě s relativní četností 30% a očekávána s 30,103%. Pro číslici pět je rozdíl již významnější (naměřeno 10% a očekáváno 7,918%). Například dataset o devíti hodnotách nemá vůbec smysl uvažovat, jelikož by v něm při dodržení Benfordova zákona některé číslice ani nebyly obsaženy. Tento odstavec bude v kontextu k citačním datům rozvinut v kapitole 3.3.

Vzorec pro výpočet pravděpodobnosti výskytu první signifikantní číslice *d* dle Benfordova zákona je:

$$F_d = \log_{10}(\frac{d+1}{d})$$

kde $F_d$ je hodnota pravděpodobnosti. Desítkový logaritmus udává, že výpočet je platný pro desítkovou soustavu. A číslice $d \in D = (1, 2\ldots 9)$. (Benford, 1938)

Pro číslici jedna tedy platí: $F_1 = \log_{10}(\frac{1+1}{1}) = 0{,}30103$. Pravděpodobnost, že se jednička vyskytuje na místě první signifikantní číslice je tedy přibližně 30,103%. Součet pravděpodobností všech prvních signifikantních číslic datasetu je roven 1, tedy 100% pravděpodobnost. Je to dáno tím, že čísla 0 či 0,00 jsou z datasetu a priori vyřazena, neobsahují totiž signifikantní číslici. V tabulkách dále bude pravděpodobnost vyjadřována ve tvaru desetinného čísla z jednotkového intervalu. Pouze v případě Obrázku 2. je pravděpodobnost uvedena ve tvaru 30,103%, jedná se o přímo převzatý obrázek z literatury.

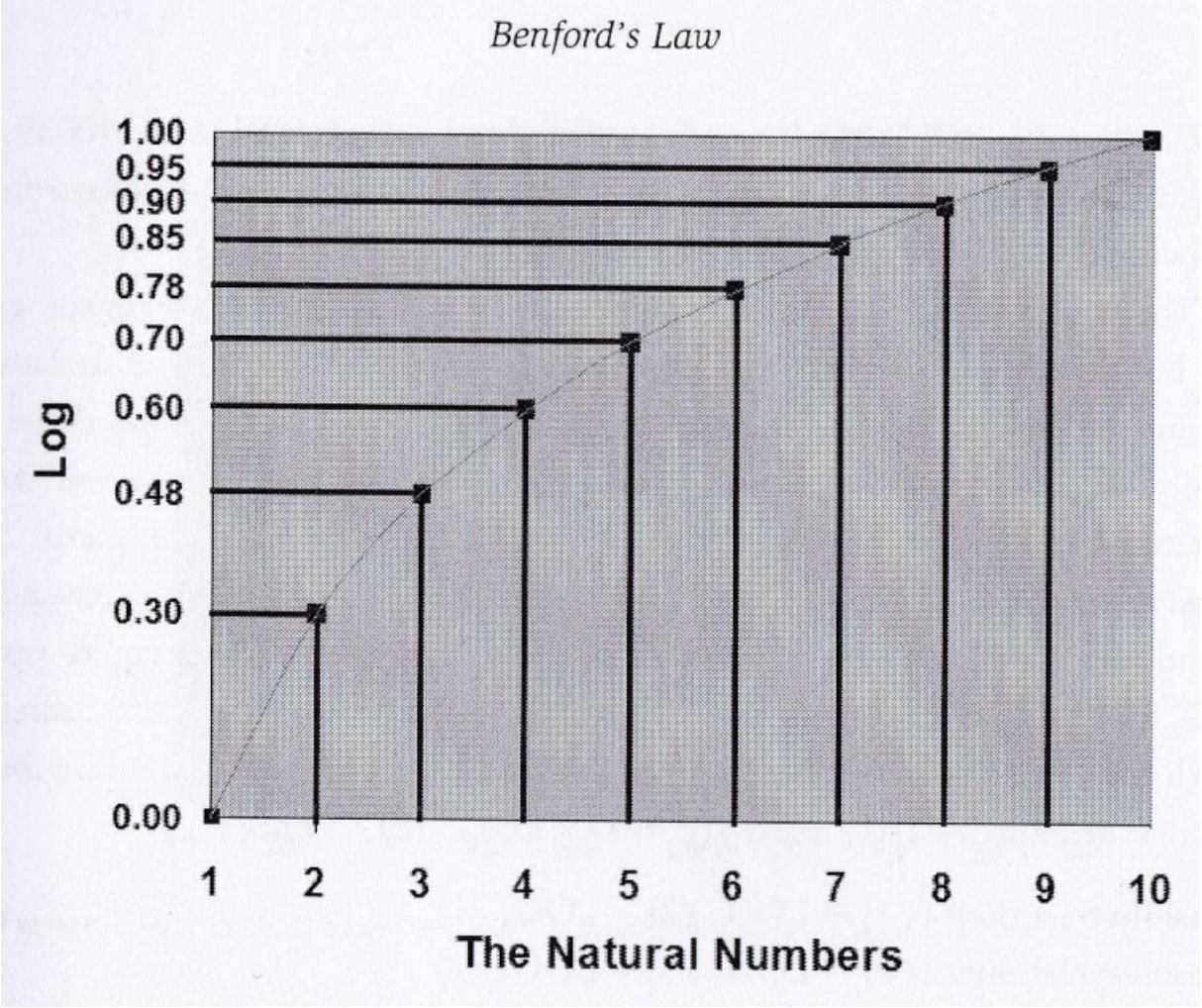
| První signifikantní číslice | $F_d = \log_{10}(\frac{d+1}{d})$ | Pravděpodobnost [%] |
|---|---|---|
| 1 | 0,30103 | 30,103 |
| 2 | 0,17609 | 17,609 |
| 3 | 0,12494 | 12,494 |
| 4 | 0,09691 | 9,691 |
| 5 | 0,07918 | 7,918 |
| 6 | 0,06695 | 6,695 |
| 7 | 0,05799 | 5,799 |
| 8 | 0,05115 | 5,115 |
| 9 | 0,04576 | 4,576 |

*Vypočtené hodnoty pro první signifikantní číslici dle Benfordova zákona. Tabulka 1.*

Z vypočtených hodnot v tabulce 1. je zřejmý klesající trend výskytu pravděpodobnosti vůči velikosti první signifikantní číslice.

Ačkoliv je vzorec udán pro desítkovou soustavu, tak na ni není omezen. Benfordův zákon je aplikovatelný (při vhodném datasetu) i na jiné soustavy. Je známo, že rozdíl logaritmů je logaritmus podílu, takže je možné zvolit alternativní způsob zápisu pro pravděpodobnost první signifikantní číslice $F_d = \log_n(d+1) - \log_n(d)$, kde *d* je číslice od jedné do devíti a *n* udává

soustavu, ve které se výpočet pohybuje. Výsledkem tohoto vzorce je pravděpodobnost výskytu první signifikantní číslice. Tento zápis také lépe ilustruje, proč lze Benfordův zákon nazvat i logaritmickou distribucí, k tomu dále Obrázek 1.




*Logaritmický průběh (Kossovsky, 2015). Obrázek 1.*

Je možno sledovat i další signifikantní číslice v případě, že to vlastnosti datasetu umožňují. U čtvrté a další signifikantní číslice se však rozložení stává silně uniformním.

| $d$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Prob$(D_1 = d)$ | 0 | 30.10 | 17.60 | 12.49 | 9.69 | 7.91 | 6.69 | 5.79 | 5.11 | 4.57 |
| Prob$(D_2 = d)$ | 11.96 | 11.38 | 10.88 | 10.43 | 10.03 | 9.66 | 9.33 | 9.03 | 8.75 | 8.49 |
| Prob$(D_3 = d)$ | 10.17 | 10.13 | 10.09 | 10.05 | 10.01 | 9.97 | 9.94 | 9.90 | 9.86 | 9.82 |
| Prob$(D_4 = d)$ | 10.01 | 10.01 | 10.00 | 10.00 | 10.00 | 9.99 | 9.99 | 9.99 | 9.98 | 9.98 |

*Pravděpodobnosti výskytů n-tých signifikantních číslic (Berger, 2015, s. 15). Obrázek 2.*

Pro některé praktické aplikace je možné vyhotovit i pravděpodobnost výskytu první signifikantní dvojice. To je realizováno vzorcem:

$$F_{da} = \log_{10}(1 + \frac{1}{d * 10 + a * 1})$$

Kde *d* je rovno první signifikantní číslici a *a* je rovno druhé signifikantní číslici. Pro 34 tedy platí: $F_{34} = \log_{10}\left(1 + \frac{1}{3*10+4*1}\right) = 0{,}01259$. Pravděpodobnost, že se 34 vyskytuje jakožto první signifikantní dvojice je tedy přibližně 1,259%.

Níže uvedená Tabulka 2. zobrazuje přehled možných prvních signifikantních dvojic. Vynechána jsou čísla 00 až 09 ze stejného důvodu jako u první signifikantní číslice. Nulice uvedená zleva není považována za signifikantní číslici. Přirozenou vlastností je, že součet pravděpodobností všech dvojic začínajících na jedničku (10, 11, …, 19) je právě pravděpodobností, že první číslice je jednička. Z toho je možné dále odvodit, že pravděpodobnost pouze druhé signifikantní číslice je součtem pravděpodobností všech dvojic končících na stejnou číslici, což je možno vyjádřit vzorcem $F_a = \sum_{d_1=1}^{9} \log_{10}(1 + \frac{1}{d*10+a*1})$, (Nigrini, 2012, s. 5). Například pravděpodobnost, že druhá signifikantní číslice je právě nula, je součtem pravděpodobností všech prvních signifikantních dvojic končících na nulu. Tedy 11,968%, což je po hrubém zaokrouhlení v souladu s tabulkou na Obrázku 2.

| První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 0,04139 | 20 | 0,02119 | 30 | 0,01424 | 40 | 0,01072 | 50 | 0,00860 |
| 11 | 0,03779 | 21 | 0,02020 | 31 | 0,01379 | 41 | 0,01047 | 51 | 0,00843 |
| 12 | 0,03476 | 22 | 0,01931 | 32 | 0,01336 | 42 | 0,01022 | 52 | 0,00827 |
| 13 | 0,03218 | 23 | 0,01848 | 33 | 0,01296 | 43 | 0,00998 | 53 | 0,00812 |
| 14 | 0,02996 | 24 | 0,01773 | 34 | 0,01259 | 44 | 0,00976 | 54 | 0,00797 |
| 15 | 0,02803 | 25 | 0,01703 | 35 | 0,01223 | 45 | 0,00955 | 55 | 0,00783 |
| 16 | 0,02633 | 26 | 0,01639 | 36 | 0,01190 | 46 | 0,00934 | 56 | 0,00769 |
| 17 | 0,02482 | 27 | 0,01579 | 37 | 0,01158 | 47 | 0,00914 | 57 | 0,00755 |
| 18 | 0,02348 | 28 | 0,01524 | 38 | 0,01128 | 48 | 0,00895 | 58 | 0,00742 |
| 19 | 0,02228 | 29 | 0,01472 | 39 | 0,01100 | 49 | 0,00877 | 59 | 0,00730 |
| Suma 1a | 0,30103 | Suma 2a | 0,17609 | Suma 3a | 0,12494 | Suma 4a | 0,09691 | Suma 5a | 0,07918 |

| První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost | První sig. dvojice | Pravděpodobnost | Celková suma |
|---|---|---|---|---|---|---|---|---|
| 60 | 0,00718 | 70 | 0,00616 | 80 | 0,00540 | 90 | 0,00480 | |
| 61 | 0,00706 | 71 | 0,00607 | 81 | 0,00533 | 91 | 0,00475 | |
| 62 | 0,00695 | 72 | 0,00599 | 82 | 0,00526 | 92 | 0,00470 | |
| 63 | 0,00684 | 73 | 0,00591 | 83 | 0,00520 | 93 | 0,00464 | |
| 64 | 0,00673 | 74 | 0,00583 | 84 | 0,00514 | 94 | 0,00460 | |
| 65 | 0,00663 | 75 | 0,00575 | 85 | 0,00508 | 95 | 0,00455 | |
| 66 | 0,00653 | 76 | 0,00568 | 86 | 0,00502 | 96 | 0,00450 | |
| 67 | 0,00643 | 77 | 0,00560 | 87 | 0,00496 | 97 | 0,00445 | |
| 68 | 0,00634 | 78 | 0,00553 | 88 | 0,00491 | 98 | 0,00441 | |
| 69 | 0,00625 | 79 | 0,00546 | 89 | 0,00485 | 99 | 0,00436 | |
| Suma 6a | 0,06695 | Suma 7a | 0,05799 | Suma 8a | 0,05115 | Suma 9a | 0,04576 | 1 |

*Vypočtené hodnoty pravděpodobností pro výskyt první dvojice signifikantních číslic dle Benfordova zákona. Tabulka 2.*

V dalších praktických aplikacích je možno jako doplňkové ověření přítomnosti Benfordova zákona v datasetu použít zjištění četnosti výskytu posledních dvou číslic. Podle Benfordova zákona je totiž očekávána pravděpodobnost přesně 1% pro každou dvojici. Toto pravděpodobnostní rozložení je ale nezbytně podmíněno tím, že čísla v datasetu mají alespoň čtyři číslice. Pro první dvě číslice je pravděpodobnostní rozložení dle Benfordova zákona zobrazeno v Tabulce 2. a je na první pohled jasné, že ho rozhodně nelze považovat za rovnoměrné.

Další zajímavostí Benfordova zákona je to, že pokud dataset Benfordův zákon sleduje, tak je možno s datasetem provádět různé operace a Benfordův zákon zůstane zachován, například

při převedení čísel v datasetu do jiné číselné soustavy (Berger, 2015). Pokud jsou data převedena z desítkové do osmičkové soustavy, tak ve vzorci $F_d = \log_{10}(\frac{d+1}{d})$ dojde pouze ke změně z desítkového logaritmu na osmičkový. Následně může být stanovena pravděpodobnost první signifikantní číslice z intervalu od jedné do osmi, devítka se v osmičkové soustavě nevyskytuje. Naopak osmička se v intervalu vyskytuje, jelikož interval začíná jedničkou na rozdíl od jiných konvencí (typicky programování), kde by interval začínal nulou. Opět nula není ve vztahu k Benfordovu zákonu považována za signifikantní číslici. Stejně tak umocnění všech čísel datasetu konstantou větší než jedna či vynásobení nenulovou konstantou, zachová pravděpodobností rozložení signifikantních číslic dle Benfordova zákona. Tento jev je možné snadno empiricky ověřit, což se také stalo na příkladu délky řek. Je lhostejno, zdali jsou délky řek na zemi měřeny v kilometrech, metrech, mílích či stopách. Benfordův zákon je v tomto příkladu platný bez ohledu na jednotky. (Kossovsky, 2015)

# 2. Hlavní bibliometrické zákony

Mezi hlavní tři bibliometrické zákony patří Lotkův, Bradfordův a Zipfův. Je nezbytně nutné uvést, že pojem „zákon“ je mírně zavádějící, avšak etablovaný. Jedná se totiž o statistické aproximace a nejsou vždy a za všech okolností platné. Také je možné tyto zákony nazvat zákony infometrickými a scientometrickými. Zatímco bibliometrie a scientometrie jsou podmnožinami infometrie, tak bibliometrie a scientometrie jsou do značné části překrývající se disciplíny. Přesto, že jsou zde zákony zvány bibliometrickými, tak mají značný přesah mimo informační vědu. (Bawden, 2017)

Lotkův zákon:

Také může být zván Zákonem vědecké produktivity. Výpočet je stanoven pro určitý soubor dokumentů a popisuje, kolik autorů je tvůrcem jednoho, dvou, nebo více dokumentů.

Obecný vzorec vypadá takto: $f_{(y)} = C/y^a$. *C* je konstanta typická pro konkrétní dataset. Pokud není uvažován konkrétní dataset, ale jedná se pouze o ilustraci poměrů mezi počty autorů s počty jimi vytvořených dokumentů, tak je konstanta rovna jedné. Proměnná *y* je rovna počtu článků publikovaných autory s *y* publikovaných článků. Při zjištění, kolik autorů vytvořilo právě jeden dokument v datasetu, tak je *y* rovno jedné. Exponent *a* je také konstanta přibližně rovna dvěma. (Bawden, 2017)

V Grafu 1. je níže zobrazena distribuce Lotkova zákona. Pro Graf 1. platí: $C = 1$, $\alpha = 2$.


*Poměrné zastoupení autorů vůči počtu jimi publikovaných dokumentů dle Lotkova zákona. Graf 1.*

Je patrné, že distribuce je silně nerovnoměrná a zešikmená. Také nápadně připomíná citační křivky ve většině vědních disciplín (jen několik málo prací má velký počet citací a velké množství prací získalo jen malý počet citací).

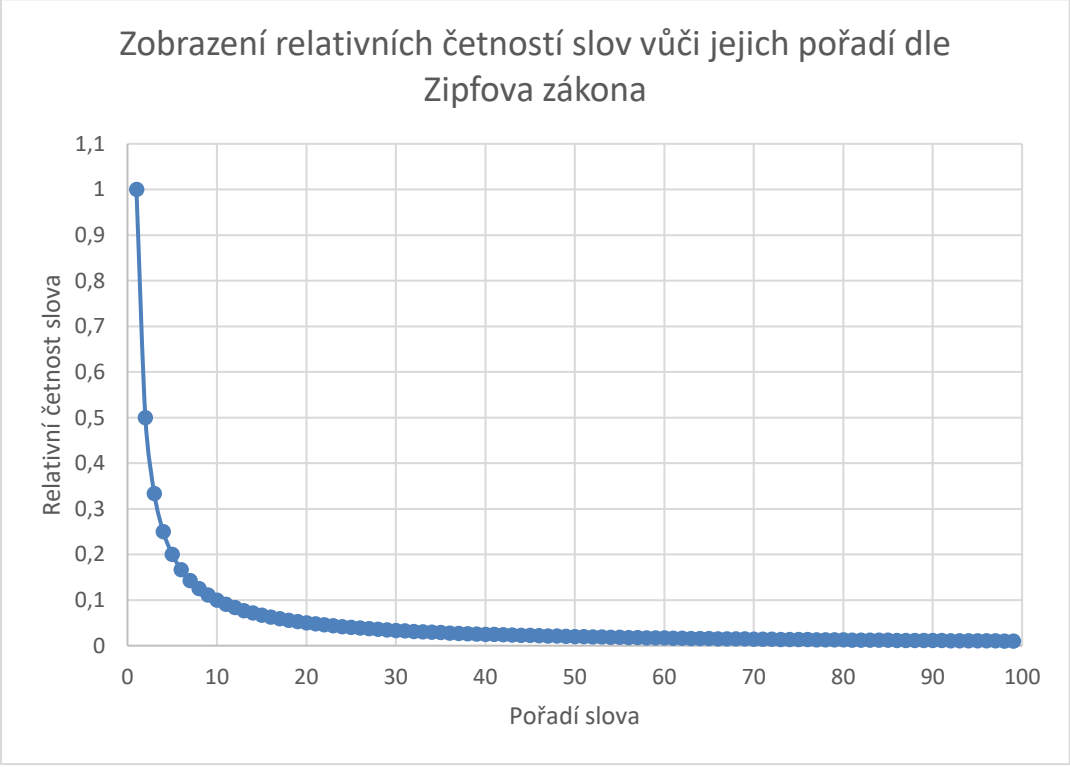
<u>Bradfordův zákon:</u>

Pojednává o rozptýlenosti relevantních dokumentů (vědeckých článků) v relevantních zdrojích (časopisech). Pokud jsou seřazeny zdroje určitého tématu podle počtu v nich obsažených relevantních dokumentů a takto seřazené zdroje jsou rozděleny na tři části podle poměru $1 : n : n^2$ ($n$ je konstanta závislá na datasetu), tak v každé ze tří částí bude obsaženo přibližně stejné množství relevantních dokumentů. Takzvané jádro o malém počtu zdrojů bude obsahovat velké množství relevantních dokumentů, zatímco v „okrajové“ části se bude relativně malé množství relevantních dokumentů nacházet ve velkém množství zdrojů.

Zipfův zákon:

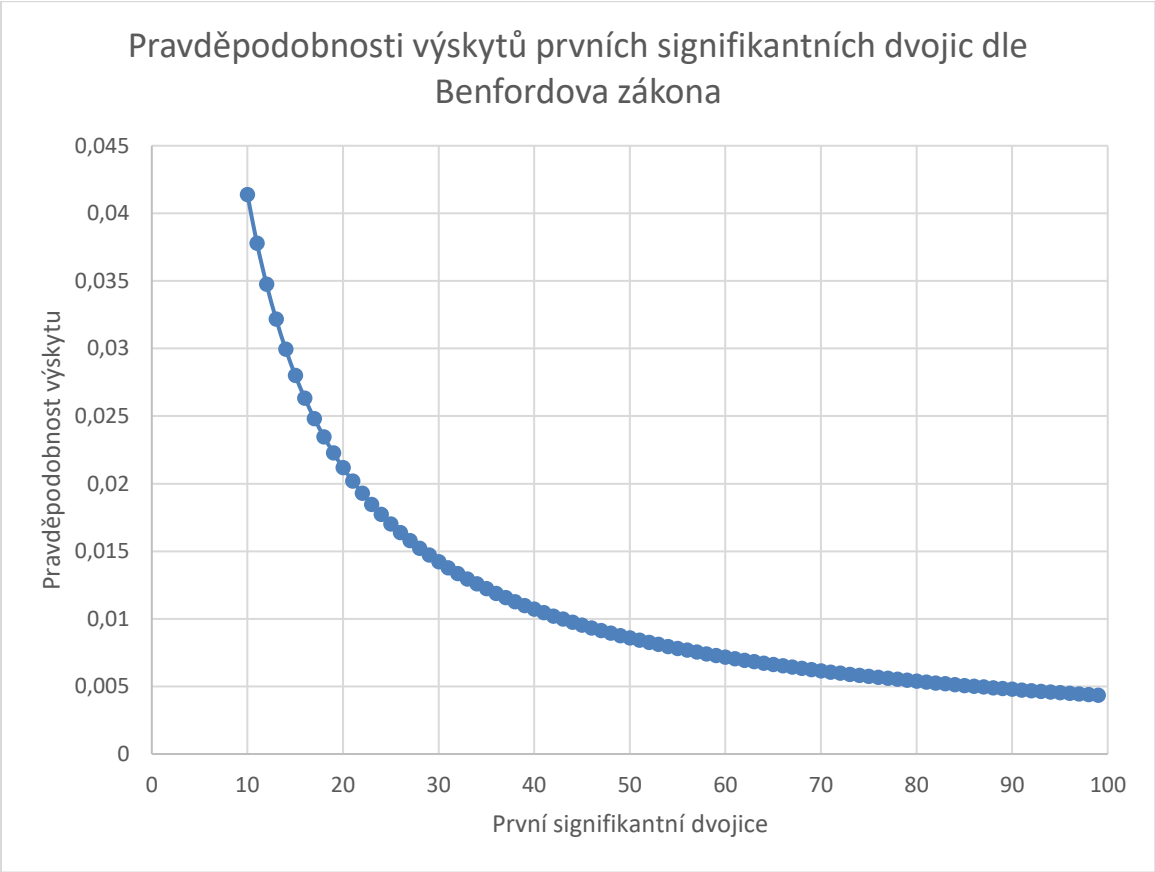
Posledním zde zmíněným hlavním bibliometrickým zákonem je Zipfův zákon, který byl původně odvozen z četností výskytu slov v dokumentu. Tento dokument musí být napsán v přirozeném jazyce (Esperanto je do značné míry odvozeno od přirozených jazyků a platí pro ně Zipfův zákon). Podle tohoto zákona je součin četnosti a pořadí (seřazeno dle četnosti) každého slova dokumentu přibližně konstantní. Pokud by se nejčetnější slovo v dokumentu vyskytovalo stokrát, tak druhé nejčetnější slovo by se vyskytovalo padesátkrát, třetí přibližně třicettřikrát atd. Pro obecný Zipfův zákon platí vzorec $y = k/r$, kde *y* udává četnost slova, *k* je konstanta datasetu a *r* je pořadí daného slova při řazení četnosti slov od nejčetnějšího k nejméně četnému. (Bawden, 2017)

Graf 2. níže ilustruje Zipfův zákon. Konstanta *k* je rovna jedné. Osa y zobrazuje relativní četnosti slova ve vztahu k tomu nejčetnějšímu. První slovo při řazení dle nejčetnějšího po nejméně četné je na ose y 100 % (zobrazeno jako jedna). Druhé slovo je tedy v dokumentu obsaženo s 50 % četnosti nejčetnějšího slova. Osa x udává pořadí slova při řazení od nejčetnějšího po nejméně četné.




*Zobrazení relativních četností slov vůči jejich pořadí dle Zipfova zákona. Graf 2.*

Pro porovnání je zde uveden Graf 3. pravděpodobností výskytů prvních signifikantních dvojic dle Benfordova zákona.




*Pravděpodobnosti výskytů prvních signifikantních dvojic dle Benfordova zákona. Graf 3.*

Pokud jsou porovnány Grafy 1. Lotkův zákon, 2. Zipfův zákon a 3. Benfordův zákon, tak je možné spatřit jisté podobnosti, například klesající tendence křivek. Mezi obecnými zákony Lotkovým a Zipfovým je hlavní rozdíl ve strmostech křivek, který je dán především tím, že Lotkův zákon je mocninnou funkcí. Benfordův zákon oproti dvěma etablovaným bibliometrickým zákonům je méně strmý, avšak stále obsahuje klesající tendenci a ve tvaru jeho křivky je patrný „dlouhý chvost“.

Fundované porovnání uvedených zákonů by vyžadovalo podrobný výzkum, který přesahuje rámec této práce.

# 3. Popis analyzovaných citačních dat

V této práci je hlavním cílem sledování relativních četností výskytů prvních signifikantních číslic u počtů citací záznamů v porovnání s pravděpodobnostmi dle Benfordova zákona. Již publikované články (Alves, 2014), (Alves, 2016) či (Campanario, 2011) jsou zaměřeny na počty citací či počty článků u časopisů a autoři si tak mohli dovolit pracovat se vzorkem dat. U počtů citací záznamů to však možné není. Dostupné citační databáze jako Web of Science, Scopus, Google Scholar či Dimensions totiž neumožňují skutečně náhodný výběr. Záznamy, aby mohly být zobrazeny uživatelům, jsou vždy nějakým způsobem řazeny (podle data vydání, citovanosti, oboru, relevance k vyhledávacímu dotazu atd.) a všechna tato řazení již mohou ovlivnit strukturu dat. Získání jakéhokoliv vzorku nesplňuje požadavek náhodného výběru a není tedy validním postupem. Je tedy zapotřebí získat celé populace a pro tento účel byla jako nejvhodnější vybrána databáze Web of Science.

Z autorovi známých citačních databází pouze Web of Science, Scopus, Dimensions a Microsoft Academic umožňují pro stahování dat použít třídění podle oborů, let a typů dokumentů. Další citační databáze jsou svým rozsahem příliš malé nebo jsou dostupné pouze skrze předplatné, kterým autor nedisponuje. Google Scholar není vybaven oborovým tříděním a jen v omezené míře v něm lze rozlišovat typ dokumentu. Dimensions, Microsoft Academic mají jen velmi omezené možnosti stahování. Scopus sice nabízí stahování až po 20 000 záznamech a je teoreticky možné rozložit celou databázi na takto malé části a postupně ji stáhnout, časově se však jedná o práci v řádech měsíců, což by vzhledem k „živosti“ databáze mohlo ovlivnit pozorované výsledky.

Autor, jakožto zaměstnanec Knihovny Akademie věd České republiky, disponuje jejími zdroji a tedy i API přístupem k Web of Science. Tento přístup umožňuje postupně a částečně automatizovaně získat potřebné záznamy. Přibližně 60% dat použitých v této diplomové práci bylo původně staženo pro účely hodnocení Akademie věd České republiky, jejich opětovným použitím bylo ušetřeno značné množství času. Tímto bylo možné stažení dat pro sledování Benfordova zákona zvládnout v rozsahu pěti týdnů.

## 3.1 Zdroj – Web of Science

Web of Science (dříve Web of Knowledge) je citační databází společnosti Clarivate Analytics (dříve Institute of Scientific Information). Její obsah je dostupný pouze za předplatné a stejně

tak množství kolekcí dostupných je stanoveno dle výše předplatného. Se zdroji Knihovny Akademie věd České republiky jsou k dispozici kolekce Web of Science Core Collection: Citation Indexes a Core Collection: Chemical Indexes.

První index v některých sub-indexech sahá až do roku 1900 a obsahuje sub-indexy: Science Citation Index Expanded, Social Sciences Citation Index, Art & Humanities Citation Index, Conference Proceedings Citatin Index – Science, Conference Proceedings Citation Index – Social Science & Humanities, Book Citation Index – Science, Book Citation Index – Social Sciences & Humanities a Emerging Sources Citation Index.

Druhý index sahá až do roku 1840. A obsahuje sub-indexy Current Chemical Reactions a Index Chemicus.

Celkem tyto kolekce indexují přes 159 milionů záznamů a 1,7 miliardy citací (Web of Science, 2020).

## 3.2 Způsob získání dat

### 3.2.1 Teoretický postup

Ke stažení dat byla vybrána databáze Web of Science Core Collection. Pro účely testování Benfordova zákona byly zvoleny roky 2014 až 2018, typy dokumentů Article, Proceedings paper a Review, a za oborové členění bylo zvoleno WoS Categories (254 oborů).

Volba těchto parametrů je čistě pragmatická. Mají totiž úzkou vazbu na data stažená pro potřeby hodnocení Akademie věd České republiky za roky 2015 - 2019 a autor je právě osobou, která obstarává a zpracovává všechna bibliometrická data pro toto hodnocení. Také zdroje Knihovny Akademie věd České republiky umožňují pracovat právě s těmito daty (týká se typů dokumentů a oborů). Za zdroje jsou považovány dostupné kolekce, API a přístup k prémiovému účtu, jehož prostřednictvím API do databáze Web of Science přistupuje.

Volba roků je záměrně zvolená pod poločas citovanosti[4] většiny oborů. Pokud data ukáží, že je meziroční rozdíl ve sledování Benfordova zákona, tak právě na těchto letech by to mělo být nejlépe vidět. Pokud by byl zvolen desetiletý odstup, tak by tento rozdíl nemusel být patrný.

[4] Poločas citovanosti je medián doby obdržení citací. Například pokud je poločas citovanosti u článku roven 4,5, tak získal právě polovinu svých citací za první 4 roky a 6 měsíců své existence ve sledované databázi.

Volba typů dokumentů je také ovlivněna počtem záznamů v oboru a typu dokumentu a počtu citací v oboru a dokumentu. Jak bylo uvedeno dříve, je zapotřebí mít dostatečně statisticky robustní data. Nejrobustnější je typ dokumentu Article, který obsahuje velké (v rámci databáze Web of Science dokonce největší) množství záznamů a největší množství citací. Proceedings paper je široce zastoupenou a nezanedbatelnou částí databáze Web of Science. Ačkoliv počty citací jsou zde nižší než u druhých dvou zvolených typů dokumentů, tak stále obsahuje více záznamů než Review. Jedná se také po podstatnou formu mezivědecké komunikace. Review, ačkoliv je počtem záznamů nejmenší ze tří zvolených typů dokumentů, tak je velmi intenzivně citován a je tedy zajímavý z hlediska vysokých počtů citací u záznamů.

Z početnějších typů dokumentů v databázi Web of Science, které nebyly vybrány, je možné zmínit Letter, Meeting abstract, Editorial material a Book chapter. Tyto obory nebyly vybrány vzhledem k nízkým počtů výskytů nebo nízkému počtu citací[5]: Letter – 460 394 obdržených citací na 173 881 dokumentů, Meeting abstract – 166 281 obdržených citací na 1 260 754 dokumentů, Editorial material – 1 259 746 obdržených citací na 443 436 dokumentů a Book chapter – 31 665 obdržených citací na 53 653 dokumentů. Oproti zvoleným: Article – 65 157 037 obdržených citací na 6 344 455 dokumentů, Proceedings paper – 1 948 473 obdržených citací na 975 512 dokumentů a Review – 1 0652 996 citací na 449 229 dokumentů.

Počty citací, počty dokumentů a možnost použít část dat z Hodnocení Akademie věd České republiky byly důvody pro výběr výše zmíněných tří typů dokumentů.

Pokročilým dotazem do databáze Web of Science PY = (2014 OR 2015 OR 2016 OR 2017 OR 2018) s dalším omezením na výše zmíněné typy dokumentů bylo zjištěno, že výsledný počet unikátních záznamů je přibližně 11,5 milionu. Při teoretické rychlosti prémiové API 200 záznamů za sekundu mělo dojít ke stažení dat za méně než 17 hodin. Reálný stav tomu však neodpovídal.

### 3.2.2 Reálný postup

Prémiová API je omezena na sto tisíc záznamů pro jeden dotaz. Jako nejvhodnější rozdělení se tedy zdá členění na jednotlivé roky a obory pomocí logických dotazů se snahou přiblížit se

[5] Součet všech citací, které záznamy daného typu dokumentu, publikovaných v letech 2014 - 2018, obdržely v době získání dat. Údaje pochází z analytické nadstavby InCites, pro InCites Dataset + ESCI, pro území OECD, z obsahu indexovaného k 31. 5. 2020.

maximálnímu počtu záznamů na dotaz, ale nepřekročit jej. Pokud by došlo k překonání sto tisíc záznamů v dotazu, tak by nebylo možné takový dotaz spustit. Pokud by nebyly obory spojovány do dotazů tak, aby se celkový počet záznamů v dotazu blížil sto tisícům, tak by bylo zapotřebí zadávat stahování pro každý jednotlivý a sebemenší obor, což by prodloužilo dobu stahování dat.

Úspory času bylo dosaženo kombinováním oborů v rámci jednoho roku v jednom dotazu. Například obory jako Folklore, Dance a Poetry mají za rok 2018 přibližně 1301 záznamů. Je tedy vhodné nahradit tři separátní dotazy PY = 2018 AND WC = Folklore; PY = 2018 AND WC = Dance; PY = 2018 AND WC = Poetry, dotazem PY = 2018 AND WC = (Folklore OR Dance OR Poetry).

Pro stažení dat byly použity logické dotazy zadávané do API. Příkladem může být PY = (2014) AND WC = (biology) AND DT = (Article). Tento dotaz má několik vlastností: DT jakožto tag filtrující typ dokumentu není nikde na webu Web of Science zdokumentován. Byl objeven až po náhodném rozhovoru se správcem systému ALEPH Knihovny Akademie věd České Republiky, který se zmínil, že nemá ve svém systému všechny použitelné tagy zdokumentované. Na základě tohoto rozhovoru pak autor vyzkoušel několik pravděpodobných tagů a některé skutečně fungovaly. Právě nedokumentovaný tag DT umožnil výrazně zkrátit dobu stahování. Autor odhaduje, že došlo ke zkrácení přibližně o 20 až 25%. Při zadání dotazu PY = 2018 vrátí Web of Science přibližně 3,1 milionu záznamů a při následné filtraci na výše uvedené typy dokumentů vrátí Web of Science 2,43 milionu záznamů.

PY je tag určující rok publikování výstupu, tak jak je uveden u záznamů ve Web of Science. WC prohledává pole Web of Science Categories plnotextově. Nejedná se tedy o slovníkovou shodu přesně definovaných výrazů. Takový přístup vytváří problém, kdy dotaz na obor Biology: WC = Biology vrátí všechny výsledky, které mají v rámci názvu svého oboru slovo biology. Dotaz PY = 2014 AND WC = biology AND DT = (Article OR Proceedings paper OR Review) vrátí přibližně 116 tisíc výsledků, ale mezi nimi jsou zahrnuty obory Biochemistry Molecular Biology, Cell Biology, Biotechnology Applied Microbiology atd. Skutečně ryzí Biology je v takovém dotazu přítomno přibližně ve 12 tisících výsledcích. Tento jev prodlužuje dobu stahování přibližně o 10 –

15 % (subjektivní odhad autora). Tato vlastnost je ošetřena až následně ve skriptu STEAK rozložením oborů a typů dokumentů záznamů a následnou deduplikací. Tímto jevem je zatížena přibližně pětina oborů ve Web of Science, většinou v mnohem menší míře než obor Biology.

Další úspory času je dosaženo výše zmíněnou kombinací oborů do jednoho dotazu. Každý záznam může být řazen až do šesti oborů. Při stahování po každém jednotlivém oboru by za každý obor byl záznam stažen znovu. Tento jev prakticky zdvojnásobí množství stažených záznamů. Kombinace oborů do jednoho dotazu efekt vícenásobného oborového řazení alespoň trochu potlačuje. (Zde autor nemá přesnou představu. Analýza přesné úspory by byla časově náročná a pro účely této práce není důležitá).

Se zvážením uvedeného v této kapitole je možné odhadnout, že velikost celého datasetu bez jakýchkoliv úsporných opatření by byla přibližně 28 milionů záznamů za pětileté období. Je také nezbytné uvést, že deklarovaná teoretická rychlost API neodpovídá reálné rychlosti a to hned ze dvou důvodů. Zaprvé společnost Clarivate Analytics konstruovala tuto API spíše na dotazy v řádech jednotek až desítek tisíců záznamů. Zadruhé servery společnosti Clarivate jsou velmi vytěžovány z řad ostatních uživatelů (zvláště v tomto kalendářním roce, autor subjektivně zaznamenal zvýšení počtu nezpracovaných dotazů, které do jisté míry signalizují vytížení serverů), tím velké množství subdotazů[6] servery neodpoví a musí být zadány znovu. Variabilita v rychlosti API koreluje s pracovní dobou v hustě osídlených oblastech či vědecky silně rozvinutých státech. Místo teoretické rychlosti 200 záznamů za sekundu je reálná přibližně 33 - 43 záznamů za sekundu. To je bez jakýchkoliv úspor přibližně 180 - 235 hodin čistého stahovacího času.

Zadávání dotazů není možné automatizovat. Přihlašovací údaje, které umožňují API přistupovat k serveru, je nutné zadávat náhodně v intervalu 6 až 20 hodin, protože dochází k odhlašování. Přibližně dvakrát za den dojde k nespecifikované chybě ze strany serveru, která způsobí, že daný dotaz musí být zopakován a data stažena znovu. Dobu stahování tedy prodlužuje i nezbytnost ručního zadávání každého jednotlivého dotazu. Některé obory se z neznámých důvodů nestáhly celé a bylo nutné jejich stažení kontrolovat a opakovat (přibližně 50 oborů za rok).

[6] Každý dotaz požadující vice jak 100 záznamů je rozdělen na subdotazy o 100 záznamech. Teoreticky je možné požadovat 2 subdotazy za sekundu, čím by mělo dojít k obdržení 200 záznamů za sekundu.

Při zahrnutí všeho zde uvedeného a po absolvování stažení je možno uvést, že stažení citačních dat jednoho roku trvalo přibližně týden. Data byla stažena: pro roky 2014 v rozmezí 24. - 29. 2.; 2015 v rozmezí 1. - 6. 3.; 2016 v rozmezí 6. 3. - 11. 3.; 2017 v rozmezí 11. 3. - 18. 3.; 2018 v rozmezí 18. 3. - 22. 3. Databáze Web of Science je sice aktualizována každý den a mohlo by se zdát, že v rámci jednoho roku jsou data stažená v pondělí nekonzistentní vůči datům staženým v neděli, ale není tomu tak. Denní aktualizace mění data stejně přirozeně a náhodně, jak jsou přirozeně a náhodně data již v databázi obsažená. Právě inkrementální způsob budování databáze (indexace nových záznamů, vznik citačních vazeb) je to, co vytváří podobu databáze. Také týden je příliš krátká doba než aby došlo k nějakému statisticky významnému zkreslení. Takové tvrzení by však nemuselo platit při rozdílu jednoho roku, což bude dále ověřováno.

### 3.2.3 Zpracování dat

Celkem bylo staženo přibližně 28 milionu záznamů. Ty bylo nutné deduplikovat a upravit do zpracovatelné formy pomocí skriptu STEAK, který si autor vytvořil v programovacím jazyce Python. Tento skript také umí řešit oborové řazení záznamů a také jejich řazení do typů dokumentů. Skript je také optimalizován přesně na data, která jsou výstupem API.

Stažené, deduplikované a seřezané výstupy byly kontrolovány vůči databázi Web of Science. (Funkce Analyze results/Web of Science Categories u nalezených výsledků ve webovém rozhraní databáze.) V případě, že nesouhlasil počet stažených záznamů s počtem záznamů v online databázi o větší množství záznamů, tak byl konkrétní obor stažen separátně znovu. Kontrola byla prováděna vždy poměrně vůči velikosti oboru. U velkého oboru s padesáti tisíci záznamy byl rozdíl deseti prací tolerován z důvodu „živosti“ a proměnlivosti databáze. U oboru s padesáti záznamy nebyl tolerován ani rozdíl jednoho záznamu. Z neznámého důvodu server Web of Science občas neposkytl všechny záznamy některých oborů. Například obor Womens Science byl za stejných podmínek stažen až na třetí pokus. Tímto způsobem muselo být opětovně staženo dalších 3 milionu záznamů (od 4. 4. do 12. 4. 2020).

Data z Web od Science by měla obsahovat pouze takové záznamy, které mají maximálně dva typy dokumentů současně a maximálně šest oborů. Běžné jsou záznamy typu Article a Proceedings paper zároveň. Zároveň by výstup neměl být zařazen do více jak šesti oborů. Pokud by oborů vyžadoval více, tak by měl být zařazen do oboru Multidisciplinary Sciences. Ve stažených datech se však vyskytovaly záznamy se třemi typy dokumentů (k Article a Proceedings paper

přibyl navíc Retracted). Také se ve stažených datech vyskytlo několik desítek záznamů, které měly neuvěřitelných jedenáct oborů (zde se nejspíš jen jednalo o druh dočasné chyby). Těchto několik desítek záznamů bylo zpracováno ručně.

Pro určité analýzy bylo zapotřebí tuto násobnost záznamu víceoborovostí nebo dvojím typem dokumentu odstranit zachováním pouze jednoho záznamu. Například u analýzy 3. - agregát let a typů dokumentů, separace oborů došlo k odstranění informace o roku a typu dokumentu. Zůstaly zachovány informace UT WOS (jednoznačný identifikátor záznamu), počet citací a obor. Záznam typu WOS:123456789;Article|Proceedings paper;15;Toxicology|Biophysics by pro analýzu 3. byl rozložen na záznamy WOS:123456789;15;Toxicology a WOS:123456789;15;Biophysics.

Nastaly i takové situace, kdy byl záznam stažen na začátku týdne v jednom ze svých oborů s určitým počtem citací a na konci týdne byl totožný záznam stažen v jiném ze svých oborů, ale už s vyšším počtem citací. Tento jev byl vzhledem k živosti databáze a víceoborovosti záznamů poměrně běžný. V takovém případě byl zachován záznam s vyšším počtem citací.

Takto získané a zpracované záznamy byly prvotně roztříděny na jednotlivé roky, typy dokumentů a obory (přibližně 3840 separátních datasetů). Následně byly odstraněny necitované záznamy, jelikož takové záznamy nemají žádnou signifikantní číslici. V dalším kroku byly vytvořeny (znovu) deduplikované agregáty, podle požadavků jednotlivých analýz uvedených v kapitolách 3.3.1 až 3.3.6.

Tabulky v práci uvedené již prezentují zaokrouhlené výsledky. Během výpočtů však nebylo zaokrouhlování použito. Při vypočtu MAD z níže uvedených tabulek, by tak mohly vzniknout drobné odchylky dané zaokrouhlováním prezentovaných hodnot.

## 3.3 Popis citačních dat v kontextu Benfordova zákona

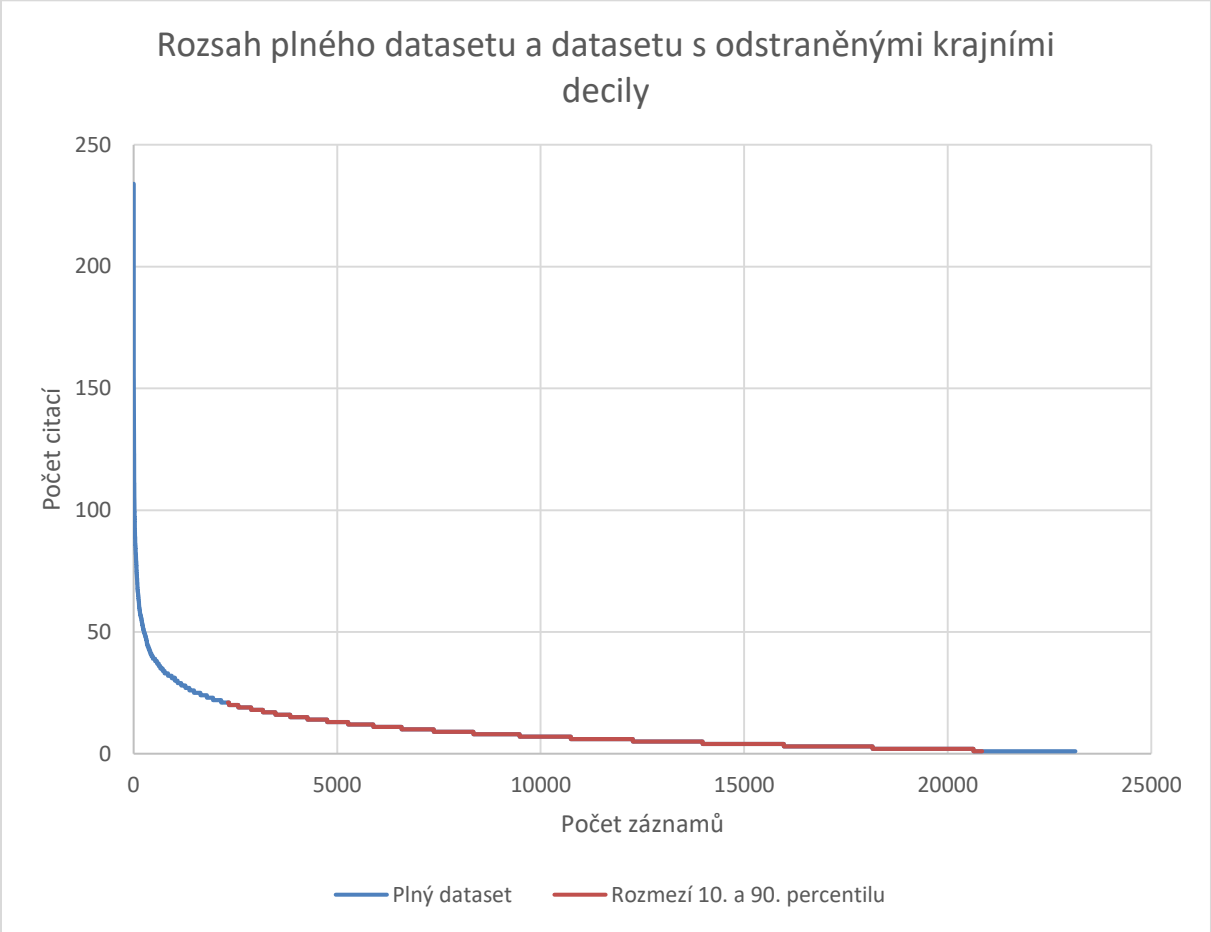
Jak bylo nastíněno v bodě 3. z kapitoly 1.2, je nezbytné, aby množství čísel v datasetech bylo dostatečně velké a zároveň jejich rozsah řádů dostatečně široký. Pro rozsah považuje Kossovsky $F_{diff} = \log(max) - \log(min)$ , $F_{diff} > 3$ za dostatečný (Kossovsky, 2015). V kontextu citačních dat je minimum v datasetu prakticky vždy jedna citace. Platí tedy, že pokud *min* = 1*,* tak $\log(1) = 0$. Z tohoto a výše uvedeného vzorce vyplývá, že nejvhodnější datasety citačních dat pro sledování Benfordova zákona dosahují maximálních hodnot tisíc citací a více.

($\log(1000) = 3$). Pro počet čísel v datasetu uvádí Nigrini více jak tisíc čísel jako optimální. Dodává, že i menší množství je možné testovat, ale je nutno počítat s většími odchylkami od Benfordova zákona (Nigrini, 2012).

Pro potřeby této diplomové práce zůstal Nigriniho požadavek zachován, ačkoliv tím ve 2. a 3. analýze datasetů (viz. kapitoly 3.3.2 a 3.3.3) došlo k odstranění jistého počtu pozorovatelných oborů. Kossovského požadavek na rozsah řádů však musel být z hlediska charakteru citačních dat snížen na $F_{diff} > 2$. Pokud by tomu tak nebylo a ve 2. analýze datasetů by byly odstraněny všechny obory, které nemají maximum větší jak tisíc citací, tak by například pro analýzu 3. agregát typů dokumentů a let bylo možné analyzovat pouze 89 z 254 oborů. Snížením Kossovského požadavku na $F_{diff} > 2$ je ve stejné analýze možné sledovat 236 z 254 oborů.

Dalším Kossovského doporučením je neuvažovat rozsah maximální a minimální hodnoty datasetu, ale pracovat s hodnotami 90. a 10. percentilu pro $F_{diff}$ (Kossovsky, 2015, s. 33). Tím by byl potlačen vliv odlehlých hodnot. U citačních dat k tomuto doporučení nelze přistoupit. Při využití dat určených pro hodnocení Akademie věd České republiky (v tomto hodnocení je krom jiných sledována hodnota 1. decilu, tedy počtu citací potřebných pro zařazení do 1. decilu, ten zde obsahuje 10% nejcitovanějších výstupů) bylo zjištěno, že v roce 2015 by pouze osm oborů (u všech oborů byly typu dokumentu Review, zároveň obsahovaly více jak 1000 záznamů) splnilo $F_{diff} = \log(10.precentil) - \log(90.percentil)$, $F_{diff} > 2$ (Web of Science, 2020).

Pro ilustraci je zde uveden Graf 4., který modrou křivkou zobrazuje celou citační křivku oboru Chemistry Analytical, typu dokumentu Article a roku publikování 2017. Počet citovaných dokumentů je 23 136 (necitované nejsou zobrazeny). Nejcitovanější práce byla citována 234 krát. Červená křivka zobrazuje tytéž data, jen s odstraněnými výstupy nad 10. percentilem a pod 90. percentilem. Tato ilustrace zjevně zešikmené distribuce citační křivky poukazuje na problematičnost „vynechání“ části dat dle Kossovského. Citační křivky oborů jsou totiž velice podobné.




*Rozsah plného datasetu a datasetu s odstraněnými krajními decily. Graf 4.*

### 3.3.1 Data pro analýzu 1. – globální agregát

V této analýze byly všechny záznamy za roky 2014 – 2018, všechny typy dokumentů a všechny obory vloženy do jednoho datasetu. V rámci něj proběhla deduplikace, aby se žádný zde přítomný záznam nevyskytoval dva či více krát. Duplicity jsou způsobeny tím, že jeden záznam může mít přiřazeny až dva typy dokumentů (typická je kombinace Article a Proceedings paper) a může být zařazen až do šesti oborů. Na tomto datasetu bylo možné stanovit míru shody s Benfordovým zákonem v celé agregaci populací dat.

### 3.3.2 Data pro analýzu 2. – maximální separace

Oproti předchozí analýze zde došlo k maximálnímu možnému rozdělení dat na jednotlivé roky, typy dokumentů a obory. Zjišťována byla míra shody s Benfordovým zákonem pro každý jednotlivý dataset, který splnil výše uvedené podmínky počtů záznamů a maximálních hodnot citací. I přesto, že došlo k rozdělení stažených dat na části, tak se i v těchto nejmenších jednotkách jedná o celé populace. Například pracuje se všemi články (Article) z oboru Biology, které indexuje databáze Web of Science za rok 2014. Jedná se tedy o celou populaci dané databáze. Jakékoliv agregáty jsou pak kombinacemi těchto populací.

### 3.3.3 Data pro analýzu 3. – agregát let a typů dokumentů, separace oborů

Analýza 3. poskytla datasety agregovaných let a typů dokumentů se zachovanou separací oborů. Tím bylo umožněno stanovení míry Benfordova zákona mezi obory s mnohem robustnější datovou základnou než v analýze 2. V rámci oborů proběhla deduplikace typů dokumentů.

### 3.3.4 Data pro analýzu 4. – agregát typů dokumentů a oborů, separace let

V této analýze bylo možné porovnat míru shody vyhovění Benfordovu zákonu mezi jednotlivými roky, při maximálním možném zachování ostatních podmínek. Na takto robustních datasetech je očekáván zanedbatelný vliv rozdílů velikostí datasetů. Za rok 2014 dataset obsahuje přibližně 1,59 milionu záznamů, rok 2015 1,76 milionu záznamů, za rok 2016 1,78 milionu záznamů, za rok 2017 1,74 milionu záznamů a rok 2018 obsahuje přibližně 1,57 milionu záznamů. Tendence počtu záznamů není lineárně rostoucí, je třeba poznamenat, že jsou odstraněny necitované práce. Počty veškerých záznamů jsou v milionech po letech přibližně takovéto: 1,98; 2,28; 2,38; 2,47 a 2,47. (Web of Science, 2020). Datasety jsou velmi robustní i v rámci šíře řádů a dokonce je splněn Kossovského požadavek na $F_{diff} > 3$. Proběhla deduplikace v rámci let na typy dokumentů a zároveň na obory.

### 3.3.5 Data pro analýzu 5. – agregát let a oborů, separace typů dokumentů

Tato analýza má za úkol zobrazit rozdíly v míře sledování Benfordova zákona mezi typy dokumentů. Opět je splněn Kossovského požadavek na šíři řádů $F_{diff} > 3$. Zde je výrazně větší rozdíl v počtu záznamů jednotlivých typů dokumentů, ale dle Nigriniho požadavku na počty záznamů by stále nemělo docházet ke zkreslení. Proběhla deduplikace multiplicitních víceoborových záznamů.

### 3.3.6 Data pro analýzu 6. - agregát oborů, separace let a typů dokumentů

Tato analýza poskytuje podrobnější pohled na analýzy 4. a 5. Slouží také jako kontrola pro absolutní počty analýzy 5. Pro analýzu 4. jako kontrola součtu absolutních počtů sloužit nemůže z důvodu duplicitních záznamů dvojího typu dokumentu. Proběhla deduplikace víceoborových záznamů.

# 4. Výběr metody analýzy

Existuje několik teorií popisujících požadavky na vznik a vlastnosti datasetu vyhovujícího Benfordovu zákonu (Kossovsky, 2017). Autor však neobjevil žádný obecně přijímaný a široce platný konsensus ve vědecké obci. Pro účely této práce je nutno nebrat v úvahu příčinnou souvislost mezi pozorovanými hodnotami a teoretickými hodnotami dle Benfordova zákona. Žádná taková kauzalita zde není hledána, a pokud existuje, tak na ni není brán zřetel.

Sledována je tedy korelace, do jaké míry vyhovují získaná data Benfordovu zákonu a jak se míra tohoto vyhovění za určitých podmínek mění. K tomuto účelu je nejvhodnější použít statistické metody. Pro zodpovězení otázek hloubkovou rešerší nebylo v prohledávaných zdrojích dostatečné množství vědeckých výstupů, které by toto téma dostatečně zpracovaly. V oblasti scientometrie bylo sice publikováno několik článků, ale ty pracují pouze s malým množstvím scientometrických indikátorů.

Pro tuto práci byla jako zdroj zdrojů také využita webová stránka benfordonline.net s rozsáhlým, avšak neúplným seznamem literatury týkající se Benfordova zákona (Berger, 2019). Zde se z 1399 uvedených článků nachází pouze čtyři relevantní: (Alves, 2016), (Alves, 2014) které se zabývají čistě časopiseckými indikátory, (Campanario, 2011) pracující pouze s jediným statistickým testem a na velmi malém vzorku, jehož výběr není nijak popsán, (Tseng, 2017) který taktéž pracuje pouze s časopiseckými indikátory. (Nalezené výsledky se vztahují k době poslední provedené rešerše v říjnu 2019). Několik dalších článků sice v názvech obsahuje slovo „bibliometrics“, ale jedná se o soupisy literatury zabývající se daným zákonem, pro tuto práci nerelevantní.

Dále byla prohledána databáze Web of Science Core Collection bez časového omezení na string TS = ((benford OR newcomb OR benford-newcomb OR newcomb-benford) AND (biblio* OR Sciento*)) s velmi podobnými výsledky jako u benfordonline.net. Nakonec byl prohledán Google Books a Amazon kvůli monografiím, z nichž budou použity tři (Kossovky, 2015), (Kossovky, 2017) a (Nigrini, 2012) a to zejména kvůli podrobnému popisu některých aspektů Benfordova zákona a podrobně popsaným statistickým metodám, které jsou přenositelné i na jiné obory, než na které jsou monografie orientovány.

Pro sledování korelace citačních dat s Benfordovým zákonem bylo použito pouze výskytů prvních číslic. Bylo by vhodné a statisticky silnější použít i výskyty prvních dvojic a posledních dvojic. To však vzhledem k charakteru dat a jejich „chudosti“ na číslice není možné.

Jedná se pouze o přibližné srovnání, ale dle rychlé analýzy ve Web of Science InCites je možné říct, že za roky 2014-2018 jsou typy dokumentů Article, Proceedings paper a Review necitovány v 27,53% případů. Dle hodnot mediánů citací pro obory převzatých z Hodnocení Akademie věd České republiky 2015-2019 je možné říct, že většina oborů má hodnotu mediánu nižší než 10 citací (Web of Science, 2020), takže se objevují pouze jednomístné číslice ve více než polovině celého datasetu.[7]

Pro potřeby testu na první dvě číslice a poslední dvě číslice je navíc zapotřebí, aby tyto číslice nebyly shodné, takže jsou od datasetu vyžadována alespoň trojmístná čísla, v ideálním případě pětimístná. (Kossovsky, 2015, s. 110). Tento požadavek na počet číslic je v rámci získaného datasetu neslučitelný s požadavkem na dostatečný počet čísel. Z tohoto důvodu jsou testy na první a poslední dvojici vynechány. Zjištění míry vyhovění Benfordovu zákonu se tak stává více indikativní a méně vhodné například pro forenzní analýzu.

## 4.1 Zvolené statistiky

### 4.1.1 MAD

Popisná statistika Mean Absolute Deviation neboli průměrná absolutní odchylka[8] zkoumá průměrnou absolutní odchylku mezi relativní četností prvních signifikantních číslic v datasetech a pravděpodobnostmi výskytu prvních signifikantních číslic dle Benfordova zákona. V běžném použití je takto možné sledovat rozptýlenost dat v datasetu od průměru v datasetu. V kontextu zjištění míry vyhovění Benfordovu zákonu je zkoumána průměrná absolutní odchylka od ideálu Benfordova zákona, nikoliv od průměru pozorovaných hodnot. Vzorec je následující:

$$MAD = \frac{\sum_{d=1}^{K} |AP_d - EP_d|}{K}$$

[7] Efekty tohoto jevu by bylo vhodné sledovat v další práci, například disertační.
[8] V anglofonním světě také Mean Absolute Error, Average Absolute Deviation. V českém jazyce Průměrná absolutní odchylka či Průměrná absolutní chyba.

Kde $AP_d$ je relativní četnost pozorované číslice, $EP_d$ je pravděpodobnost očekávané číslice dle Benfordova zákona, *d* je sledovaná číslice a *K* je počet sledovaných číslic.

Ze vzorce je snadno patrné, že pomocí dílčích hodnot $MAD_d = |AP_d - EP_d|$ lze sledovat nikoliv odchylku všech číslic, ale každé jednotlivé číslice, což může při další analytické práci poskytnout lepší vhled do zkoumaného datasetu. Pro účely této práce však nebude přistupováno k takto jemné granularitě. Ve prospěch použití MAD pro všechny číslice hovoří Nigrini, který dokonce stanovil jisté hranice (dále Nigriniho intervaly), ve kterých se je možné stanovit míra vyhovění Benfordovu zákonu. Zároveň varuje, že zjištěné intervaly jsou empiricky zjištěné z jím zkoumaných dat, do jisté míry subjektivní a nejedná se o robustní statistické testy (Nigrini, 2012, s. 158). Tuto popisnou statistiku také doporučují Hindls a Hronová jakožto vhodný popis pro datasety, které nemohou být považovány za vzorek (Hindls, 2015).

Hodnoty Nigriniho intervalů jsou:

- Blízká shoda: 0,000 až 0,006
- Přijatelná shoda: 0,006 až 0,012
- Marginálně přijatelná shoda: 0,012 až 0,015
- Bez shody: více než 0,015

Jednotky jsou bezrozměrné. Intervaly jsou Nigrinim stanoveny empiricky z 25 datasetů z různých oblastí a různých velikostí. Tyto datasety byly testovány pomocí Chí-kvadrát testů a Kolmogorov-Smirnoff testů a byly tak stanoveny hranice pro MAD. (Nigrini, 2012, s. 160)

Výhodou této popisné statistky je její nezávislost na velikosti datasetu. Vlastností Benfordova zákona je to, že výsledná pravděpodobnost výskytu první signifikantní číslice je dána logaritmem zlomku, což je iracionální číslo, které nemůže být vyjádřeno podílem dvou celých čísel (Kossovsky, 2015, s. 115). Při zvyšujícím se počtu čísel v datasetu může podíl celých čísel konvergovat k danému iracionálnímu číslu (výsledku logaritmu podílu dvou celých čísel). Z toho vyplývá, že u velkých datasetů je snazší dosáhnout konformity k Benfordovu zákonu než u datasetů malých. Tato popisná statistika tedy může nacházet lepší shody u velkých datasetů. Pro ošetření vlivu malých datasetů byla zachována Nigriniho podmínka přítomnosti alespoň tisíce pozorování (uvedeno na začátku kapitoly 3.3). Tím jsou vyřazeny takové obory, které by mohly svým nízkým počtem záznamů zkreslovat tento popis.

## 4.2 Další statistiky

V této kapitole se nachází zdůvodnění, proč nebyly pro analýzu použity jiné statistické metody, které doporučuje či používá literatura zabývající se aplikací Benfordova zákon a vzhledem k tomu, že tato práce není čistě statistického charakteru, tak budou popsány jen takové statistiky, které se nejčastěji objevovaly ve zdrojové literatuře.

### 4.2.1 Chí-kvadrát

Tento často používaný statistický test lze pozorovat u článků, které se zabývají Benfordovým zákonem ve scientometrii. Jedná se o takzvaný Test dobré shody poprvé publikovaný Karlem Pearsonem roku 1900 (Pearson, 2012).

Vzorec pro výpočet je následovný:

$$\chi^2{}_{(n-1)} = \sum_{i=1}^{n} \left(\frac{(N_{Obs} - N_{Ben})^2}{N_{Ben}}\right)$$

kde *n-1* značí počet stupňů volnosti (v dekadickém logaritmu pro první signifikantní číslic lze sledovat devět prvků, tj. *n* = 9), $N_{Obs}$ udává počet pozorování dané signifikantní číslice v datasetu, $N_{Ben}$ udává počet očekávaných absolutních četností prvních signifikantních číslic, které jsou dány pravděpodobností výskytu dle Benfordova zákona ve vztahu k velikosti sledovaného datasetu. (Pearson, 2012)

Výsledná hodnota je pak porovnána s tabelovanými hodnotami. Pro devět signifikantních číslic platí $\chi^2{}_{(n-1)}$, *n* = 9 tedy 9 -1 = 8 stupňů volnosti. Dále pak pro hladinu významnosti 5% platí, že kritická hodnota je rovna 15,507 – výsledek z tabelovaných hodnot. Přijetí (nevyvrácení) nulové hypotézy je možno pouze v případě, že výsledná hodnota $\chi^2$ je nižší než kritická hodnota. Nulová hypotéza je standardně definována jako vyhovění sledovaných dat Benfordovu zákonu při dané hladině významnosti. Pokud je výsledná hodnota $\chi^2$ vyšší než tabelovaná pro danou hladinu spolehlivosti a stupně volnosti, tak je nulová hypotéza vyvrácena a předpokládá se, že data nesledují Benfordův zákon. (Mir, 2018)

Tento Test dobré shody má však několik podstatných vlastností, které neumožňují jeho použití v této práci. Práce s hladinou významnosti naznačuje, že Chí-kvadrát není určen pro práci s celou populací, ale předpokládá vzorek z populace. Navíc v případě Benfordova zákona očekává, že populace je logaritmického rozložení a dokonale ji sleduje, což není vždy splněno.

Také je v testu Chí-kvadrát zakomponována vlastnost zohlednění počtu pozorování. Se zvyšujícím se počtem pozorování se stává test vysoce citlivým, což je v pořádku. Myšlenka, která stojí za tímto testem, se silně dotýká Zákonu velkých čísel a je tedy logické, že u vysokého počtu pozorování budou hodnoty výskytu pozorované číslice konvergovat k těm předpokládaným (teoretickým). Toto se však u citačních dat neděje a při aplikaci tohoto testu na tyto datasety, které obsahující přibližně více jak deset tisíc záznamů už dochází ve většině případů k zamítnutí nulové hypotézy. (Kossovsky, 2015)

Z výše uvedeného vyplývá, že pro účely této práce je test Chí-kvadrát nevhodný. V této práci se pracuje s datasetem o až 8,46 milionech záznamů a i v jednotlivých analýzách se jedná o celé populace, nikoliv o náhodně zvolené vzorky.

Je možné namítat, že obsah použité databáze a jejích indexů není celosvětová populace všech vědecký výstupů a dokonalá citační síť. Taková námitka by však narazila na neexistenci přesných hranic definice vědeckého výstupu a „všeho ostatního“ co také obsahuje citace. Není možné vytvořit klasifikační systém, který by jednoznačně a spolehlivě určil, který dokument lze považovat za vědecký výstup a který nikoliv. Celosvětová populace tedy není zachycena, uvažuje se tedy o obsahu databází jako celých populacích.

Pokud jsou zohledněna kritéria výběru vědeckých výstupů do databáze Web of Science, tak je nasnadě, že za těchto podmínek výběru je obsah databáze roven celé populaci databáze. I v případě, že by v této úvaze byly překročeny hranice Web of Science, tak velikost této databáze a způsob výběru vědeckých výstupu je natolik signifikantní a zástupný, že pro potřeby Benfordova zákona dostatečně zastupuje hypotetickou „celosvětovou populaci“. Autor však uznává, že by bylo zajímavé provést stejnou analýzu na datech z jiných citačních databází.

### 4.2.2 Z-test

Z-Test je velmi podobný Chí-kvadrát testu. Stejně jako on je vhodný spíše pro testování vzorku z populace. Také zohledňuje počet celkových pozorování a u vysokých hodnot se stává velmi

citlivým a snadno zavrhuje pozorování, která se nechovají dle zákonu velkých čísel. Je náchylnější na chybu na chybu 1. typu než Chí-kvadrát. Tedy je větší šance, že výsledek je falešně pozitivní (při vyhovění Benfordovu zákonu se standardně pro nulovou hypotézu stanoví předpoklad, že data vyhovují Benfordovu zákonu). Nigrini uvádí příklad: U datasetu s jedním milionem pozorování je pro statistický test přijatelných 72 významných odchylek (při hladině významnosti 5 %), pro dataset se dvěma miliony pozorování 75 významných odchylek (Nigrini, 2012, s. 151). To ilustruje nevhodnost tohoto testu pro velké datasety, které jsou v této práci použity.

Zatímco Chí-kvadrátu je testována celková shoda pro všechny číslice, tak pomocí Z-Testu je určena pro každou konkrétní číslici zvlášť. Použití Z-Testu poskytuje mnohem více informací než Chí-kvadrátu. Nevyhodnocuje totiž celý dataset, ale umožní indikovat odchylku u konkrétní číslice a pak je snazší nalézt konkrétní anomálii v pozorovaných datech.

Vzorec pro výpočet hodnoty testu u konkrétní číslice vypadá takto:

$$Z_d = \frac{|AP_d - EP_d| - (\frac{1}{2N})}{\sqrt{\frac{EP_d(1 - EP_d)}{N}}}$$

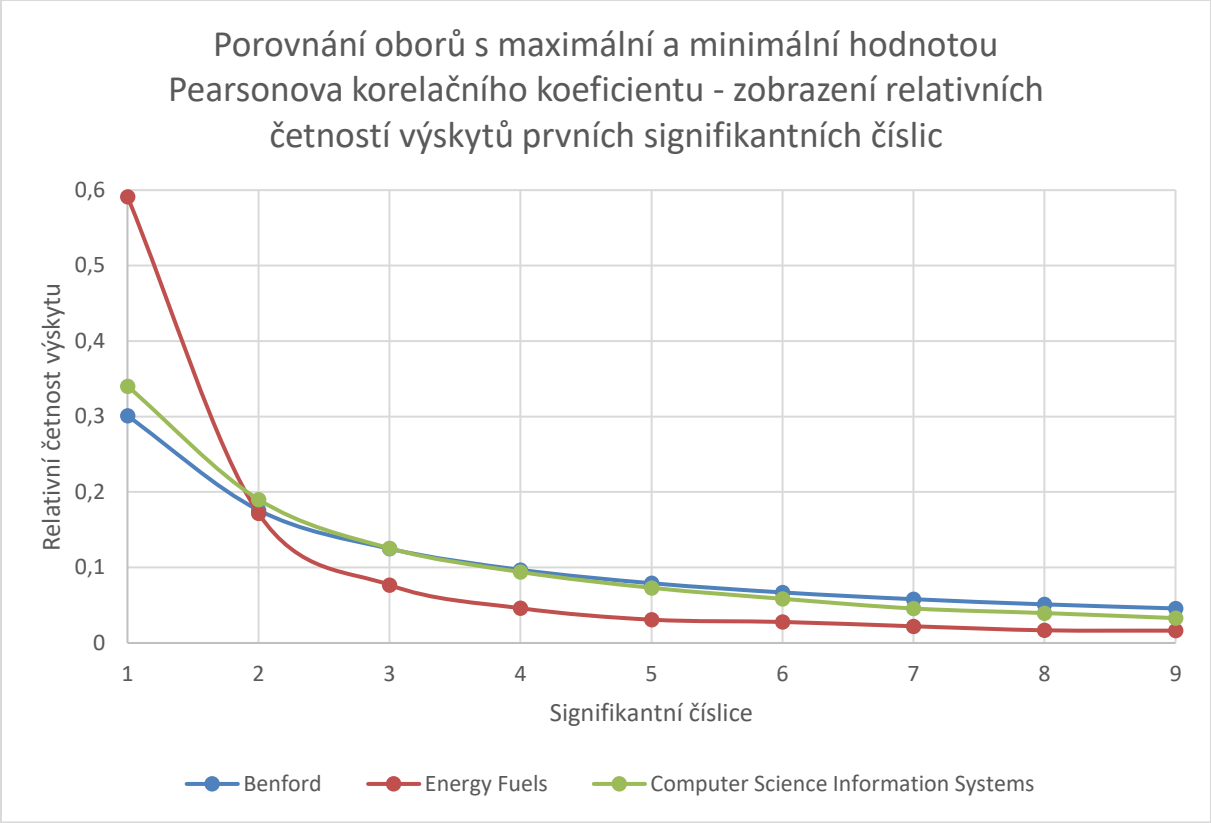
Kde *d* je sledovaná číslice, $AP_d$ je relativní hodnota pozorovaných výskytů číslice *d*, $EP_d$ je relativní hodnota očekávaných výskytů číslice dle Benfordova zákona a *N* je celkový počet pozorování. Pro 5% hladinu spolehlivosti je zde kritická hodnota rovna 1,96. Pokud by hodnota vypočteného $Z_d$ byla vyšší než kritická hodnota, tak je nulová hypotéza vyvrácena. Nulová hypotéza je stanovena stejně jako u Chí-kvadrát testu.

### 4.2.3 Pearsonův korelační koeficient

Pearsonův korelační koeficient popisuje lineární závislost dvou veličin. Rozsah hodnot tohoto koeficientu se pohybuje od -1 do 1. -1 značí nepřímou závislost, 1 zcela přímou závislost a 0 značí, že není přítomna lineární závislost. Může se však jednat o nelineární (například kvadratickou) závislost, k tomu však slouží jiné koeficienty.

Tento koeficient byl původně zvažován pro použití v této práci a testován na datech. Brzy se však ukázalo, že nepřináší dostatek informací pro smysluplné porovnání v rámci analýz. Například v analýze 2. dosahuje největší korelace s Benfordovým zákonem obor Computer Science

Information Systems, pro rok 2016 a typ dokumentu Article, a to s hodnotou 0,99994. Ve stejné analýze naopak nejmenší korelace dosahuje obor Energy Fuels, pro rok 2018, typu dokumentu Proceedings paper, a to s hodnotou 0,96187. V dané analýze se však nachází 1355 kombinací oboru, roku a typu dokumentu. Rozdíl těchto dvou krajních oborů je u Pearsonova korelačního koeficientu pouhých 3,807 procentního bodu. Následující Graf 5. ilustruje, že mezi křivkami je znatelný rozdíl. Na základě těchto zjištění byl pro tuto práci Pearsonův korelační koeficient vyhodnocen jako nevyhovující.




*Porovnání oborů s maximální a minimální hodnotou Pearsonova korelačního koeficientu - zobrazení relativních četností výskytů prvních signifikantních číslic. Graf 5.*

# 5. Výsledky analýz

## 5.1 Analýza 1. - Globální data

| Analýza 1. - globální agregát dat | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |
| Kompletní dataset absolutní počty | 2966277 | 1552706 | 1034620 | 767742 | 604964 | 492027 | 409802 | 344079 | 295016 | 8467233 | 16256 | |
| Kompletní dataset relativní počty | 0,35032 | 0,18338 | 0,12219 | 0,09067 | 0,07145 | 0,05811 | 0,04840 | 0,04064 | 0,03484 | | | 0,012574 |

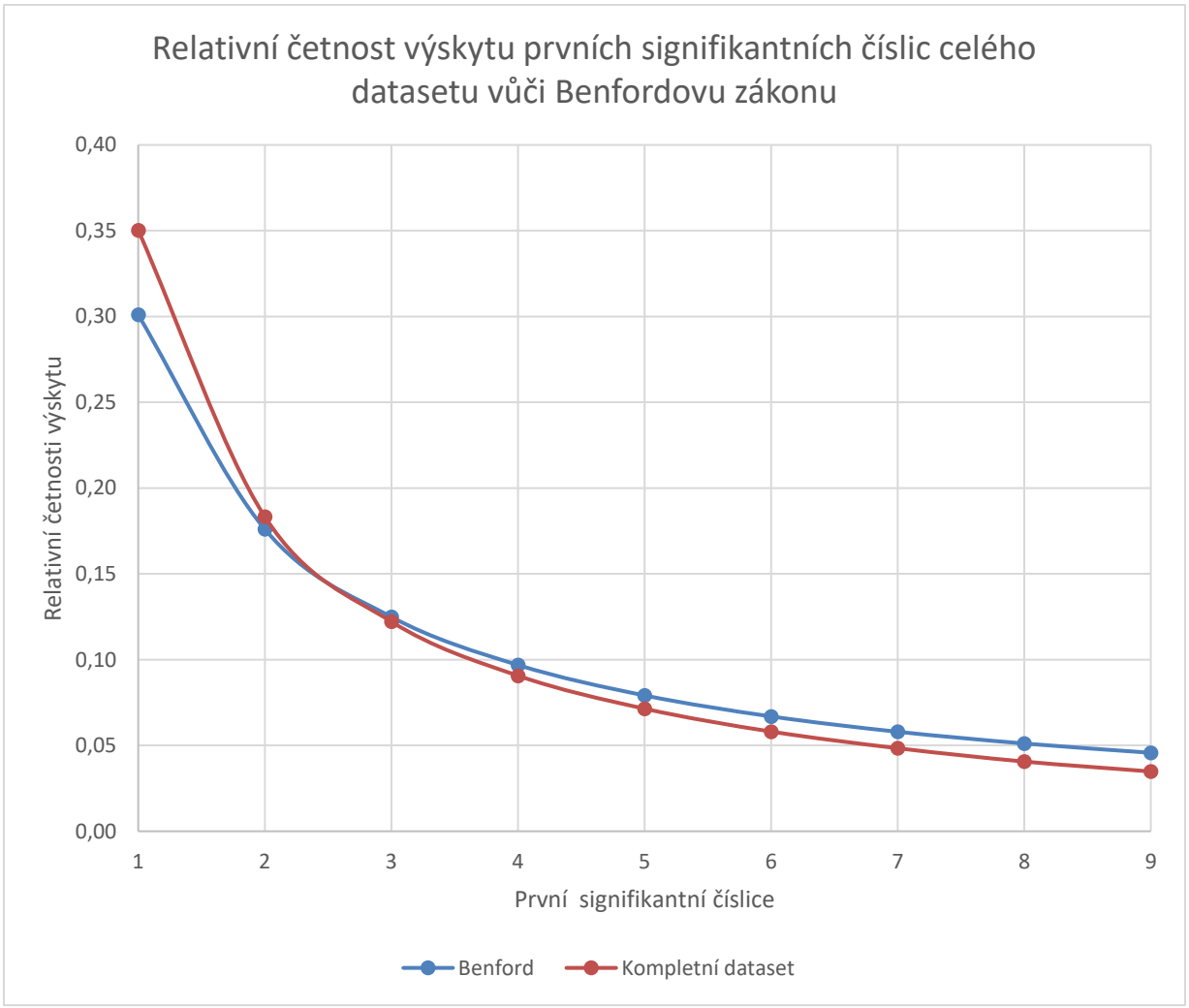
*Analýza 1. - globální agregát dat. Tabulka 3.*

V této analýze se nachází kompletní dataset let 2014 až 2018, všech 254 oborů a typů dokumentů Article, Proceedings paper a Review, stažených z databáze Web of Science. Celkem se tedy jednalo o 8,46 milionu deduplikovaných, alespoň jednou citovaných záznamů.

V Tabulce 3. jsou uvedeny hodnoty relativních četností pro jednotlivé číslice z globálních dat. Vypočtená hodnota MAD 0,0125 vůči Nigriniho intervalům značí, že se jedná o marginálně přijatelnou shodu. Pro připomenutí (sekce 4.1.1):

- Blízká shoda: 0,000 až 0,006
- Přijatelná shoda: 0,006 až 0,012
- Marginálně přijatelná shoda: 0,012 až 0,015
- Bez shody: více jako 0,015

Na Grafu 6. níže je graficky znázorněna podobnost křivek celého datasetu a Benfordova zákona.




*Relativní četnost výskytu prvních signifikantních číslic celého datasetu vůči Benfordovu zákonu. Graf 6.*

## 5.2 Analýza 2. – maximální separace

Tabulka pro tuto analýzu byla vzhledem ke své velikosti umístěna do přílohy s názvem Priloha_1_analyza_2.pdf. Obsahuje 1355 kombinací oborů, let a typů dokumentů. 2364 kombinací oborů, let a typů dokumentů bylo z této analýzy vyřazeno pro nesplnění Nigriniho požadavku na velikost datasetu nebo nesplnění modifikovaného Kossovského požadavku na rozsah maximálních a minimálních hodnot. Pro vyřazené kombinace byly zjištěny počty kombinací po jednotlivých letech:

- 2014 – 433 kombinací
- 2015 – 440 kombinací
- 2016 – 452 kombinací
- 2017 – 494 kombinací
- 2018 – 545 kombinací

Za stejných podmínek pro typy dokumentů bylo vyřazeno: Article – 303 kombinací, Proceedings paper – 1037 kombinací a Review – 1024 kombinací.

Nejnižší hodnoty MAD a tedy nejlepší shody s Benfordovým zákonem dosáhl obor Physics Applied v roce 2018 a typu dokumentu Review (MAD je rovno 0,00229). Naopak největší hodnoty MAD dosáhl obor Green Sustainable Science Technology v roce 2018 typu dokumentu Proceedings paper (MAD je rovno 0,07816).

Při bližším pohledu na oborové členění a řazení od největší po nejmenší shodu, je patrné že nejlepšího vyhovění Benfordovu zákonu dosahují takzvané „tvrdé“ vědy. První obor z kategorie humanitních a sociálních věd, s hodnotou MAD 0,00528, je Psychology Experimental v roce 2017 a typu dokumentu Article. Tento obor se nachází až na 81. pozici.

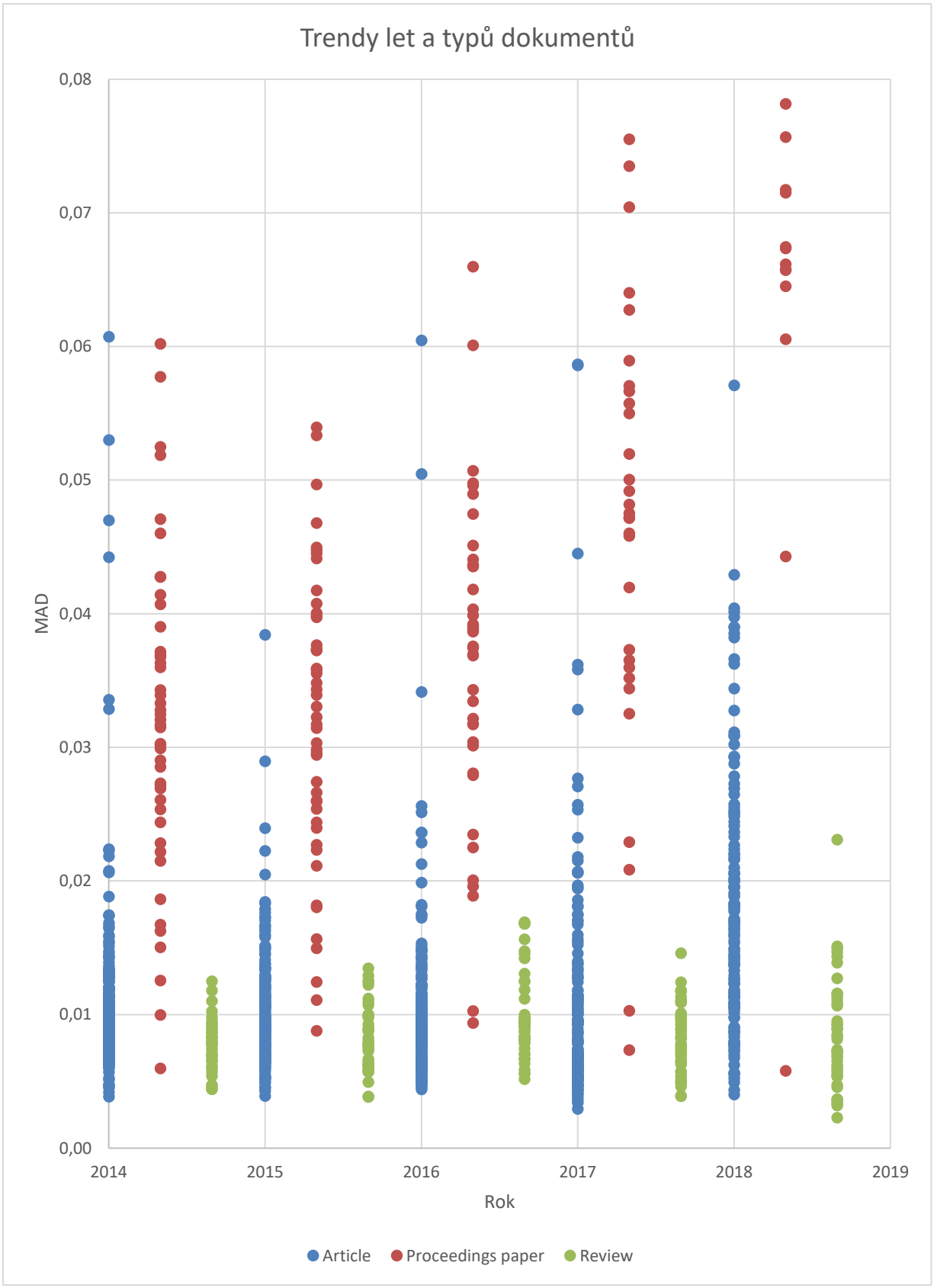
Z 1355 kombinací pouze 150 (11,07% z celkového počtu kombinací této analýzy[9]) kombinací má hodnotu MAD 0,006 a nižší. Zde se jedná o nejlepší shodu s Benfordovým zákonem a dle Nigriniho intervalů o blízkou shodu. 722 kombinací (53,28%) je hodnotou MAD v intervalu

[9] Je třeba si uvědomit, že Nigriniho intervaly nejsou rovnoměrně rozložené. Intervaly 0-0,006, 0,006-0,012, 0,012-0,015 a 0,015+ nejsou stejně velké. Procentuální hodnoty jsou tak pouze orientační.

(0,006, 0,012][10], tedy shodu přijatelnou. 146 (10,77%) kombinací se hodnotou MAD nachází v intervalu (0,012, 0,015], dosahují tedy shody marginální. 337 (24,87%) kombinací převyšuje svou hodnotou MAD horní hranici Nigriniho intervalu 0,015 a jsou tedy bez shody.

Graf 7. níže zobrazuje rozprostření oborů (přesněji kombinace oboru, roku a typu dokumentu) mezi roky a typy dokumentů. Každá modrá tečka reprezentuje konkrétní obor s typem dokumentu Article a její pozici na ose Y udává hodnota MAD daného oboru. Barva určuje typ dokumentů. Osa X reprezentuje roky vydání, pro grafickou přehlednost je k roku pro Proceedings paper přičtena hodnota 0,33 a pro Review hodnota 0,66. Tím je zajištěno, že se datové sady navzájem nepřekrývají a lze snáze sledovat.

[10] Polootevřený interval značí, že hodnota 0,006 není v intervalu možných hodnot obsažena, ale jakákoliv větší hodnota (např. 0,006000001) již intervalu náleží. Zároveň je interval omezen shora hodnotou 0,012, která interval náleží.




*Trendy let a typů dokumentů. Graf 7.*

Například trend pro obory s typem dokumentu Proceedings paper se z Grafu 7. jeví jako vzrůstající průměrná hodnota MAD s rostoucími roky. Zároveň také dochází ke snižování počtu oborů, které splňují Nigriniho a Kossovského podmínky. Počty těchto oborů předem vyřazených z analýzy pro Proceedings paper jsou v letech následující:

- 2014 – 45 oborů
- 2015 – 48 oborů
- 2016 – 41 oborů
- 2017 – 30 oborů
- 2018 – 13 oborů

Průměrná hodnota MAD pro stejný typ dokumentu:

- 2014 – 0,03187
- 2015 – 0,03202
- 2016 – 0,03661
- 2017 – 0,04593
- 2018 – 0,0619

## 5.3 Analýza 3. – agregát let a typů dokumentů, separace oborů

| Analýza 3. - agregát let a typů dokumentů, separace oborů | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Obor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | Cmax | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |
| Physics Atomic Molecular Chemical | 0,31727 | 0,17334 | 0,12090 | 0,09558 | 0,07727 | 0,06530 | 0,05783 | 0,04882 | 0,04368 | 81428 | 1541 | 0,003610 |
| Parasitology | 0,31745 | 0,17482 | 0,12297 | 0,09575 | 0,07702 | 0,06616 | 0,05638 | 0,04829 | 0,04115 | 30276 | 1033 | 0,003648 |
| Virology | 0,31896 | 0,17411 | 0,12218 | 0,09444 | 0,07625 | 0,06746 | 0,05785 | 0,04725 | 0,04150 | 31618 | 1033 | 0,004100 |
| Chemistry Physical | 0,31953 | 0,17389 | 0,12096 | 0,09389 | 0,07688 | 0,06505 | 0,05618 | 0,04919 | 0,04443 | 310608 | 3487 | 0,004111 |
| Mycology | 0,30531 | 0,16701 | 0,12900 | 0,10359 | 0,08232 | 0,06868 | 0,05887 | 0,04534 | 0,03987 | 9682 | 465 | 0,004618 |
| Cell Biology | 0,32208 | 0,17414 | 0,11986 | 0,09140 | 0,07526 | 0,06471 | 0,05749 | 0,05075 | 0,04430 | 153712 | 2875 | 0,004677 |
| Immunology | 0,32216 | 0,17140 | 0,11790 | 0,09278 | 0,07821 | 0,06640 | 0,05728 | 0,05032 | 0,04354 | 114559 | 1832 | 0,004695 |
| Hematology | 0,32041 | 0,17786 | 0,12327 | 0,09332 | 0,07812 | 0,06488 | 0,05374 | 0,04682 | 0,04159 | 50152 | 2206 | 0,004699 |
| Biochemistry Molecular Biology | 0,32322 | 0,17255 | 0,11926 | 0,09322 | 0,07849 | 0,06504 | 0,05616 | 0,04855 | 0,04352 | 274936 | 13065 | 0,004931 |
| Marine Freshwater Biology | 0,32223 | 0,17086 | 0,12244 | 0,09784 | 0,07862 | 0,06706 | 0,05400 | 0,04711 | 0,03984 | 53412 | 481 | 0,004945 |
| Physiology | 0,32252 | 0,16824 | 0,12000 | 0,09356 | 0,08026 | 0,06640 | 0,05624 | 0,04962 | 0,04316 | 53525 | 786 | 0,005016 |
| Chemistry Analytical | 0,32078 | 0,17901 | 0,12208 | 0,09350 | 0,07665 | 0,06403 | 0,05462 | 0,04755 | 0,04178 | 120275 | 803 | 0,005039 |
| Peripheral Vascular Disease | 0,32385 | 0,17579 | 0,12179 | 0,09511 | 0,07590 | 0,06533 | 0,05353 | 0,04825 | 0,04044 | 47112 | 4997 | 0,005070 |
| Infectious Diseases | 0,32389 | 0,17022 | 0,12117 | 0,09274 | 0,07840 | 0,06589 | 0,05730 | 0,04848 | 0,04192 | 72831 | 1715 | 0,005079 |
| Reproductive Biology | 0,32051 | 0,17033 | 0,12494 | 0,09228 | 0,08249 | 0,06394 | 0,05874 | 0,04778 | 0,03898 | 21347 | 855 | 0,005233 |
| Microbiology | 0,32463 | 0,17278 | 0,12224 | 0,09388 | 0,07619 | 0,06610 | 0,05474 | 0,04695 | 0,04249 | 108052 | 1551 | 0,005245 |
| Biophysics | 0,32483 | 0,17555 | 0,12206 | 0,09459 | 0,07585 | 0,06430 | 0,05393 | 0,04770 | 0,04118 | 63821 | 718 | 0,005289 |
| Toxicology | 0,32270 | 0,16752 | 0,11801 | 0,09361 | 0,07821 | 0,06858 | 0,05722 | 0,05180 | 0,04235 | 55188 | 601 | 0,005324 |
| Neurosciences | 0,32520 | 0,17528 | 0,11962 | 0,09238 | 0,07616 | 0,06427 | 0,05530 | 0,04882 | 0,04296 | 190065 | 3440 | 0,005372 |
| Endocrinology Metabolism | 0,32584 | 0,17467 | 0,11999 | 0,09196 | 0,07605 | 0,06452 | 0,05623 | 0,04839 | 0,04234 | 88171 | 3239 | 0,005514 |
| Critical Care Medicine | 0,32525 | 0,17685 | 0,12220 | 0,09640 | 0,07364 | 0,06296 | 0,05632 | 0,04640 | 0,03996 | 24999 | 1583 | 0,005551 |
| Chemistry Inorganic Nuclear | 0,32491 | 0,17728 | 0,12259 | 0,09607 | 0,07682 | 0,06415 | 0,05395 | 0,04478 | 0,03944 | 59579 | 783 | 0,005571 |
| Chemistry Organic | 0,32626 | 0,17352 | 0,12039 | 0,09436 | 0,07633 | 0,06373 | 0,05529 | 0,04753 | 0,04259 | 90649 | 790 | 0,005606 |
| Developmental Biology | 0,32495 | 0,17313 | 0,12132 | 0,09213 | 0,07412 | 0,06410 | 0,05932 | 0,04919 | 0,04173 | 19049 | 1482 | 0,005611 |
| Polymer Science | 0,32631 | 0,17598 | 0,12203 | 0,09260 | 0,07606 | 0,06390 | 0,05525 | 0,04709 | 0,04078 | 94980 | 959 | 0,005618 |
| Clinical Neurology | 0,32370 | 0,17887 | 0,12386 | 0,09496 | 0,07598 | 0,06355 | 0,05388 | 0,04587 | 0,03933 | 134235 | 3389 | 0,005654 |
| Biochemical Research Methods | 0,32526 | 0,17757 | 0,12086 | 0,09411 | 0,07682 | 0,06322 | 0,05513 | 0,04593 | 0,04110 | 82864 | 9815 | 0,005712 |
| Biotechnology Applied Microbiology | 0,32660 | 0,17629 | 0,12145 | 0,09364 | 0,07664 | 0,06414 | 0,05526 | 0,04594 | 0,04004 | 134545 | 9815 | 0,005727 |
| Behavioral Sciences | 0,32701 | 0,17137 | 0,11734 | 0,09030 | 0,07701 | 0,06592 | 0,05690 | 0,04994 | 0,04422 | 34464 | 718 | 0,005773 |
| Gastroenterology Hepatology | 0,32721 | 0,17627 | 0,12231 | 0,09427 | 0,07741 | 0,06364 | 0,05369 | 0,04510 | 0,04009 | 62456 | 1492 | 0,005856 |
| Oncology | 0,32741 | 0,17551 | 0,12072 | 0,09325 | 0,07525 | 0,06387 | 0,05473 | 0,04774 | 0,04152 | 222642 | 16256 | 0,005862 |
| Materials Science Coatings Films | 0,32749 | 0,16714 | 0,11613 | 0,09224 | 0,07891 | 0,06660 | 0,05780 | 0,04920 | 0,04450 | 36383 | 1020 | 0,005880 |
| Andrology | 0,31647 | 0,17963 | 0,11715 | 0,09100 | 0,08693 | 0,06248 | 0,05806 | 0,04856 | 0,03973 | 2945 | 195 | 0,005954 |
| Chemistry Medicinal | 0,32810 | 0,16596 | 0,11833 | 0,09347 | 0,07685 | 0,06594 | 0,05723 | 0,04943 | 0,04469 | 67270 | 1849 | 0,006015 |
| Rheumatology | 0,32822 | 0,17521 | 0,12165 | 0,09225 | 0,07434 | 0,06584 | 0,05435 | 0,04677 | 0,04136 | 24011 | 1317 | 0,006043 |
| Mineralogy | 0,32710 | 0,17726 | 0,12138 | 0,09499 | 0,07330 | 0,06246 | 0,05588 | 0,04717 | 0,04046 | 15497 | 394 | 0,006052 |
| Chemistry Multidisciplinary | 0,32467 | 0,17990 | 0,12360 | 0,09371 | 0,07516 | 0,06313 | 0,05372 | 0,04593 | 0,04018 | 325624 | 14857 | 0,006098 |
| Meteorology Atmospheric Sciences | 0,32651 | 0,17808 | 0,12056 | 0,09353 | 0,07573 | 0,06474 | 0,05467 | 0,04624 | 0,03992 | 65003 | 2751 | 0,006105 |
| Psychology Biological | 0,32859 | 0,17152 | 0,12148 | 0,09680 | 0,07848 | 0,06332 | 0,05498 | 0,04464 | 0,04018 | 8512 | 528 | 0,006126 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Psychology Developmental | 0,32575 | 0,17911 | 0,11978 | 0,09383 | 0,07648 | 0,06176 | 0,05605 | 0,04784 | 0,03940 | 25940 | 509 | 0,006164 |
| Substance Abuse | 0,32869 | 0,17186 | 0,12021 | 0,09435 | 0,07520 | 0,06725 | 0,05390 | 0,04951 | 0,03904 | 19109 | 722 | 0,006214 |
| Social Sciences Biomedical | 0,32898 | 0,17603 | 0,12536 | 0,09491 | 0,07635 | 0,06150 | 0,05346 | 0,04480 | 0,03861 | 16162 | 510 | 0,006304 |
| Public Environmental Occupational Health | 0,32691 | 0,17868 | 0,12457 | 0,09627 | 0,07842 | 0,06381 | 0,05140 | 0,04330 | 0,03664 | 148464 | 1206 | 0,006328 |
| Genetics Heredity | 0,32757 | 0,17805 | 0,12154 | 0,09462 | 0,07620 | 0,06240 | 0,05329 | 0,04571 | 0,04061 | 103650 | 13065 | 0,006334 |
| Psychology Experimental | 0,32905 | 0,17664 | 0,12143 | 0,09332 | 0,07530 | 0,06207 | 0,05567 | 0,04588 | 0,04063 | 36572 | 957 | 0,006348 |
| Physics Fluids Plasmas | 0,32517 | 0,17899 | 0,12596 | 0,09752 | 0,07848 | 0,06211 | 0,05210 | 0,04288 | 0,03680 | 44841 | 430 | 0,006371 |
| Urology Nephrology | 0,32868 | 0,17725 | 0,12551 | 0,09527 | 0,07639 | 0,06350 | 0,05239 | 0,04406 | 0,03693 | 50769 | 1249 | 0,006531 |
| Paleontology | 0,32241 | 0,16735 | 0,13042 | 0,09696 | 0,08167 | 0,06312 | 0,05432 | 0,04427 | 0,03948 | 12943 | 155 | 0,006534 |
| Psychiatry | 0,33003 | 0,17667 | 0,11974 | 0,09409 | 0,07661 | 0,06360 | 0,05446 | 0,04452 | 0,04028 | 94358 | 724 | 0,006573 |
| Transplantation | 0,33114 | 0,17346 | 0,12215 | 0,09662 | 0,07724 | 0,06066 | 0,05279 | 0,04492 | 0,04103 | 20327 | 752 | 0,006690 |
| Integrative Complementary Medicine | 0,33115 | 0,17560 | 0,11932 | 0,09273 | 0,07780 | 0,06516 | 0,05367 | 0,04522 | 0,03934 | 18354 | 182 | 0,006694 |
| Fisheries | 0,32833 | 0,17350 | 0,12376 | 0,09559 | 0,08206 | 0,06205 | 0,05210 | 0,04439 | 0,03822 | 26738 | 330 | 0,006706 |
| Medicine Research Experimental | 0,33148 | 0,17531 | 0,12396 | 0,09418 | 0,07584 | 0,06393 | 0,05191 | 0,04533 | 0,03805 | 127787 | 1641 | 0,006767 |
| Obstetrics Gynecology | 0,32797 | 0,17934 | 0,12528 | 0,09411 | 0,07732 | 0,06240 | 0,05096 | 0,04419 | 0,03843 | 56648 | 646 | 0,006785 |
| Respiratory System | 0,33108 | 0,17658 | 0,12142 | 0,09282 | 0,07439 | 0,06465 | 0,05217 | 0,04649 | 0,04040 | 47520 | 3186 | 0,006786 |
| Pharmacology Pharmacy | 0,33186 | 0,17047 | 0,11882 | 0,09326 | 0,07701 | 0,06409 | 0,05560 | 0,04723 | 0,04165 | 199489 | 1849 | 0,006852 |
| Psychology Clinical | 0,33182 | 0,17620 | 0,12217 | 0,09574 | 0,07732 | 0,06242 | 0,05277 | 0,04360 | 0,03797 | 39796 | 630 | 0,006865 |
| Audiology Speech Language Pathology | 0,31750 | 0,18747 | 0,12547 | 0,09948 | 0,07862 | 0,06119 | 0,05304 | 0,04302 | 0,03422 | 12274 | 198 | 0,006877 |
| Nutrition Dietetics | 0,33214 | 0,17574 | 0,12018 | 0,09195 | 0,07372 | 0,06389 | 0,05301 | 0,04719 | 0,04218 | 62329 | 622 | 0,006913 |
| Allergy | 0,32759 | 0,18067 | 0,11979 | 0,09176 | 0,07321 | 0,06627 | 0,05254 | 0,04625 | 0,04192 | 12238 | 638 | 0,006918 |
| Multidisciplinary Sciences | 0,33220 | 0,17429 | 0,11948 | 0,09050 | 0,07563 | 0,06298 | 0,05488 | 0,04810 | 0,04194 | 289119 | 11391 | 0,006926 |
| Tropical Medicine | 0,33227 | 0,17592 | 0,11941 | 0,08938 | 0,07805 | 0,06284 | 0,05745 | 0,04572 | 0,03895 | 20384 | 585 | 0,006942 |
| Physics Condensed Matter | 0,32659 | 0,18178 | 0,12491 | 0,09697 | 0,07621 | 0,06199 | 0,05127 | 0,04255 | 0,03774 | 142323 | 3487 | 0,006958 |
| Neuroimaging | 0,32535 | 0,18311 | 0,12280 | 0,09183 | 0,07021 | 0,06403 | 0,05596 | 0,04564 | 0,04107 | 14243 | 1013 | 0,006964 |
| Medical Laboratory Technology | 0,32468 | 0,18410 | 0,12490 | 0,09354 | 0,07689 | 0,06273 | 0,05184 | 0,04386 | 0,03747 | 14411 | 674 | 0,007035 |
| Environmental Sciences | 0,33247 | 0,17638 | 0,11921 | 0,09284 | 0,07460 | 0,06299 | 0,05353 | 0,04687 | 0,04113 | 256582 | 4476 | 0,007049 |
| Evolutionary Biology | 0,33301 | 0,16858 | 0,11735 | 0,09005 | 0,07739 | 0,06326 | 0,05693 | 0,04874 | 0,04470 | 30651 | 13065 | 0,007106 |
| Chemistry Applied | 0,33338 | 0,17474 | 0,12018 | 0,09020 | 0,07334 | 0,06412 | 0,05435 | 0,04757 | 0,04213 | 75323 | 549 | 0,007188 |
| Geography Physical | 0,33339 | 0,17472 | 0,12140 | 0,09172 | 0,07683 | 0,06405 | 0,05268 | 0,04509 | 0,04011 | 32709 | 1199 | 0,007192 |
| Ecology | 0,33344 | 0,17459 | 0,11765 | 0,09307 | 0,07449 | 0,06318 | 0,05437 | 0,04752 | 0,04168 | 91329 | 911 | 0,007203 |
| Orthopedics | 0,33076 | 0,17896 | 0,12323 | 0,09483 | 0,07681 | 0,06314 | 0,05075 | 0,04424 | 0,03728 | 62144 | 1146 | 0,007244 |
| Cell Tissue Engineering | 0,33369 | 0,17497 | 0,11654 | 0,08951 | 0,07213 | 0,06547 | 0,05425 | 0,04884 | 0,04460 | 16054 | 902 | 0,007257 |
| Biodiversity Conservation | 0,32859 | 0,18132 | 0,12195 | 0,09568 | 0,07527 | 0,06328 | 0,05365 | 0,04306 | 0,03719 | 27101 | 641 | 0,007286 |
| Limnology | 0,32983 | 0,17193 | 0,12581 | 0,10010 | 0,07459 | 0,06099 | 0,05207 | 0,04176 | 0,04293 | 9411 | 481 | 0,007301 |
| Health Policy Services | 0,33146 | 0,17856 | 0,12335 | 0,09557 | 0,07845 | 0,06219 | 0,05190 | 0,04293 | 0,03558 | 32656 | 780 | 0,007309 |
| Cardiac Cardiovascular Systems | 0,33079 | 0,17929 | 0,12274 | 0,09431 | 0,07473 | 0,06184 | 0,05241 | 0,04419 | 0,03971 | 97307 | 5152 | 0,007323 |
| Psychology Social | 0,33137 | 0,17918 | 0,12140 | 0,09389 | 0,07668 | 0,06183 | 0,05133 | 0,04538 | 0,03893 | 19522 | 325 | 0,007429 |
| Sport Sciences | 0,33240 | 0,17823 | 0,12153 | 0,09182 | 0,07465 | 0,06346 | 0,05307 | 0,04537 | 0,03947 | 46615 | 587 | 0,007446 |
| Geriatrics Gerontology | 0,33392 | 0,17673 | 0,11955 | 0,08964 | 0,07554 | 0,06357 | 0,05300 | 0,04689 | 0,04116 | 29151 | 1191 | 0,007451 |
| Health Care Sciences Services | 0,33154 | 0,17912 | 0,12388 | 0,09594 | 0,07691 | 0,06253 | 0,05166 | 0,04293 | 0,03548 | 56482 | 828 | 0,007453 |
| Forestry | 0,32989 | 0,17814 | 0,12760 | 0,09615 | 0,07671 | 0,06221 | 0,05264 | 0,04257 | 0,03410 | 25721 | 386 | 0,007460 |
| Gerontology | 0,32994 | 0,18089 | 0,12053 | 0,09051 | 0,07955 | 0,06180 | 0,05142 | 0,04679 | 0,03857 | 15324 | 1191 | 0,007573 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Food Science Technology | 0,33581 | 0,17397 | 0,12005 | 0,09072 | 0,07495 | 0,06283 | 0,05451 | 0,04615 | 0,04101 | 118071 | 1064 | 0,007730 |
| Soil Science | 0,33575 | 0,17409 | 0,12192 | 0,09017 | 0,07935 | 0,06254 | 0,05334 | 0,04448 | 0,03836 | 22368 | 382 | 0,007753 |
| Ophthalmology | 0,33095 | 0,18028 | 0,12597 | 0,09636 | 0,07507 | 0,06289 | 0,04989 | 0,04250 | 0,03607 | 43692 | 1158 | 0,007811 |
| Otorhinolaryngology | 0,32522 | 0,18646 | 0,12634 | 0,09565 | 0,07770 | 0,06160 | 0,05055 | 0,04266 | 0,03382 | 27142 | 644 | 0,007990 |
| Psychology | 0,33565 | 0,17758 | 0,11997 | 0,09233 | 0,07400 | 0,06269 | 0,05054 | 0,04602 | 0,04121 | 34485 | 630 | 0,008026 |
| Plant Sciences | 0,33268 | 0,18059 | 0,12268 | 0,09350 | 0,07501 | 0,06120 | 0,05351 | 0,04305 | 0,03776 | 114101 | 1849 | 0,008034 |
| Materials Science Biomaterials | 0,33445 | 0,17899 | 0,11579 | 0,08757 | 0,07280 | 0,06233 | 0,05603 | 0,04934 | 0,04269 | 42801 | 593 | 0,008072 |
| Engineering Chemical | 0,33201 | 0,18152 | 0,12385 | 0,08971 | 0,07337 | 0,06037 | 0,05278 | 0,04593 | 0,04046 | 164904 | 2082 | 0,008089 |
| Electrochemistry | 0,33107 | 0,18250 | 0,12296 | 0,09186 | 0,07241 | 0,06038 | 0,05039 | 0,04610 | 0,04233 | 73604 | 1020 | 0,008100 |
| Rehabilitation | 0,33235 | 0,17754 | 0,12863 | 0,09637 | 0,07601 | 0,06188 | 0,05178 | 0,04108 | 0,03436 | 40481 | 1905 | 0,008103 |
| Geochemistry Geophysics | 0,33533 | 0,17868 | 0,12091 | 0,09147 | 0,07503 | 0,06264 | 0,05182 | 0,04574 | 0,03836 | 51483 | 515 | 0,008198 |
| Astronomy Astrophysics | 0,33416 | 0,18006 | 0,12166 | 0,09176 | 0,07432 | 0,06026 | 0,05284 | 0,04558 | 0,03935 | 102414 | 5908 | 0,008246 |
| Surgery | 0,33498 | 0,17980 | 0,12296 | 0,09498 | 0,07512 | 0,06168 | 0,05118 | 0,04260 | 0,03669 | 165796 | 3186 | 0,008369 |
| Nanoscience Nanotechnology | 0,33038 | 0,18462 | 0,12494 | 0,09300 | 0,07261 | 0,06089 | 0,05157 | 0,04364 | 0,03835 | 187631 | 3781 | 0,008419 |
| Materials Science Ceramics | 0,33525 | 0,18002 | 0,11782 | 0,09224 | 0,07455 | 0,06219 | 0,05301 | 0,04513 | 0,03979 | 28653 | 489 | 0,008477 |
| Psychology Mathematical | 0,31698 | 0,18711 | 0,13575 | 0,09748 | 0,07164 | 0,06084 | 0,04711 | 0,04645 | 0,03664 | 3057 | 957 | 0,008523 |
| Oceanography | 0,33983 | 0,17373 | 0,12036 | 0,09192 | 0,07338 | 0,06202 | 0,05424 | 0,04514 | 0,03938 | 33649 | 372 | 0,008623 |
| Family Studies | 0,33854 | 0,17741 | 0,12402 | 0,09642 | 0,07832 | 0,06176 | 0,04845 | 0,04155 | 0,03352 | 14199 | 209 | 0,008629 |
| Ergonomics | 0,33992 | 0,17546 | 0,12244 | 0,09230 | 0,07466 | 0,06453 | 0,04940 | 0,04552 | 0,03577 | 7996 | 360 | 0,008642 |
| Pediatrics | 0,33178 | 0,18135 | 0,12834 | 0,09608 | 0,07520 | 0,06145 | 0,04861 | 0,04240 | 0,03478 | 75926 | 509 | 0,008757 |
| Spectroscopy | 0,33799 | 0,17873 | 0,12055 | 0,09343 | 0,07497 | 0,06282 | 0,05141 | 0,04312 | 0,03698 | 32933 | 612 | 0,008799 |
| Anatomy Morphology | 0,32727 | 0,18034 | 0,13406 | 0,09553 | 0,07490 | 0,06253 | 0,05037 | 0,03903 | 0,03597 | 9787 | 255 | 0,008802 |
| Entomology | 0,32552 | 0,18631 | 0,12828 | 0,09858 | 0,07901 | 0,06207 | 0,04941 | 0,03921 | 0,03160 | 28920 | 438 | 0,008827 |
| Dermatology | 0,33466 | 0,18230 | 0,12236 | 0,09361 | 0,07457 | 0,06352 | 0,05051 | 0,04284 | 0,03564 | 33500 | 420 | 0,008852 |
| Water Resources | 0,33525 | 0,18171 | 0,12248 | 0,09430 | 0,07336 | 0,06278 | 0,05205 | 0,04172 | 0,03634 | 77112 | 740 | 0,008854 |
| Materials Science Paper Wood | 0,32430 | 0,18640 | 0,13133 | 0,09240 | 0,07658 | 0,06325 | 0,04959 | 0,04281 | 0,03334 | 9297 | 688 | 0,008883 |
| Geosciences Multidisciplinary | 0,33824 | 0,17891 | 0,12231 | 0,09051 | 0,07366 | 0,06226 | 0,05159 | 0,04445 | 0,03806 | 120271 | 1493 | 0,008897 |
| Agriculture Dairy Animal Science | 0,33496 | 0,17860 | 0,12798 | 0,09723 | 0,07946 | 0,06042 | 0,04951 | 0,03965 | 0,03220 | 32458 | 313 | 0,008906 |
| Materials Science Multidisciplinary | 0,33664 | 0,18088 | 0,12180 | 0,09189 | 0,07344 | 0,06131 | 0,05189 | 0,04370 | 0,03844 | 497664 | 3781 | 0,008979 |
| Anesthesiology | 0,33483 | 0,18336 | 0,12290 | 0,09490 | 0,07472 | 0,05975 | 0,04964 | 0,04482 | 0,03506 | 22644 | 608 | 0,009127 |
| Pathology | 0,33801 | 0,18028 | 0,12091 | 0,09328 | 0,07473 | 0,06221 | 0,05208 | 0,04229 | 0,03621 | 38002 | 3389 | 0,009148 |
| Radiology Nuclear Medicine Medical Imaging | 0,33775 | 0,18098 | 0,12279 | 0,09407 | 0,07359 | 0,06074 | 0,05179 | 0,04303 | 0,03526 | 104628 | 3711 | 0,009247 |
| Thermodynamics | 0,33608 | 0,18274 | 0,12161 | 0,09133 | 0,07205 | 0,06011 | 0,05246 | 0,04419 | 0,03943 | 69650 | 615 | 0,009267 |
| Dentistry Oral Surgery Medicine | 0,33718 | 0,17948 | 0,12759 | 0,09390 | 0,07623 | 0,06022 | 0,04833 | 0,04158 | 0,03550 | 46847 | 732 | 0,009375 |
| Psychology Applied | 0,34415 | 0,17617 | 0,12345 | 0,09392 | 0,07274 | 0,06011 | 0,05343 | 0,04140 | 0,03463 | 19782 | 745 | 0,009600 |
| Biology | 0,33379 | 0,18674 | 0,12435 | 0,09322 | 0,07339 | 0,05908 | 0,05009 | 0,04329 | 0,03606 | 56383 | 1761 | 0,009646 |
| Psychology Educational | 0,33582 | 0,18499 | 0,12403 | 0,09545 | 0,07392 | 0,05894 | 0,05303 | 0,04169 | 0,03214 | 12352 | 278 | 0,009708 |
| Materials Science Composites | 0,34077 | 0,17976 | 0,12690 | 0,09020 | 0,07417 | 0,05710 | 0,05064 | 0,04229 | 0,03818 | 23837 | 683 | 0,010084 |
| Environmental Studies | 0,34548 | 0,17756 | 0,11873 | 0,09123 | 0,07478 | 0,06039 | 0,05136 | 0,04287 | 0,03759 | 54611 | 1493 | 0,010205 |
| Primary Health Care | 0,33237 | 0,18871 | 0,12716 | 0,09770 | 0,07268 | 0,05998 | 0,05003 | 0,03549 | 0,03588 | 7636 | 627 | 0,010438 |
| Physics Applied | 0,34231 | 0,18248 | 0,12347 | 0,09287 | 0,07339 | 0,05984 | 0,04931 | 0,04085 | 0,03547 | 328544 | 3487 | 0,010594 |
| Physics Particles Fields | 0,33922 | 0,18669 | 0,12304 | 0,09245 | 0,07263 | 0,05678 | 0,04963 | 0,04374 | 0,03582 | 59437 | 5668 | 0,010840 |
| Zoology | 0,33024 | 0,18693 | 0,13017 | 0,10055 | 0,07552 | 0,05907 | 0,04869 | 0,03864 | 0,03018 | 53415 | 547 | 0,010873 |
| Physics Mathematical | 0,33607 | 0,18651 | 0,12904 | 0,09614 | 0,07394 | 0,06033 | 0,04789 | 0,03815 | 0,03192 | 45767 | 1081 | 0,011015 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Engineering Environmental | 0,34145 | 0,18545 | 0,12131 | 0,08877 | 0,07134 | 0,05847 | 0,05038 | 0,04336 | 0,03948 | 82481 | 850 | 0,011061 |
| Metallurgy Metallurgical Engineering | 0,34721 | 0,17975 | 0,12160 | 0,09084 | 0,07276 | 0,05995 | 0,04896 | 0,04263 | 0,03630 | 79801 | 1678 | 0,011075 |
| Transportation | 0,34483 | 0,18216 | 0,12203 | 0,09193 | 0,07303 | 0,06038 | 0,04909 | 0,04191 | 0,03463 | 20471 | 483 | 0,011082 |
| Instruments Instrumentation | 0,33894 | 0,18670 | 0,12647 | 0,09250 | 0,07359 | 0,05811 | 0,04812 | 0,04083 | 0,03473 | 89207 | 803 | 0,011123 |
| Mechanics | 0,34587 | 0,18157 | 0,12110 | 0,09108 | 0,07251 | 0,05990 | 0,05026 | 0,04137 | 0,03635 | 111342 | 457 | 0,011181 |
| Psychology Multidisciplinary | 0,34536 | 0,18099 | 0,12644 | 0,09336 | 0,07162 | 0,05940 | 0,04642 | 0,04187 | 0,03453 | 46163 | 1065 | 0,011275 |
| Crystallography | 0,34565 | 0,18223 | 0,12226 | 0,09443 | 0,07107 | 0,05866 | 0,05019 | 0,04099 | 0,03453 | 28931 | 14857 | 0,011279 |
| Geology | 0,34422 | 0,18466 | 0,12471 | 0,08961 | 0,07319 | 0,06014 | 0,04734 | 0,04153 | 0,03460 | 16013 | 236 | 0,011503 |
| Medicine Legal | 0,34251 | 0,18659 | 0,12504 | 0,09167 | 0,07303 | 0,05764 | 0,04746 | 0,04312 | 0,03294 | 9229 | 201 | 0,011572 |
| Development Studies | 0,34904 | 0,17737 | 0,12795 | 0,08873 | 0,07422 | 0,05989 | 0,04795 | 0,04143 | 0,03343 | 10887 | 268 | 0,011622 |
| Veterinary Sciences | 0,34057 | 0,18338 | 0,12866 | 0,09883 | 0,07592 | 0,05814 | 0,04614 | 0,03836 | 0,03000 | 57929 | 200 | 0,011659 |
| Materials Science Textiles | 0,34848 | 0,18158 | 0,12015 | 0,09719 | 0,07102 | 0,05831 | 0,04749 | 0,04117 | 0,03461 | 12193 | 688 | 0,011825 |
| Quantum Science Technology | 0,34282 | 0,18743 | 0,12175 | 0,09792 | 0,07592 | 0,05672 | 0,04799 | 0,03861 | 0,03084 | 9273 | 847 | 0,012030 |
| Nursing | 0,33984 | 0,18622 | 0,13078 | 0,09645 | 0,07315 | 0,05942 | 0,04746 | 0,03560 | 0,03106 | 40725 | 229 | 0,012174 |
| Mathematical Computational Biology | 0,34264 | 0,18674 | 0,12770 | 0,09217 | 0,07370 | 0,05704 | 0,04658 | 0,04052 | 0,03290 | 35466 | 9815 | 0,012227 |
| Energy Fuels | 0,34704 | 0,18685 | 0,12196 | 0,08853 | 0,06875 | 0,05728 | 0,04896 | 0,04209 | 0,03853 | 206156 | 2082 | 0,012615 |
| Emergency Medicine | 0,34496 | 0,18640 | 0,12749 | 0,09147 | 0,07573 | 0,05838 | 0,04472 | 0,03912 | 0,03172 | 16770 | 538 | 0,012621 |
| Mining Mineral Processing | 0,34707 | 0,18584 | 0,12902 | 0,09534 | 0,06729 | 0,05799 | 0,04723 | 0,03859 | 0,03163 | 13657 | 394 | 0,013305 |
| Nuclear Science Technology | 0,33932 | 0,19110 | 0,13098 | 0,09769 | 0,07517 | 0,05566 | 0,04492 | 0,03602 | 0,02913 | 38754 | 601 | 0,013361 |
| Ornithology | 0,34047 | 0,19083 | 0,12944 | 0,09451 | 0,08078 | 0,05755 | 0,04402 | 0,03332 | 0,02908 | 4952 | 167 | 0,013395 |
| Green Sustainable Science Technology | 0,35316 | 0,18446 | 0,12049 | 0,08795 | 0,06829 | 0,05691 | 0,04947 | 0,04093 | 0,03834 | 67109 | 881 | 0,013444 |
| Engineering Civil | 0,35459 | 0,18389 | 0,12010 | 0,09076 | 0,07007 | 0,05780 | 0,04916 | 0,03934 | 0,03430 | 96162 | 670 | 0,013635 |
| Microscopy | 0,34434 | 0,18404 | 0,13536 | 0,08927 | 0,07895 | 0,05435 | 0,04678 | 0,03922 | 0,02769 | 5814 | 202 | 0,013707 |
| Acoustics | 0,34941 | 0,18958 | 0,12330 | 0,09322 | 0,07245 | 0,05628 | 0,04581 | 0,03781 | 0,03214 | 29993 | 493 | 0,013749 |
| Agricultural Engineering | 0,34780 | 0,19129 | 0,12051 | 0,08366 | 0,06998 | 0,05549 | 0,05150 | 0,04159 | 0,03818 | 18778 | 804 | 0,013770 |
| Engineering Manufacturing | 0,35667 | 0,18346 | 0,11787 | 0,08791 | 0,06870 | 0,06086 | 0,04778 | 0,04092 | 0,03584 | 40962 | 683 | 0,014003 |
| Engineering Biomedical | 0,35412 | 0,18602 | 0,12127 | 0,09164 | 0,07057 | 0,05612 | 0,04883 | 0,03964 | 0,03179 | 74857 | 1169 | 0,014003 |
| Geography | 0,36228 | 0,17998 | 0,12149 | 0,08895 | 0,06892 | 0,05753 | 0,04769 | 0,03914 | 0,03402 | 25014 | 1328 | 0,014474 |
| Agronomy | 0,35794 | 0,18546 | 0,12176 | 0,08991 | 0,06997 | 0,05514 | 0,04805 | 0,03899 | 0,03280 | 46248 | 373 | 0,014728 |
| Construction Building Technology | 0,35660 | 0,18700 | 0,12047 | 0,08881 | 0,06871 | 0,05667 | 0,04754 | 0,03948 | 0,03471 | 50908 | 490 | 0,014775 |
| Social Work | 0,35045 | 0,18777 | 0,13144 | 0,09652 | 0,07115 | 0,05699 | 0,04078 | 0,03719 | 0,02772 | 13634 | 229 | 0,015020 |
| Operations Research Management Science | 0,35929 | 0,18665 | 0,12137 | 0,08919 | 0,07066 | 0,05664 | 0,04435 | 0,03779 | 0,03406 | 48234 | 567 | 0,015294 |
| Medical Informatics | 0,35474 | 0,19014 | 0,12688 | 0,08901 | 0,07176 | 0,05616 | 0,04662 | 0,03603 | 0,02868 | 20122 | 657 | 0,015487 |
| Engineering Ocean | 0,37094 | 0,17650 | 0,11596 | 0,09617 | 0,06764 | 0,05414 | 0,04925 | 0,03726 | 0,03213 | 8589 | 372 | 0,015627 |
| Medicine General Internal | 0,35657 | 0,18979 | 0,12642 | 0,09148 | 0,07023 | 0,05516 | 0,04509 | 0,03587 | 0,02939 | 127403 | 5045 | 0,015715 |
| Mathematics Interdisciplinary Applications | 0,35358 | 0,19384 | 0,12664 | 0,09308 | 0,06913 | 0,05590 | 0,04216 | 0,03624 | 0,02943 | 43597 | 957 | 0,016000 |
| Physics Nuclear | 0,35568 | 0,19373 | 0,12403 | 0,08942 | 0,07125 | 0,05405 | 0,04473 | 0,03823 | 0,02887 | 28953 | 5668 | 0,016064 |
| Materials Science Characterization Testing | 0,36210 | 0,18531 | 0,12717 | 0,08850 | 0,06924 | 0,05374 | 0,04500 | 0,03668 | 0,03227 | 14068 | 386 | 0,016116 |
| Management | 0,36719 | 0,18587 | 0,12012 | 0,08776 | 0,06798 | 0,05570 | 0,04529 | 0,03774 | 0,03236 | 68731 | 3386 | 0,016874 |
| Optics | 0,37269 | 0,18235 | 0,11883 | 0,08724 | 0,06928 | 0,05515 | 0,04518 | 0,03735 | 0,03193 | 157732 | 2874 | 0,017314 |
| Agriculture Multidisciplinary | 0,36721 | 0,18792 | 0,12110 | 0,08599 | 0,06616 | 0,05475 | 0,04642 | 0,03769 | 0,03275 | 31816 | 316 | 0,017335 |
| Engineering Geological | 0,36956 | 0,18650 | 0,12257 | 0,08995 | 0,06693 | 0,05268 | 0,04616 | 0,03631 | 0,02934 | 24981 | 605 | 0,017543 |
| Criminology Penology | 0,37111 | 0,18516 | 0,11775 | 0,09094 | 0,06914 | 0,05624 | 0,04313 | 0,03831 | 0,02822 | 14955 | 240 | 0,017588 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Physics Multidisciplinary | 0,35759 | 0,19527 | 0,12875 | 0,09150 | 0,06864 | 0,05167 | 0,04241 | 0,03430 | 0,02988 | 99293 | 3714 | 0,017678 |
| Statistics Probability | 0,35847 | 0,19404 | 0,12994 | 0,09361 | 0,06935 | 0,05157 | 0,04188 | 0,03346 | 0,02768 | 42913 | 13324 | 0,017865 |
| Hospitality Leisure Sport Tourism | 0,36915 | 0,18980 | 0,11992 | 0,08639 | 0,06706 | 0,05401 | 0,04504 | 0,03756 | 0,03108 | 20848 | 406 | 0,018184 |
| Engineering Aerospace | 0,36981 | 0,18939 | 0,12288 | 0,09321 | 0,06611 | 0,05362 | 0,04338 | 0,03463 | 0,02697 | 22176 | 559 | 0,018242 |
| Regional Urban Planning | 0,38125 | 0,17890 | 0,12062 | 0,08283 | 0,06816 | 0,05434 | 0,04516 | 0,03586 | 0,03289 | 15478 | 3386 | 0,018451 |
| Economics | 0,36714 | 0,19462 | 0,12458 | 0,08919 | 0,06821 | 0,05244 | 0,04160 | 0,03403 | 0,02819 | 105648 | 852 | 0,018809 |
| Demography | 0,37164 | 0,18359 | 0,13151 | 0,08772 | 0,06683 | 0,05838 | 0,04148 | 0,02904 | 0,02980 | 6509 | 322 | 0,018818 |
| Remote Sensing | 0,37224 | 0,19073 | 0,12062 | 0,08841 | 0,06867 | 0,05229 | 0,04193 | 0,03534 | 0,02976 | 40119 | 1199 | 0,019079 |
| Social Sciences Mathematical Methods | 0,36534 | 0,19638 | 0,12647 | 0,09201 | 0,06508 | 0,05403 | 0,03913 | 0,03684 | 0,02472 | 12216 | 782 | 0,019141 |
| Transportation Science Technology | 0,37394 | 0,18932 | 0,12322 | 0,08884 | 0,06604 | 0,05219 | 0,04333 | 0,03538 | 0,02773 | 32056 | 708 | 0,019143 |
| Information Science Library Science | 0,37252 | 0,19125 | 0,12147 | 0,08807 | 0,06773 | 0,05165 | 0,04381 | 0,03434 | 0,02917 | 24377 | 814 | 0,019255 |
| Engineering Petroleum | 0,36961 | 0,19104 | 0,12820 | 0,08879 | 0,06624 | 0,05386 | 0,04007 | 0,03514 | 0,02704 | 9134 | 213 | 0,019288 |
| Business | 0,37566 | 0,18864 | 0,11812 | 0,08718 | 0,06555 | 0,05224 | 0,04517 | 0,03674 | 0,03070 | 51627 | 3386 | 0,019373 |
| Engineering Industrial | 0,37329 | 0,19118 | 0,11974 | 0,08573 | 0,06648 | 0,05237 | 0,04423 | 0,03462 | 0,03236 | 39428 | 1286 | 0,019411 |
| Industrial Relations Labor | 0,37004 | 0,19415 | 0,12523 | 0,08371 | 0,06610 | 0,05215 | 0,04518 | 0,03803 | 0,02541 | 6021 | 852 | 0,019413 |
| Public Administration | 0,37451 | 0,18895 | 0,12611 | 0,09069 | 0,06684 | 0,05144 | 0,04192 | 0,03186 | 0,02768 | 11236 | 336 | 0,019447 |
| Urban Studies | 0,37635 | 0,18956 | 0,11702 | 0,08444 | 0,06983 | 0,05263 | 0,04390 | 0,03472 | 0,03155 | 15467 | 709 | 0,019732 |
| Sociology | 0,37867 | 0,18872 | 0,12188 | 0,08840 | 0,06589 | 0,05298 | 0,04218 | 0,03406 | 0,02721 | 24253 | 415 | 0,020061 |
| Business Finance | 0,37306 | 0,19338 | 0,12612 | 0,08939 | 0,06759 | 0,04969 | 0,04228 | 0,03315 | 0,02535 | 31011 | 749 | 0,020111 |
| Anthropology | 0,37631 | 0,19247 | 0,12332 | 0,08458 | 0,06543 | 0,05417 | 0,03977 | 0,03440 | 0,02954 | 17499 | 198 | 0,020368 |
| Engineering Mechanical | 0,38244 | 0,18704 | 0,11675 | 0,08507 | 0,06500 | 0,05287 | 0,04348 | 0,03662 | 0,03074 | 125997 | 452 | 0,020523 |
| Mathematics Applied | 0,36716 | 0,19786 | 0,12969 | 0,08934 | 0,06873 | 0,05236 | 0,03987 | 0,03090 | 0,02409 | 109283 | 620 | 0,020589 |
| Horticulture | 0,38157 | 0,18891 | 0,12115 | 0,09054 | 0,06327 | 0,05201 | 0,04096 | 0,03452 | 0,02706 | 18919 | 268 | 0,020747 |
| Engineering Marine | 0,39391 | 0,17688 | 0,11846 | 0,08533 | 0,06423 | 0,05179 | 0,04584 | 0,03340 | 0,03016 | 7395 | 141 | 0,020816 |
| Computer Science Interdisciplinary Applications | 0,37985 | 0,19125 | 0,12239 | 0,08565 | 0,06601 | 0,05155 | 0,04086 | 0,03416 | 0,02827 | 113072 | 13324 | 0,020883 |
| Medical Ethics | 0,36818 | 0,19216 | 0,13605 | 0,08583 | 0,06610 | 0,05227 | 0,04330 | 0,02818 | 0,02793 | 3903 | 114 | 0,020962 |
| Linguistics | 0,38022 | 0,19213 | 0,12332 | 0,08952 | 0,06642 | 0,05061 | 0,04174 | 0,03069 | 0,02536 | 27960 | 338 | 0,021163 |
| Imaging Science Photographic Technology | 0,38055 | 0,19199 | 0,12293 | 0,08762 | 0,06689 | 0,05041 | 0,03918 | 0,03358 | 0,02685 | 34637 | 2743 | 0,021204 |
| Ethics | 0,37974 | 0,19078 | 0,12916 | 0,08769 | 0,06221 | 0,05266 | 0,04053 | 0,02977 | 0,02745 | 11621 | 232 | 0,021694 |
| Communication | 0,38274 | 0,19508 | 0,12285 | 0,08428 | 0,06619 | 0,04983 | 0,04006 | 0,03241 | 0,02656 | 25487 | 396 | 0,022378 |
| Computer Science Software Engineering | 0,38056 | 0,19908 | 0,12647 | 0,08759 | 0,06405 | 0,04918 | 0,03832 | 0,02973 | 0,02503 | 70806 | 14679 | 0,023123 |
| Social Issues | 0,39245 | 0,19306 | 0,12049 | 0,08811 | 0,06429 | 0,05108 | 0,03722 | 0,02884 | 0,02447 | 10748 | 165 | 0,024085 |
| Engineering Multidisciplinary | 0,39156 | 0,19554 | 0,12173 | 0,08507 | 0,06236 | 0,04853 | 0,03864 | 0,03081 | 0,02576 | 81883 | 1697 | 0,024438 |
| Archaeology | 0,39117 | 0,19690 | 0,12576 | 0,08825 | 0,05978 | 0,04871 | 0,03699 | 0,02891 | 0,02352 | 13733 | 140 | 0,024837 |
| Ethnic Studies | 0,40256 | 0,18822 | 0,11888 | 0,08862 | 0,06124 | 0,05457 | 0,03314 | 0,02810 | 0,02468 | 5552 | 139 | 0,025257 |
| Computer Science Artificial Intelligence | 0,39428 | 0,19817 | 0,12348 | 0,08309 | 0,06179 | 0,04711 | 0,03655 | 0,03035 | 0,02517 | 163345 | 10408 | 0,025629 |
| Robotics | 0,39087 | 0,20046 | 0,12651 | 0,08674 | 0,05945 | 0,04749 | 0,03612 | 0,02966 | 0,02271 | 38875 | 1087 | 0,025729 |
| Computer Science Information Systems | 0,39092 | 0,20120 | 0,12625 | 0,08495 | 0,06193 | 0,04719 | 0,03500 | 0,02878 | 0,02377 | 160330 | 1700 | 0,025846 |
| Agricultural Economics Policy | 0,38900 | 0,20068 | 0,12893 | 0,08020 | 0,06396 | 0,04602 | 0,03519 | 0,03164 | 0,02437 | 5910 | 269 | 0,025901 |
| Education Scientific Disciplines | 0,39921 | 0,19515 | 0,11866 | 0,08268 | 0,06279 | 0,04793 | 0,03874 | 0,02920 | 0,02564 | 25037 | 4476 | 0,026053 |
| Computer Science Cybernetics | 0,39885 | 0,19762 | 0,12364 | 0,08284 | 0,05987 | 0,04921 | 0,03561 | 0,02850 | 0,02387 | 17721 | 487 | 0,026521 |
| Automation Control Systems | 0,40276 | 0,19407 | 0,11865 | 0,08331 | 0,06042 | 0,04628 | 0,03753 | 0,03120 | 0,02577 | 97350 | 6373 | 0,026603 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Engineering Electrical Electronic | 0,40081 | 0,19719 | 0,12175 | 0,08290 | 0,06070 | 0,04633 | 0,03683 | 0,02917 | 0,02431 | 534768 | 5029 | 0,026863 |
| Telecommunications | 0,40046 | 0,19805 | 0,12147 | 0,08290 | 0,06118 | 0,04630 | 0,03616 | 0,02955 | 0,02394 | 137848 | 3335 | 0,026975 |
| Computer Science Hardware Architecture | 0,39757 | 0,20082 | 0,12584 | 0,08511 | 0,06024 | 0,04516 | 0,03601 | 0,02687 | 0,02239 | 55648 | 14679 | 0,027149 |
| International Relations | 0,40829 | 0,19199 | 0,11651 | 0,08429 | 0,06146 | 0,04808 | 0,03645 | 0,03028 | 0,02265 | 21235 | 317 | 0,027369 |
| Political Science | 0,40499 | 0,19566 | 0,11742 | 0,08326 | 0,06220 | 0,04539 | 0,03632 | 0,03076 | 0,02400 | 39789 | 547 | 0,027449 |
| Education Educational Research | 0,40704 | 0,19512 | 0,11907 | 0,08293 | 0,06214 | 0,04730 | 0,03574 | 0,02805 | 0,02260 | 97777 | 472 | 0,027785 |
| Mathematics | 0,39099 | 0,20714 | 0,12988 | 0,08722 | 0,06191 | 0,04424 | 0,03410 | 0,02528 | 0,01924 | 106885 | 620 | 0,027988 |
| Womens Studies | 0,40965 | 0,19556 | 0,11437 | 0,08184 | 0,05804 | 0,04501 | 0,03855 | 0,03295 | 0,02401 | 9286 | 137 | 0,028465 |
| Social Sciences Interdisciplinary | 0,42077 | 0,19229 | 0,11914 | 0,07977 | 0,05730 | 0,04708 | 0,03364 | 0,02749 | 0,02251 | 42774 | 610 | 0,030208 |
| Computer Science Theory Methods | 0,41759 | 0,20541 | 0,12329 | 0,08069 | 0,05655 | 0,04147 | 0,03151 | 0,02421 | 0,01930 | 179549 | 13866 | 0,032417 |
| History Philosophy Of Science | 0,42146 | 0,20844 | 0,11955 | 0,07885 | 0,05638 | 0,04149 | 0,03154 | 0,02246 | 0,01982 | 11351 | 171 | 0,033951 |
| History Of Social Sciences | 0,39431 | 0,22192 | 0,13879 | 0,08632 | 0,05714 | 0,03531 | 0,03163 | 0,01815 | 0,01643 | 4078 | 185 | 0,033993 |
| Law | 0,44575 | 0,20145 | 0,11718 | 0,07525 | 0,05362 | 0,03930 | 0,02768 | 0,02294 | 0,01682 | 30380 | 260 | 0,037796 |
| Language Linguistics | 0,45114 | 0,20021 | 0,11736 | 0,07357 | 0,05177 | 0,03912 | 0,02969 | 0,02165 | 0,01549 | 20783 | 407 | 0,038717 |
| Area Studies | 0,46990 | 0,20687 | 0,10950 | 0,07131 | 0,04848 | 0,03353 | 0,02660 | 0,01947 | 0,01433 | 14584 | 103 | 0,044366 |
| Cultural Studies | 0,48015 | 0,21007 | 0,10618 | 0,06863 | 0,04770 | 0,03197 | 0,02410 | 0,01713 | 0,01408 | 7883 | 348 | 0,047355 |
| Philosophy | 0,48648 | 0,20658 | 0,11166 | 0,07149 | 0,04460 | 0,03146 | 0,02102 | 0,01520 | 0,01150 | 20786 | 120 | 0,047987 |
| Music | 0,52117 | 0,19785 | 0,09651 | 0,06537 | 0,04146 | 0,02281 | 0,02522 | 0,01557 | 0,01404 | 4559 | 131 | 0,053755 |
| Architecture | 0,54332 | 0,20509 | 0,09818 | 0,06136 | 0,03273 | 0,02269 | 0,01543 | 0,01227 | 0,00893 | 5378 | 280 | 0,060288 |
| History | 0,54668 | 0,20626 | 0,10160 | 0,05764 | 0,03363 | 0,02107 | 0,01509 | 0,01010 | 0,00792 | 25245 | 185 | 0,061293 |
| Art | 0,56750 | 0,19650 | 0,09344 | 0,04723 | 0,03143 | 0,02147 | 0,02027 | 0,01271 | 0,00945 | 5822 | 140 | 0,063751 |
| Religion | 0,56309 | 0,20100 | 0,09535 | 0,05224 | 0,03565 | 0,02041 | 0,01419 | 0,01005 | 0,00802 | 16214 | 157 | 0,063771 |
| Film Radio Television | 0,58789 | 0,18439 | 0,09206 | 0,05168 | 0,03338 | 0,01696 | 0,01454 | 0,00915 | 0,00996 | 3715 | 152 | 0,065590 |
| Humanities Multidisciplinary | 0,58098 | 0,19655 | 0,09481 | 0,04776 | 0,02806 | 0,02033 | 0,01371 | 0,01040 | 0,00741 | 12689 | 272 | 0,066756 |
| Literature | 0,61174 | 0,20510 | 0,07952 | 0,04058 | 0,02564 | 0,01536 | 0,01179 | 0,00617 | 0,00411 | 7294 | 216 | 0,075492 |

*Analýza 3. - agregát let a typů dokumentů, separace oborů. Tabulka 4.*

V této analýze byly agregovány roky a typy dokumentů, zachována byla separace oborů. Na základě Nigriniho a Kossovského podmínek bylo vyřazeno 18 z 254 oborů. Opět je patrné, že při seřazení oborů podle hodnoty MAD od nejmenší po největší jsou první příčky zastoupeny „tvrdými“[11] vědami. Teprve na 29. pozici se nachází první zástupce kategorie humanitních a sociálních věd obor Behavioral Sciences.

Při interpretaci výsledků této analýzy je zapotřebí zvýšené opatrnosti. Například obor umístěný na 221. pozici (16. odspodu) Computer Science Theory Methods. Na první pohled zde při srovnání hodnot MAD s Nigriniho intervaly není dosaženo shody s Benfordovým zákonem. Při

[11] „Tvrdými“ vědami jsou zde myšleny přibližně přírodní vědy s vysokou citační intenzitou v oboru. Často jsou popisovány jako vysoce exaktní s aplikovanými výstupy. Autor v této práci odmítá použít jakoukoliv definici s přesným dělením oborů, jelikož vědecké disciplíny považuje spíše za škálu s neurčitě definovanými pozicemi oborů. Představa o „tvrdosti” věd je tedy ponechána na čtenáři.

konzultaci s podrobnějšími daty z analýzy 2. je patrný vliv vzrůstajících let na zvyšující se hodnotu MAD a shodně pro typy dokumentů Article a Proceedings paper. Review se v analýze 2. nevyskytuje z důvodu nízkého počtu dokumentů. Jsou sice intenzivně citovány, je jich však pouze 530 za sledované pětileté období. Dále v tomto oboru typ dokumentů Article dosahuje lepší shody s Benfordovým zákonem než Proceedings paper. Vliv poměrného zastoupení typů dokumentů v rámci oboru na celkovou vlastnost oboru sledovat Benfordův zákon bude ilustrován v kapitolách 5.4, 5.5 a 5.6)

Z 236 oborů této analýzy 33 oborů (13,98%) dle MAD a Nigriniho intervalů dosahuje blízké shody s Benfordovým zákonem, 105 oborů (44,49%) dosahuje přijatelné shody, 18 oborů (7,63%) marginální shody. Konečně 80 oborů (33,89%) je vyhodnoceno bez shody.

## 5.4 Analýza 4. – agregát typů dokumentů a oborů, separace let

| Analýza 4. - agregát typů dokumentů a oborů, separace let | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Absolutní počty | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ |
| 2014 | 549609 | 292454 | 192013 | 141317 | 112537 | 94044 | 80777 | 70200 | 62599 | 1595550 | 9815 |
| 2015 | 615261 | 320125 | 209638 | 155768 | 125648 | 104411 | 89340 | 76908 | 67589 | 1764688 | 16256 |
| 2016 | 623813 | 317919 | 212179 | 159998 | 128611 | 106756 | 90724 | 76976 | 66199 | 1783175 | 13065 |
| 2017 | 609743 | 313589 | 213832 | 162095 | 128094 | 104142 | 85862 | 70474 | 59479 | 1747310 | 14679 |
| 2018 | 567851 | 308619 | 206958 | 148564 | 110074 | 82674 | 63099 | 49521 | 39150 | 1576510 | 6706 |

*Analýza 4. - agregát typů dokumentů a oborů, separace let – absolutní počty. Tabulka 5.*

| Analýza 4. - agregát typů dokumentů a oborů, separace let | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Relativní počty | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | |
| 2014 | 0,34446 | 0,18329 | 0,12034 | 0,08857 | 0,07053 | 0,05894 | 0,05063 | 0,04400 | 0,03923 | 0,011252 |
| 2015 | 0,34865 | 0,18141 | 0,11880 | 0,08827 | 0,07120 | 0,05917 | 0,05063 | 0,04358 | 0,03830 | 0,011764 |
| 2016 | 0,34983 | 0,17829 | 0,11899 | 0,08973 | 0,07212 | 0,05987 | 0,05088 | 0,04317 | 0,03712 | 0,011333 |
| 2017 | 0,34896 | 0,17947 | 0,12238 | 0,09277 | 0,07331 | 0,05960 | 0,04914 | 0,04033 | 0,03404 | 0,011402 |
| 2018 | 0,36019 | 0,19576 | 0,13128 | 0,09424 | 0,06982 | 0,05244 | 0,04002 | 0,03141 | 0,02483 | 0,018927 |

*Analýza 4. - agregát typů dokumentů a oborů, separace let – relativní počty. Tabulka 6.*

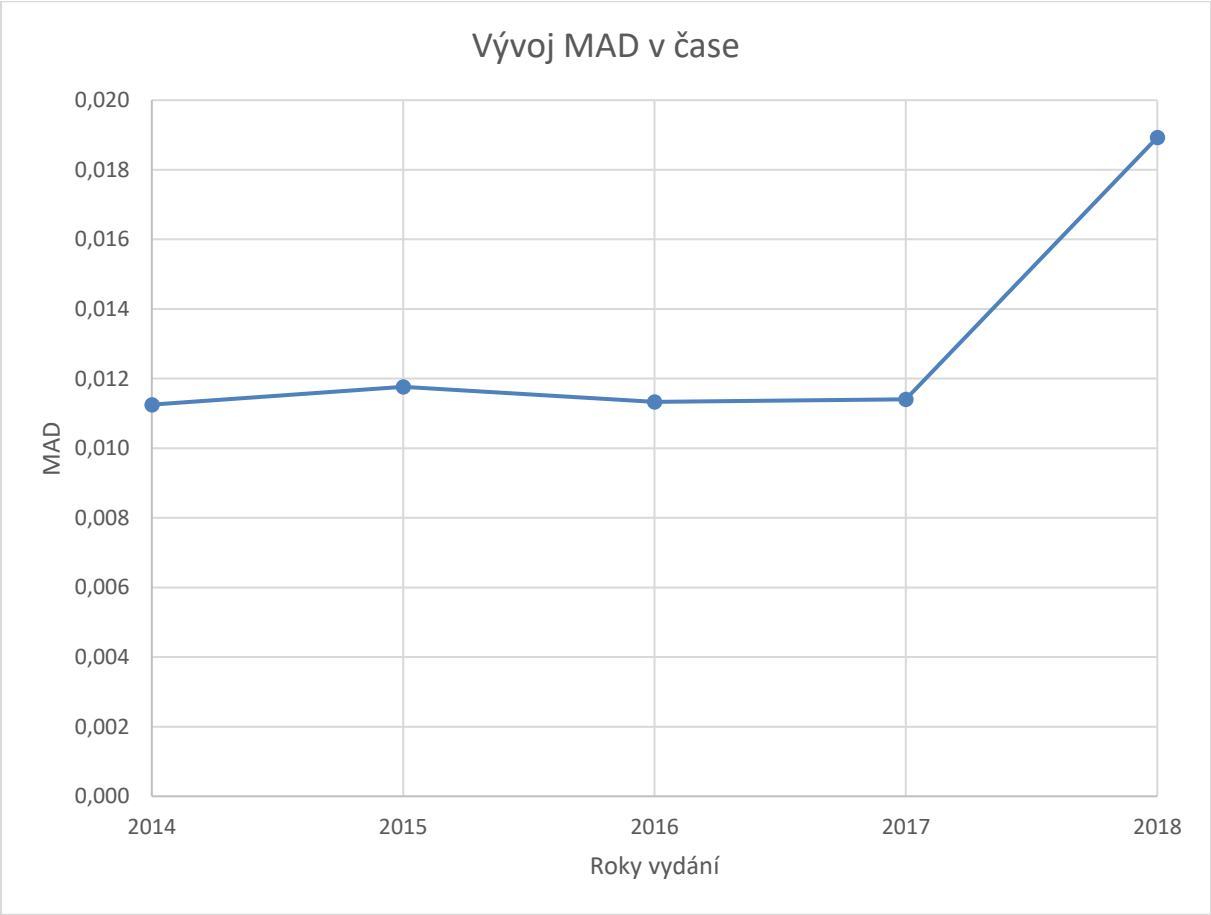
Účelem této analýzy je snaha o zjištění vlivu roku publikování záznamů na přítomnost Benfordova zákona v citačních datech. V rámci daných roků vydání byly deduplikovány veškeré záznamy, tak aby víceoborovost a duplicita typů dokumentů nezkreslovala výsledky. Počty citovatelných záznamů jsou v letech 2015 až 2017 stabilní, za roky 2014 a 2018 jsou skokově nižší. Rozdíl mezi roky 2014 a 2015 je pravděpodobně způsoben změnou akviziční politiky Web of Science, kdy byl přidán nový index ESCI. Databáze byla sice rozšířena i retrospektivně, je však možné, že změna akviziční politiky, v poměru k retrospektivě, způsobila výrazně větší nárůst počtu dokumentů od roku 2015 do současnosti (Web of Science, 2018). Rok 2018 obsahuje menší množství záznamů právě z důvodu své nedávnosti. Velké množství záznamů dosud nebylo citováno.

Jak je z Tabulky 6. patrné, pro roky 2014 až 2017 zůstává hodnota MAD relativně stabilní a je možné mluvit o přijatelné shodě. Rok 2018 je vyhodnocen jako bez shody.

Autor uvažuje, zda i nízké naplnění roku citacemi a tedy nízkou variabilitou hodnot citací v roce 2018 nemůže mít vliv právě na vysokou hodnotu MAD. Zároveň se domnívá, že pokud by stejná analýza na rok 2018 byla provedena s daty staženými na jaře roku 2021, tak by hodnota

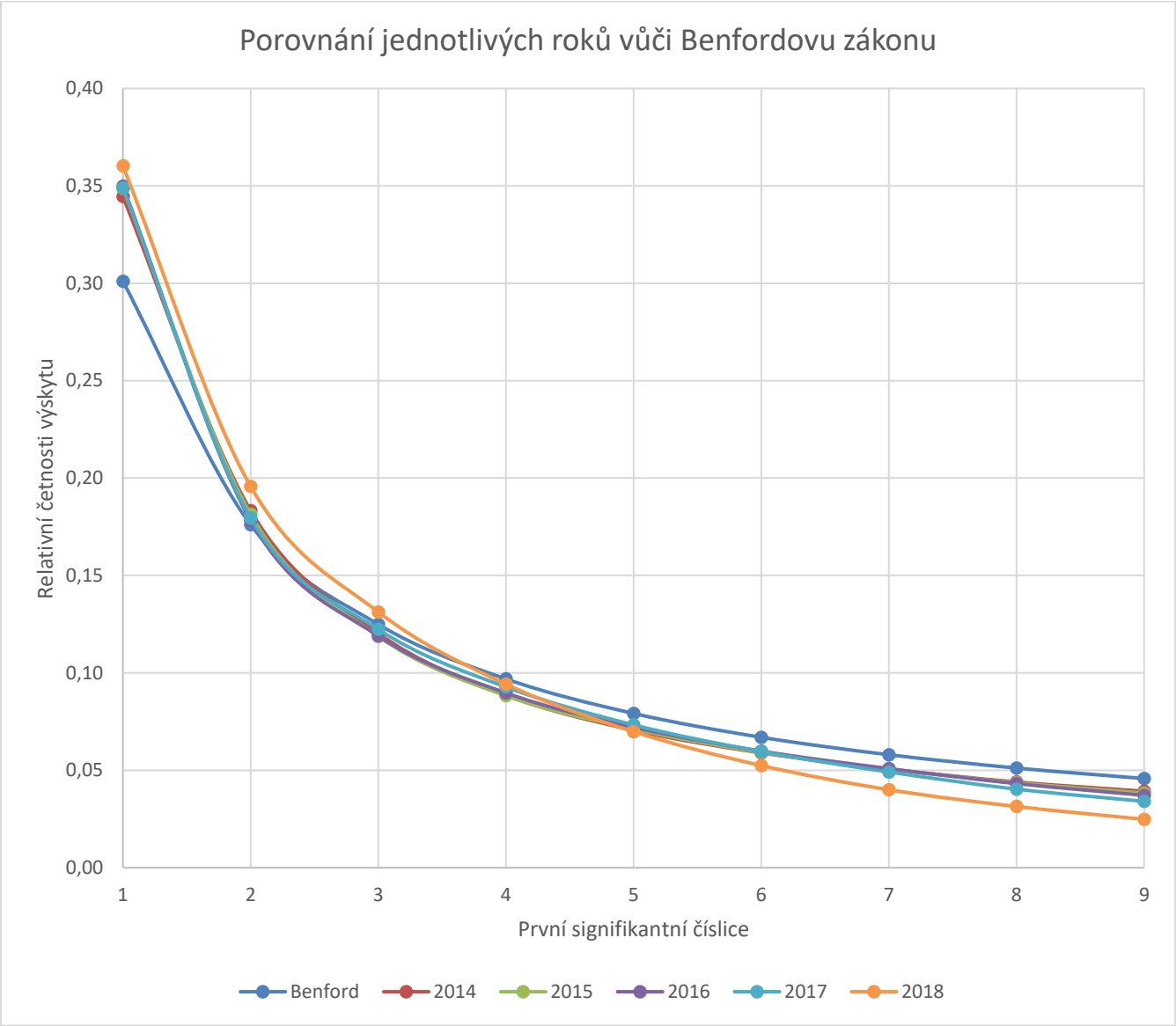
MAD byla nižší a přiblížila by se k hodnotě předchozích let. Zároveň, za stejných podmínek by předchozí roky zůstaly prakticky nezměněny.

Níže uvedený Graf 8. zobrazuje změnu hodnoty MAD mezi jednotlivými roky.




*Vývoj MAD v čase. Graf 8.*

Graf 9. porovnání let vůči Benfordovu zákonu zobrazuje vysoké podobnosti křivek různých let. Všechny roky jsou v četnosti výskytu číslice jedna výrazně odchýlené. Rok 2018 je odchýlen více než ostatní roky.




*Porovnání jednotlivých roků vůči Benfordovu zákonu. Graf 9.*

## 5.5 Analýza 5. – agregát let a oborů, separace typů dokumentů

| Analýza 5. - agregát let a oborů, separace typů dokumentů | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Absolutní počty | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ |
| Article | 2445675 | 1302314 | 883727 | 665518 | 529796 | 433885 | 362934 | 305156 | 262112 | 7191117 | 16256 |
| Proceedings Paper | 363983 | 164574 | 93058 | 59057 | 40360 | 29375 | 21835 | 16668 | 13144 | 802054 | 10408 |
| Review | 160975 | 89851 | 61463 | 46606 | 37541 | 30840 | 26529 | 23095 | 20641 | 497541 | 11391 |

*Analýza 5. - agregát let a oborů, separace typů dokumentů - absolutní počty. Tabulka 7.*

| Analýza 5. - agregát let a oborů, separace typů dokumentů | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Relativní počty | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | |
| Article | 0,34010 | 0,18110 | 0,12289 | 0,09255 | 0,07367 | 0,06034 | 0,05047 | 0,04244 | 0,03645 | 0,009795 |
| Proceedings Paper | 0,45381 | 0,20519 | 0,11602 | 0,07363 | 0,05032 | 0,03662 | 0,02722 | 0,02078 | 0,01639 | 0,040418 |
| Review | 0,32354 | 0,18059 | 0,12353 | 0,09367 | 0,07545 | 0,06198 | 0,05332 | 0,04642 | 0,04149 | 0,006002 |

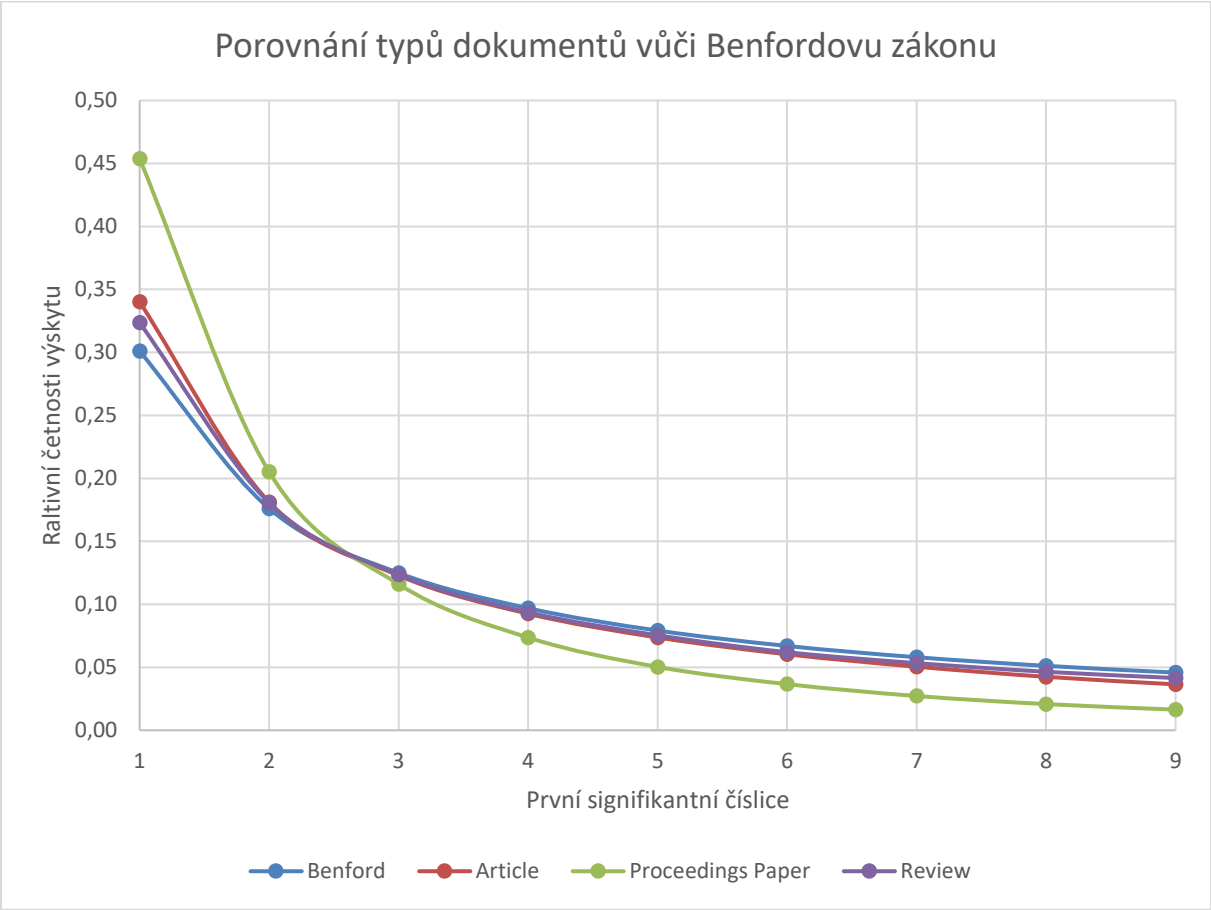
*Analýza 5. - agregát let a oborů, separace typů dokumentů - relativní počty. Tabulka 8.*

V této analýze jsou vyčleněny pouze typy dokumentů za celé pětileté období a všechny obory. Uvedeny jsou tabulky pro absolutní i relativní počty výskytů prvních signifikantních číslic. Z Tabulky 7. je viditelné, že největší množství alespoň jednou citovaných záznamů je obsaženo v typu dokumentu Article. Dále v Tabulce 8. je patrné, že typy dokumentů Article a Review dosahují přijatelné shody. Proceedings paper je však vysoko nad horní hranicí Nigriniho intervalu 0,015 a je tedy vyhodnocen jako bez shody.

Podle maximálního počtu citací by se mohlo zdát, že Article je také typ dokumentu s největší intenzitou citování. V tomto ohledu však je však výrazně intenzivněji citován Review. Ačkoliv obsahuje mnohem menší počet dokumentů, tak průměrné Review obdrží přibližně 3x víc citací než průměrný článek. V některých „tvrdých" vědách obdrží až 5,4x více citací (Miranda, 2018). Celková suma citací všech záznamů typu Article je však vyšší než celková suma citací u Review. Pouze průměrný počet citací jednotlivých záznamů je u Article nižší.

Nejnižší intenzitu citování z těchto tří typů dokumentů má právě Proceedings paper. Bylo by však chybné tvrdit, že právě nízká citační intenzita tohoto typu dokumentu je jedinou příčinou jeho nevyhovění Benfordovu zákonu. V tomto typu dokumentu pravděpodobně figurují i další aspekty citačních zvyklostí, které nejsou a ani nemohou být v těchto datech zachyceny, a které mohou do dat vnášet strukturu neslučitelnou se sledováním Benfordova zákona (viz. bod 2. Konstrukce datasetu v kapitole 1.2).

Graf 10. níže zobrazuje průběhy křivek relativních četnosti výskytu prvních signifikantních číslic jednotlivých typů dokumentů vůči pravděpodobnostem Benfordova zákona. Je zde zřetelně ilustrován odklon Proceedings paper od Benfordova zákona.




*Porovnání typů dokumentů vůči Benfordovu zákonu. Graf 10.*

## 5.6 Analýza 6. - agregát oborů, separace let a typů dokumentů

| Analýza 6. - agregát oborů, separace let a typů dokumentů | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Article | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ |
| 2014 | 418197 | 224464 | 149084 | 111517 | 90083 | 76732 | 66505 | 58282 | 52434 | 1247298 | 9815 |
| 2015 | 516070 | 269453 | 178492 | 134476 | 110486 | 92532 | 79849 | 69336 | 61082 | 1511776 | 16256 |
| 2016 | 519942 | 266356 | 181525 | 139988 | 114121 | 95684 | 81779 | 69578 | 60062 | 1529035 | 13065 |
| 2017 | 505189 | 266792 | 187210 | 144590 | 115482 | 94226 | 77863 | 63799 | 53627 | 1508778 | 14679 |
| 2018 | 486277 | 275249 | 187416 | 134947 | 99624 | 74711 | 56938 | 44161 | 34907 | 1394230 | 6706 |

*Analýza 6. - agregát oborů, separace let a typů dokumentů - Article, absolutní počty. Tabulka 9.*

| Analýza 6. - agregát oborů, separace let a typů dokumentů | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Proceedings Paper | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ |
| 2014 | 76073 | 36659 | 22003 | 14496 | 10355 | 7455 | 5744 | 4509 | 3573 | 180867 | 6070 |
| 2015 | 78442 | 38293 | 22716 | 15133 | 10382 | 7996 | 6015 | 4573 | 3665 | 187215 | 6477 |
| 2016 | 80166 | 38281 | 21962 | 13822 | 9509 | 6821 | 5002 | 3844 | 3015 | 182422 | 10408 |
| 2017 | 76920 | 32723 | 17424 | 10422 | 6839 | 4831 | 3447 | 2560 | 2001 | 157167 | 3386 |
| 2018 | 52382 | 18618 | 8953 | 5184 | 3275 | 2272 | 1627 | 1182 | 890 | 94383 | 394 |

*Analýza 6. - agregát oborů, separace let a typů dokumentů - Proceedings paper, absolutní počty. Tabulka 10.*

| Analýza 6. - agregát oborů, separace let a typů dokumentů | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Review | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ |
| 2014 | 23563 | 14676 | 10138 | 7723 | 6051 | 4880 | 4207 | 3605 | 3305 | 78148 | 5668 |
| 2015 | 30099 | 17266 | 11721 | 8722 | 6889 | 5673 | 4961 | 4298 | 3926 | 93555 | 11391 |
| 2016 | 33750 | 18774 | 12538 | 9015 | 7552 | 6289 | 5487 | 4854 | 4397 | 102656 | 4831 |
| 2017 | 36329 | 19190 | 12903 | 10014 | 8113 | 6833 | 5991 | 5327 | 4808 | 109508 | 1824 |
| 2018 | 37234 | 19945 | 14163 | 11132 | 8936 | 7165 | 5883 | 5011 | 4205 | 113674 | 2041 |

*Analýza 6. - agregát oborů, separace let a typů dokumentů - Review, absolutní počty. Tabulka 11.*

Pro přesnější představu o velikostech datasetů jsou na této straně uvedeny tabulky (9., 10. a 11.) s absolutními počty výskytů prvních signifikantních číslic v datasetech. Rozděleny jsou na separátní tabulky typů dokumentů a v rámci nich na jednotlivé roky. Byla zajištěna pouze oborová deduplikace a tím vyřešena víceoborovost záznamů.

Následující tři tabulky (12., 13. a 14) zobrazují relativní výskyty prvních signifikantních číslic. Tato analýza byla přidána dodatečně, aby poskytla větší granularitu pohledu na typy dokumentů a roky. Rozbor těchto tabulek je výjimečně umístěn před tabulky (oproti ostatním analýzám). Tato analýza je rozdělena na tři části dle typů dokumentů.

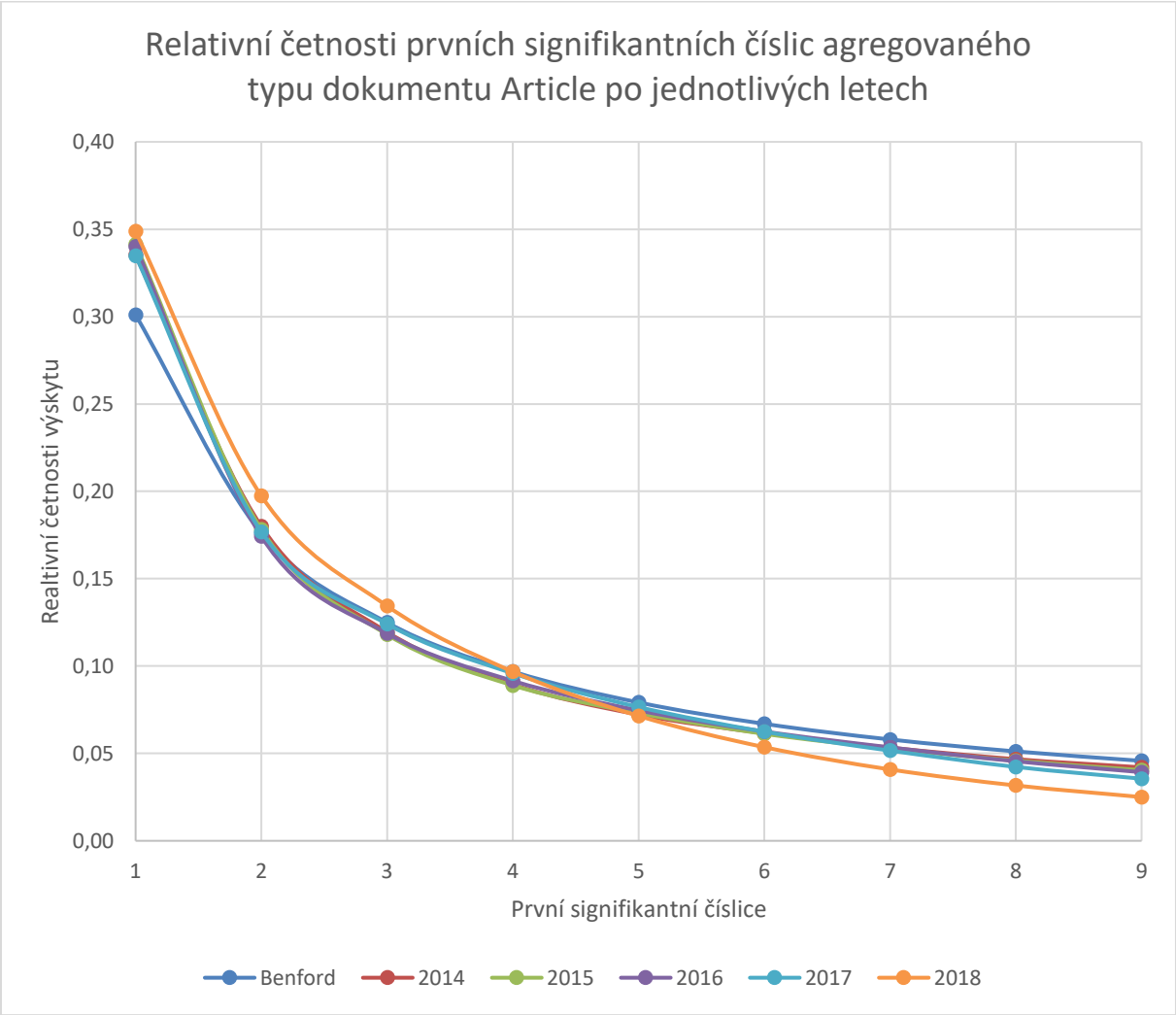
Article: Meziroční nárůst alespoň jednou citovaných záznamů v letech 2014 až 2016 je způsoben rozšiřováním databáze Web of Science a tendencí celosvětového růstu množství vědeckých výstupů. V Tabulce 12. je pro roky 2017 a 2018 naopak viditelný úbytek záznamů. To je dáno tím, že větší část záznamů ještě nestihla být citována. N v Tabulce 12. tedy reprezentuje počet alespoň jednou citovaných záznamů v příslušném roce, pro daný typ dokumentu a stažených v období od 24. 2. do 12. 4. 2020 (včetně oprav).

Hodnoty MAD v Tabulce 12. jsou poměrně vyrovnané a dle Nigriniho intervalů splňují jednotlivé roky přijatelné shody. Výjimkou je rok 2018, který značně vybočuje a je vyhodnocen bez shody s Benfordovým zákonem. Trend je shodný s výsledky analýzy 4.

| Analýza 6. - agregát oborů, separace let a typů dokumentů | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Article | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |
| 2014 | 0,33528 | 0,17996 | 0,11953 | 0,08941 | 0,07222 | 0,06152 | 0,05332 | 0,04673 | 0,04204 | 1247298 | 9815 | 0,008471 |
| 2015 | 0,34137 | 0,17824 | 0,11807 | 0,08895 | 0,07308 | 0,06121 | 0,05282 | 0,04586 | 0,04040 | 1511776 | 16256 | 0,009440 |
| 2016 | 0,34005 | 0,17420 | 0,11872 | 0,09155 | 0,07464 | 0,06258 | 0,05348 | 0,04550 | 0,03928 | 1529035 | 13065 | 0,008670 |
| 2017 | 0,33483 | 0,17683 | 0,12408 | 0,09583 | 0,07654 | 0,06245 | 0,05161 | 0,04229 | 0,03554 | 1508778 | 14679 | 0,007675 |
| 2018 | 0,34878 | 0,19742 | 0,13442 | 0,09679 | 0,07145 | 0,05359 | 0,04084 | 0,03167 | 0,02504 | 1394230 | 6706 | 0,017458 |

*Analýza 6. - agregát oborů, separace let a typů dokumentů - Article, relativní počty. Tabulka 12.*

Níže uvedený Graf 11. zobrazuje relativní četnosti výskytů prvních signifikantních číslic všech pěti let sledovaného období v porovnání s pravděpodobnostmi výskytů prvních signifikantních číslic dle Benfordova zákona.



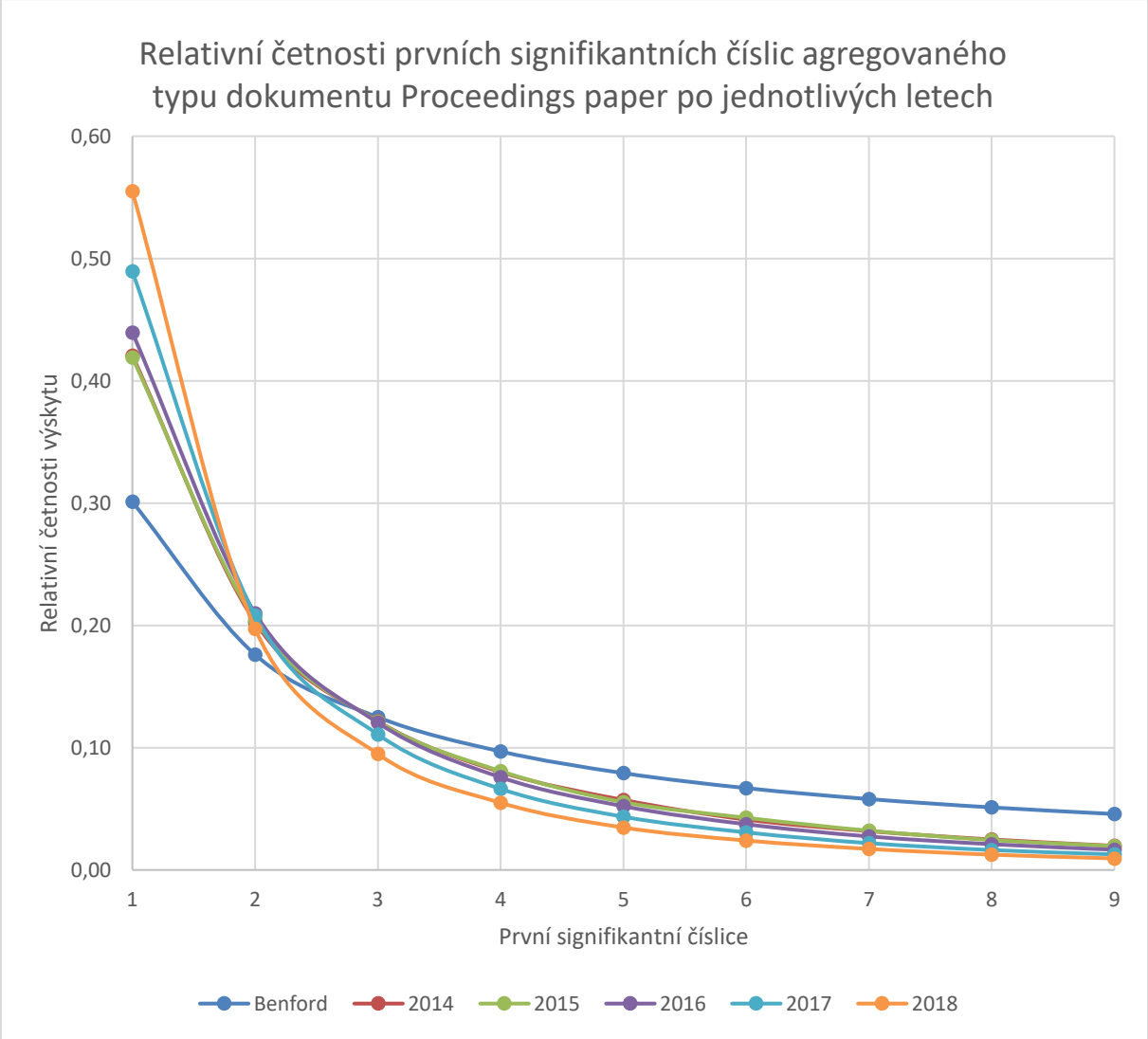

*Relativní četnosti prvních signifikantních číslic agregovaného typu dokumentu Article po jednotlivých letech. Graf 11.*

Proceedings paper: Meziroční změny v počtech alespoň jednou citovaných záznamů jsou způsobeny stejnými vlivy jako u typu dokumentu Article. Oproti Article je zde patrný větší poločas citovanosti, tzn. mezi vydáním a citováním článku je větší prodleva. Záznamy z let 2017 a 2018 byly v době stažení dat nedostatečně citovány a tím větší část nebyla citována ani jednou. Tento typ dokumentu podle MAD a hodnot Nigriniho intervalů nesleduje Benfordův zákon. Data uvedená v Tabulce 13. umožňují zjištění, že odchylky jsou v letech 2017 a 2018 výrazně větší.

| Analýza 6. - agregát oborů, separace let a typů dokumentů | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proceedings Paper | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |
| 2014 | 0,42060 | 0,20268 | 0,12165 | 0,08015 | 0,05725 | 0,04122 | 0,03176 | 0,02493 | 0,01975 | 180867 | 6070 | 0,032481 |
| 2015 | 0,41899 | 0,20454 | 0,12134 | 0,08083 | 0,05545 | 0,04271 | 0,03213 | 0,02443 | 0,01958 | 187215 | 6477 | 0,032536 |
| 2016 | 0,43945 | 0,20985 | 0,12039 | 0,07577 | 0,05213 | 0,03739 | 0,02742 | 0,02107 | 0,01653 | 182422 | 10408 | 0,038262 |
| 2017 | 0,48942 | 0,20821 | 0,11086 | 0,06631 | 0,04351 | 0,03074 | 0,02193 | 0,01629 | 0,01273 | 157167 | 3386 | 0,049000 |
| 2018 | 0,55499 | 0,19726 | 0,09486 | 0,05493 | 0,03470 | 0,02407 | 0,01724 | 0,01252 | 0,00943 | 94383 | 394 | 0,061141 |

*Analýza 6. - agregát oborů, separace let a typů dokumentů - Proceedings paper, relativní počty. Tabulka 13.*

V Grafu 12. jsou zachyceny jednotlivé roky pro typ dokumentů Proceeding paper. Nízká podobnost se Benfordovým zákonem je zřetelná. Číslice jedna je výrazně více zastoupená, než u Benfordova zákon a i číslice čtyři a vyšší jsou zatíženy velkou zápornou odchylkou. Nejvýrazněji vychýlen je právě rok 2018.



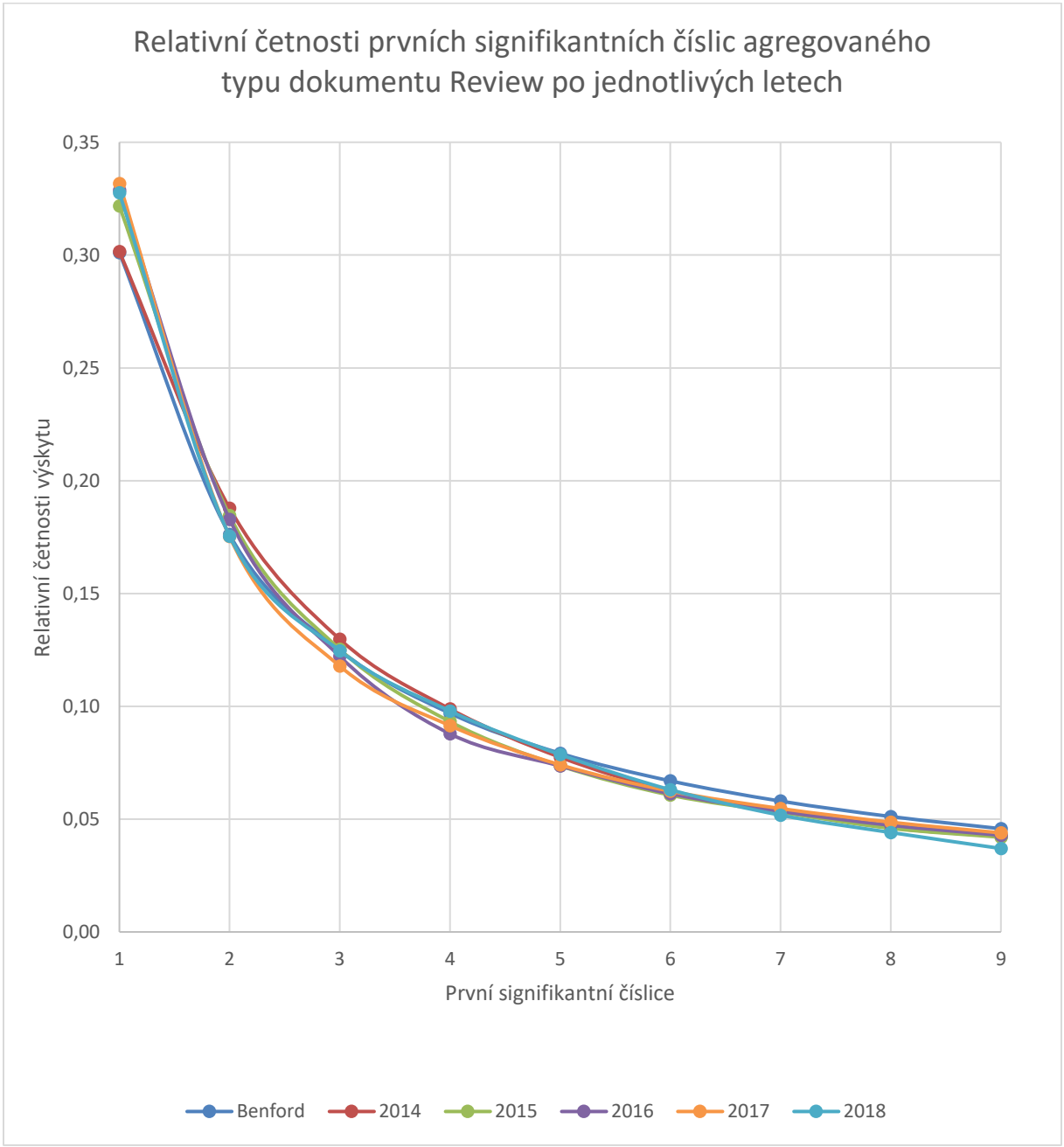

*Relativní četnosti prvních signifikantních číslic agregovaného typu dokumentu Proceedings paper po jednotlivých letech. Graf 12.*

Review: Tento typ dokumentu obsahuje až neuvěřitelně konstantní nárůst citovaných dokumentů a hodnot MAD. Poločas citovanosti je zřejmě dostatečně nízký, aby více jak roční (od 1. 1. 2018 do 31. 12. 2018 až 24. 2. do 12. 4. 2020 je rozmezí od 14 do 26 měsíců) odstup sběru dat od data vydání byl dostatečný pro ustálení hodnoty MAD vůči Benfordovu zákonu. Dle hodnot MAD v Tabulce 14. a Nigriniho intervalů se tento typ dokumentu pohybuje na pomezí blízké a přijatelné shody.

| Analýza 6. - agregát oborů, separace let a typů dokumentů | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Review | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ | MAD |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |
| 2014 | 0,30152 | 0,18780 | 0,12973 | 0,09883 | 0,07743 | 0,06245 | 0,05383 | 0,04613 | 0,04229 | 78148 | 5668 | 0,004200 |
| 2015 | 0,32173 | 0,18455 | 0,12528 | 0,09323 | 0,07364 | 0,06064 | 0,05303 | 0,04594 | 0,04196 | 93555 | 11391 | 0,006557 |
| 2016 | 0,32877 | 0,18288 | 0,12214 | 0,08782 | 0,07357 | 0,06126 | 0,05345 | 0,04728 | 0,04283 | 102656 | 4831 | 0,007673 |
| 2017 | 0,33175 | 0,17524 | 0,11783 | 0,09145 | 0,07409 | 0,06240 | 0,05471 | 0,04864 | 0,04391 | 109508 | 1824 | 0,006826 |
| 2018 | 0,32755 | 0,17546 | 0,12459 | 0,09793 | 0,07861 | 0,06303 | 0,05175 | 0,04408 | 0,03699 | 113674 | 2041 | 0,006120 |

*Analýza 6. - agregát oborů, separace let a typů dokumentů - Review, relativní počty. Tabulka 14.*

Graf 13. níže ukazuje vysokou podobnost křivek relativních zastoupení prvních signifikantních číslic pro typ dokumentu Review a jednotlivé roky vydání vůči pravděpodobnostem Benfordova zákona.




*Relativní četnosti prvních signifikantních číslic agregovaného typu dokumentu Review po jednotlivých letech. Graf 13.*

# 6. Analýza agregovaných citačních dat za veřejné vysoké školy a Akademii věd České republiky

Veřejné vysoké školy České republiky jsou kategorií, která obsahuje 26 institucí vedených v seznamu Ministerstva školství, mládeže a tělovýchovy (MŠMT, 2020). Dále se pracuje pouze s agregovanými daty za celou tuto kategorii. Nebylo provedeno dělení na jednotlivé instituce, některé by totiž nedokázaly splnit Nigriniho a Kossovského podmínky na rozsah hodnot a velikost datasetů.

Kategorie Akademie věd České republiky obsahuje všech 54 ústavů Akademie věd včetně strukturních pracovišť, jako jsou KNAV A SSČ[12] (AVČR, 2020). Dále se pracuje pouze s agregovanými daty za celou tuto kategorii. Nebylo provedeno dělení na jednotlivé ústavy, některé by také nedokázaly splnit Nigriniho a Kossovského podmínky na rozsah hodnot a velikost datasetů.

K výběru vědeckých výstupů obou výše zmíněných kategorií byla zvolena databáze RIV (RVVI, 2020). Použití této databáze zajišťuje výběr záznamů za obě kategorie při stejných podmínkách. Také stažení, filtrace a kontrola dat z této databáze je poměrně snadné. Nevýhodou tohoto zdroje je jeho nekompletnost. Při použití CRIS systému pro evidenci vědeckých výstup Akademie věd České republiky ASEP by bylo dosaženo většího množství nalezených záznamů. (ASEP za stejné období a typy dokumentů eviduje téměř 25 000 záznamů oproti 19 006 staženým pomocí RIV). Nebylo by ale možné zajistit přístup do CRIS systémů všech veřejných vysokých škol, tím ani zajistit získání dat za stejných podmínek a jejich následné porovnání.

## 6.1 Analýza citačních dat veřejných vysokých škol České republiky

Nejdříve byly staženy záznamy za veřejné vysoké školy z databáze RIV za roky 2014 až 2018. Tyto záznamy byly deduplikovány (více vysokých škol se mohlo podílet na stejném výstupu) a byla provedena kontrola na validitu identifikátoru UT WOS. V případě, že tento identifikátor chyběl či byl jiným způsobem nevyhovující, nemohl být daný výstup vyhledán v databázi Web of Science, která používá právě UT WOS jako primární identifikátor záznamů.

[12] Knihovna Akademie věd České republiky a Středisko společných činností Akademie věd České republiky jsou také zahrnuty, jelikož i infrastrukturní pracoviště mají vědecké výstupy indexované v databázi Web of Science.

V databázi Web of Science byl soubor těchto identifikátorů vyhledán a pomocí filtrů očištěn o záznamy, které měly vůči databázi RIV nekonzistentní rok vydání. Také bylo provedeno omezení na typy dokumentů Article, Proceedings paper a Review. Nalezené a filtrované záznamy byly ručně staženy po částech čítající 500 záznamů. Větší části Web of Science neumožňuje ve svém hlavním modulu exportovat[13]. API z neznámého důvodu nedokáže zpracovat dotaz na větší množství UT WOS. Ruční stahování byla jediná přijatelná volba.

Staženo bylo 56 205 záznamů. Z čehož 41 453 záznamů bylo alespoň jednou citováno. Relativní četnosti prvních signifikantních číslic u počtů citací tohoto souboru záznamů byly porovnány s pravděpodobnostmi dle Benfordova zákona v Tabulce 15. níže.

| Analýza agregovaných dat veřejných vysokých škol České republiky | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ | MAD |
| VŠ | 0,34837 | 0,18233 | 0,12363 | 0,09063 | 0,07242 | 0,05848 | 0,04924 | 0,03995 | 0,03496 | 41453 | 4565 | 0,011906 |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |

*Analýza agregovaných dat veřejných vysokých škol České republiky. Tabulka 15.*

Dle Nigriniho intervalů a hodnoty MAD je možné zařadit vybraná citační data za veřejné vysoké školy do kategorie přijatelné shody.

## 6.2 Analýza citačních dat AV České republiky

Totožným způsobem jako u veřejných vysokých školy byla opatřena data pro Akademii věd České republiky. Bylo staženo 19 006 záznamů. Z tohoto počtu 16 421 záznamů obdrželo alespoň jednu citaci. Výsledky analýzy jsou uvedeny v následující Tabulce 16.

| Analýza agregovaných dat ústavů Akademie věd České republiky | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N | $C_{max}$ | MAD |
| AVČR | 0,33506 | 0,17112 | 0,11656 | 0,08940 | 0,07685 | 0,06437 | 0,05670 | 0,04750 | 0,04245 | 16421 | 2239 | 0,007562 |
| Benford | 0,30103 | 0,17609 | 0,12494 | 0,09691 | 0,07918 | 0,06695 | 0,05799 | 0,05115 | 0,04576 | | | |

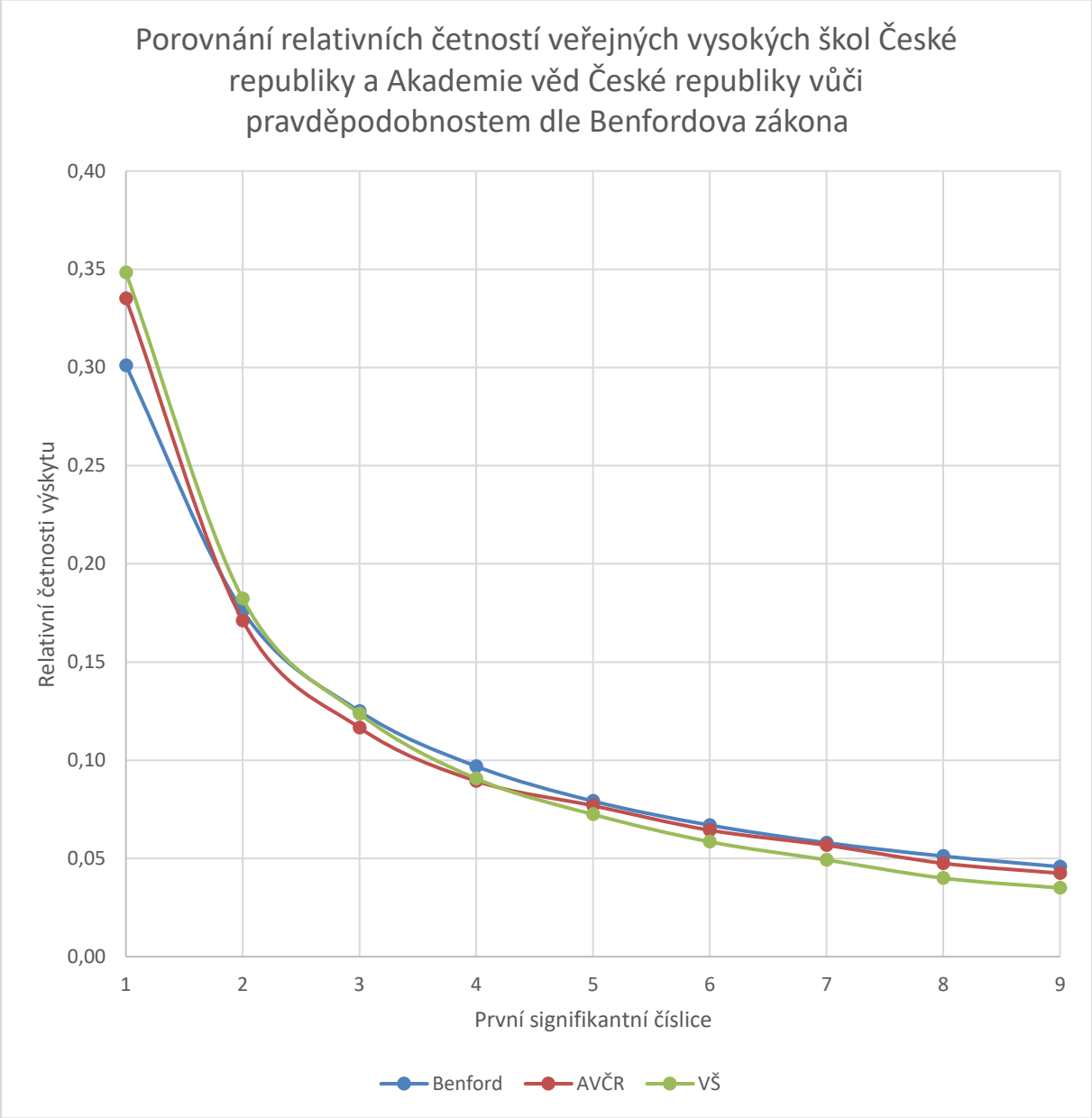
*Analýza agregovaných dat ústavů Akademie věd České republiky. Tabulka 16.*

Je zřejmé, že dle Nigriniho intervalů a hodnoty MAD se v tomto datasetu jedná o kategorii přijatelné shody.

[13] Bylo by možné vyhledané záznamy nechat exportovat do analytické nadstavby InCites, která umožňuje stažení až 40 000 záznamů. Nadstavba však není aktualizována denně, jako hlavní modul Web of Science. Také je při exportu dat do InCites vždy automaticky odstraněno několik záznamů.

## 6.3 Porovnání výsledků

Ačkoliv je dataset menší než u veřejných vysokých škol a stejně tak i nejcitovanější práce má nižší počet citací, tak je hodnota MAD u Akademie věd České republiky nižší. To značí mírně lepší shodu s Benfordovým zákonem.




*Porovnání relativních četností veřejných vysokých škol České republiky a Akademie věd České republiky vůči pravděpodobnostem dle Benfordova zákona. Graf 14.*

Autor důrazně varuje, aby nebyl výsledek této analýzy interpretován jinak než pouhé znázornění míry přítomnosti Benfordova zákona v citačních datech datasetů veřejných vysokých škol České republiky a Akademie věd České republiky. Vyšší hodnota MAD u veřejných vysokých

škol je dána faktory, které tato práce nezachycuje a nezkoumá. Také v této práci nejsou použity jemné metody forenzní analýzy, bylo by tedy chybné se snažit výsledky interpretovat jakýmkoliv jiným způsobem, aniž by byl této oblasti a datům věnován další výzkum.

# Závěr

Interpretace výsledků a dat této práce je zatížena jistými omezeními. První omezení spočívá v definici populace dat pro tuto práci. Web of Science, ačkoliv byl vybrán jako nejvhodnější zdroj, nedosahuje stejného pokrytí vědeckých výstupů jako například Google Scholar a ani ten nepokrývá celosvětovou vědeckou produkci. Je tedy nezbytné spokojit se například s kombinací konkrétního oboru, daného typu dokumentu v daném roce, evidovaného v databázi Web of Science jako s kompletní populací. V rámci databáze a kombinace již není více záznamů.

Druhé omezení vyplývá z živosti databáze Web of Science, která je aktualizována denně. Během období stahování a oprav dat od 24. 2. do 12. 4. 2020 docházelo k průběžným aktualizacím citační databáze. Aktualizace sice byly ve stažených záznamech podchyceny a jejich duplikační vliv byl odstraněn, měnily se však počty citací. Autor je přesvědčen, že ve vztahu k Benfordovu zákonu byl vliv aktualizací v takto krátkém časovém úseku zanedbatelný. Právě aditivní efekt nových záznamů a citací je stejný pro všechny kombinace oborů a typů dokumentů.

Třetí omezení je obsaženo v metodě vyhodnocování dat. MAD i Nigrinim navržené intervaly nejsou klasických statistickým testem jako je například Chí-kvadrát či Z-test. Není možné pomocí MAD a intervalů testovat nulovou hypotézu. MAD je však stále solidním nástrojem popisu odchylek citačních dat od Benfordova zákona a vzhledem k jeho nezávislosti na velikosti datasetu také jedním z mála použitelných.

Čtvrté a poslední omezení spočívá v replikovatelnosti. Pokud by došlo k pokusu o stažení dat se stejnými filtry (rok, typ dokumentu a obor) v současné době či kdykoliv později, tak by byl nejen obdržen jiný počet záznamů, ale také by velká část záznamů měla uveden jiný (většinou vyšší) počet citací než v době stažení. Samotná citační data jsou sice archivována autorem této práce, ale nemohou být poskytnuta či zveřejněna, jsou totiž vázána licenčními podmínkami společnosti Clarivate.

Z výsledků je patrné, že citační data databáze Web of Science mají velmi podobné relativní četnosti prvních signifikantních číslic, jako jsou pravděpodobnosti prvních signifikantních číslic u Benfordova zákona. U většiny analýz je také pozorován trend snižující se četnosti první signifikantní číslice s její zvyšující se hodnotou, v souladu s Benfordovým zákonem. Tato diplomová práce tedy přispěla k rozšíření palety oblastí, ve kterých se lze s tímto zákonem setkat.

Při pohledu na všechny grafy celé této práce je nutno zmínit jeden ze zajímavých jevů. Relativní četnost číslice jedna na prvním signifikantním místě je u drtivé většiny datasetů této práce vyšší než je pravděpodobnost výskytu dle Benfordova zákona. Například pro analýzu 2. je u 1277 oborů z 1355 jednička zastoupena častěji. Nejen to, největší odchylka se objevuje právě u této číslice, ostatní číslice mají odchylku od Benfordova zákon z pravidla nižší. Obdobně je na tom i číslice dva. Průnik empirických křivek relativních četností s pravděpodobnostmi Benfordova zákona se typicky objevuje mezi číslicí dva a tři. Vyšší číslice pak bývají zastoupeny s nižší relativní četností než u Benfordova zákona. U analýzy 2. má 1143 oborů relativní četnost výskytu první signifikantní číslice devět nižší než pravděpodobnost dle Benfordova zákon. Tento jev „natočení křivky“ lze pozorovat u všech analýz této práce.

Existují jisté pokusy o modifikaci vzorce Benfordova zákona pomocí koeficientu. Při použití takového koeficientu pak citační data výrazně více vyhovují Benfordovu zákonu, je kompenzováno právě „natočení křivky“. Příkladem mohou být (Egghe, 2012) nebo (Tseng, 2017). Autor se však rozhodl držet klasického Benfordova zákona. To totiž umožnilo jeho použití jakožto etalonu pro srovnání mezi analýzami. Pokud by pro každou analýzu byl použit vlastní koeficient měnící tvar křivky, jak jej aplikuje (Egghe, 2012), tak by srovnání mezi analýzami nebylo smysluplné.

Tyto články však nabízejí perspektivu pokračování této diplomové práce. Například empirické stanovení koeficientů dle Egghe pro jednotlivé kombinace oborů, let a typů dokumentů. V kombinaci s podobně modifikovaným Zipfovým zákonem by pak mohly být prováděny rekonstrukce citačních křivek jednotlivých oborů na základě malého množství dat. Tyto rekonstrukce, pokud by byly dostatečně přesné, mohou sloužit jako grafické znázornění citačních křivek a pozic konkrétních výstupů v nich. Pro účely scientometrických analýz či podkladů pro hodnocení vědecké produkce by taková grafická znázornění měla vysokou informační hodnotu a byla zároveň intuitivní.

Také je možné pokračovat analýzou dalších scientometrických indikátorů a to nikoliv na vzorcích, ale na celých populacích databáze Web of Science. V případě získání vhodných přístupů by bylo též zajímavé provést stejnou analýzu i v jiných citačních databázích a porovnat vý-

sledky. Přínosné by mohlo být i opakování této diplomové práce s časovým odstupem při porovnání výsledků. Autor je toho názoru, že při dalších, rozsáhlejších a podrobnějších studiích by mohl být Benfordův zákon zařazen mezi současné etablované bibliometrické zákony.

# Použité zdroje


ALVES, Alexandre Donizeti, Horacio Hideki YANASSE a Nei Yoshihiro SOMA. Benford's Law and articles of scientific journals: comparison of JCR® and Scopus data. *Scientometrics*. 2014, **98**(1), 173-184. DOI: 10.1007/s11192-013-1030-8. ISSN 0138-9130. Dostupné také z: http://link.springer.com/10.1007/s11192-013-1030-8

ALVES, Alexandre Donizeti, Horacio Hideki YANASSE a Nei Yoshihiro SOMA. An analysis of bibliometric indicators to JCR according to Benford's law. *Scientometrics*. 2016, **107**(3), 1489-1499. DOI: 10.1007/s11192-016-1908-3. ISSN 0138-9130. Dostupné také z: http://link.springer.com/10.1007/s11192-016-1908-3

BAWDEN, David a Lyn ROBINSON. *Úvod do informační vědy*. Doubravník: Flow, 2017. ISBN 978-80-88123-10-1.

BENFORD, Frank. The law of anomalous numbers. *Proceedings of the American philosophical society*, 1938, 551-572.

BERGER, Arno a Theodore Preston HILL. *An introduction to Benford's law*. Princeton, New Jersey: Princeton University Press, [2015]. ISBN 978-0-691-16306-2.

CAMPANARIO, Juan Miguel a María Angeles COSLADO. Benford's law and citations, articles and impact factors of scientific journals. *Scientometrics*. 2011, **88**(2), 421-432. DOI: 10.1007/s11192-011-0387-9. ISSN 0138-9130. Dostupné také z: http://link.springer.com/10.1007/s11192-011-0387-9

EGGHE, Leo a Raf GUNS. Applications of the generalized law of Benford to informetric data. *Journal of the American Society for Information Science and Technology*. 2012, **63**(8), 1662-1665. DOI: 10.1002/asi.22690. ISSN 15322882. Dostupné také z: http://doi.wiley.com/10.1002/asi.22690

KOSSOVSKY, Alex Ely. *Benford's law: theory, the general law of relative quantities, and forensic fraud detection applications*. New Jersey: World Scientific, [2015]. ISBN 978-981-4651-20-2.

KOSSOVSKY, Alex Ely. *Small is beatiful: Why the small is numerous but the big is rare in the world*. Lavergne: New York, 2017. ISBN 978-0-692-91241-6.

HINDLS, Richard a Stanislava HRONOVÁ. Benford's Law and Possibilities for Its Use in Governmental Statistics. *Statistika*. 2015, **95**(2), 11.

MIR, Tariq Ahmad. Citations to articles citing Benford's law: a Benford analysis. *ArXiv:1602.01205*. 2018, 12.

MIRANDA, Ruben a Esther GARCIA-CARPINTERO. Overcitation and overrepresentation of review papers in the most cited papers. *Journal of Informetrics*. 2018, **12**(4), 1015-1030. DOI: 10.1016/j.joi.2018.08.006. ISSN 17511577. Dostupné také z: https://linkinghub.elsevier.com/retrieve/pii/S1751157718300555

NEWCOMB, Simon. Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*. 1881, 1881(4), 39-40.

NIGRINI, Mark J. *Benford's law: applications for forensic accounting, auditing, and fraud detection* [online]. 1. Hoboken, New Jersey: Wiley, 2012 [cit. 2019-09-03]. ISBN 978-1-118-28284-7.

PEARSON, Karl. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 2009, **50**(302), 157-175. DOI: 10.1080/14786440009463897. ISSN 1941-5982. Dostupné také z: https://www.tandfonline.com/doi/full/10.1080/14786440009463897

Pracoviště AV. *Akademie věd České republiky* [online]. Praha: Akademie věd České republiky, 2020 [cit. 2020-07-25]. Dostupné z: https://www.avcr.cz/cs/o-nas/struktura/pracoviste-av/

PŘEHLED VYSOKÝCH ŠKOL V ČR. *Ministerstvo školství, mládeže a tělovýchovy* [online]. Praha: Ministerstvo školství, mládeže a tělovýchovy, 2013 [cit. 2020-07-25]. Dostupné z: https://www.msmt.cz/vzdelavani/vysoke-skolstvi/prehled-vysokych-skol-v-cr-3

REJSTŘÍK INFORMACÍ O VÝSLEDCÍCH. *Informační systém výzkumu, experimentálního vývoje a inovací* [online]. Praha: Úřad vlády České republiky, 2016 [cit. 2020-07-25]. Dostupné z: https://www.rvvi.cz/riv

TSENG, Hsiang-chi, Wei-neng HUANG a Ding-wei HUANG. Modified Benford's law for two-exponent distributions. *Scientometrics*. 2017, **110**(3), 1403-1413. DOI: 10.1007/s11192-016-2217-6. ISSN 0138-9130. Dostupné také z: http://link.springer.com/10.1007/s11192-016-2217-6

Web of Science: Emerging Sources Citation Index. *Clarivate.com* [online]. 22 Thomson Place, 36T3 Boston, MA 02210: Clarivate Analytics, 2018 [cit. 2020-07-12]. Dostupné z: https://clarivate.com/webofsciencegroup/solutions/webofscience-esci/

*Web of Science* [online]. Philadelphia, United States: Clarivate Analytics, 2020 [cit. 2020-02-13]. Dostupné z: https://www.webofknowledge.com/

# Seznam obrázků



# Seznam tabulek

# Seznam grafů



# Seznam příloh