Analiza i modelowanie wpływu czynników leksykalnych na popularność prac naukowych

JULIAN SIENKIEWICZ

Wydział Fizyki Politechniki Warszawskiej

Zjazd Fizyków Polskich, Wrocław, 11 września 2017

MOTYWACJA

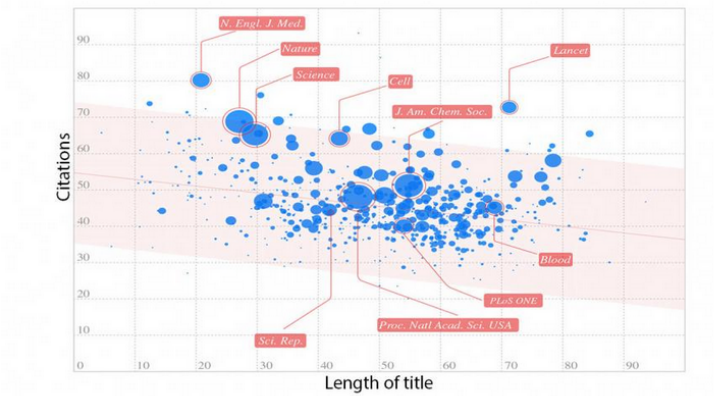
1. Powszechnie uznaje się, że liczba cytowań zebranych przez artykuł może być traktowana jako **miara uwagi** (lub popularność) uzyskanej w środowisku naukowym.
2. Istotnym jest więc zbadanie jak **właściwości tekstu** publikacji naukowych wiążą się z rozpowszechnianiem wyników naukowych w postaci liczby uzyskanych cytowań
3. Wreszcie: odniesienie się do wyników otrzymanych przez Letchforda i in. (R Soc Open Sci 2, 150266) sugerujących, iż istnieje **ujemna korelacja** pomiędzy długością tytułu oraz liczbą cytowań (tzn. im krótszy tytuł, tym więcej cytowań).

Science AAAS

Home | News | Journals | Topics | Careers

Latest News | ScienceInsider | ScienceShots | Sifter | From the Magazine | About News | Quizzes

SHARE



With a few exceptions, studies published in journals that tend toward shorter paper titles get more citations annually than those published in journals with longer paper titles. A. Letchford et al., R.Soc. Open sci. 2 (2015)

In brief, papers with shorter titles get more citations, study suggests

By **Dalmeet Singh Chawla** | Aug. 25, 2015 , 7:00 PM

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

News & Comment > News > 2016 > March > Article

NATURE | NEWS

Papers with shorter titles get more citations

Intriguing correlation mined from 140,000 papers.

Boer Deng

26 August 2015

Rights & Permissions

To William Shakespeare, brevity was the soul of wit. For scientists, it may be even more valuable, as conciseness seems to correlate with how frequently a research paper is cited.

Adrian Letchford and his colleagues at the University of Warwick in Coventry, UK, analysed the titles of 140,000 of the most highly cited peer-reviewed papers published between 2007 and 2013 as listed on Scopus, a research-paper database. They compared the lengths of the papers' titles with the number of times each paper was cited by other peer-reviewed papers— a statistic sometimes used as a crude measure of importance.

As they report in *Royal Society Open Science*[1], "journals which publish papers with shorter titles receive more citations per paper".

Szczegółowe cele

Główne czynniki

Ilościowe zbadanie jak poszczególne **cechy tekstu** publikacji naukowych, takie jak
- długość tekstu,
- złożoność tekstu,
- emocje w tekście

są związane z **liczbą cytowań**.
W ten sposób mam zamiar zidentifikować **kluczowe czynniki** wpływające na popularność naukową.

Szczegółowe cele

Główne czynniki

Ilościowe zbadanie jak poszczególne **cechy tekstu** publikacji naukowych, takie jak
- długość tekstu,
- złożoność tekstu,
- emocje w tekście

są związane z **liczbą cytowań**.
W ten sposób mam zamiar zidentifikować **kluczowe czynniki** wpływające na popularność naukową.

Różnice w cytowaniu

Wskazanie **różnic** w sposobie cytowania **najpoczytniejszych** oraz **typowych** publikacji. W tym celu zostanie użyte podejście **regresji kwantylowej**.

DANE

Portal **Web of Science**

WSTĘPNA OBRÓBKA DANYCH

Prace określone jako artykuły, opublikowane w okresie **1995—2004**, spełniające następujące dwa warunki:

1. czasopisma **nieprzerwanie aktywne** w ww. okresie (np. eliminacja czasopism `PLOS`)
2. w podanym okresie czasopismo musiało opublikować co najmniej **1.000 artykułów** (np. eliminacja `Rev Mod Phys`)

DANE

Portal **Web of Science**

WSTĘPNA OBRÓBKA DANYCH

Prace określone jako artykuły, opublikowane w okresie **1995—2004**, spełniające następujące dwa warunki:

1. czasopisma **nieprzerwanie aktywne** w ww. okresie (np. eliminacja czasopism `PLOS`)
2. w podanym okresie czasopismo musiało opublikować co najmniej **1.000 artykułów** (np. eliminacja `Rev Mod Phys`)

ZBIÓR DANYCH

- ponad **4.300.000** artykułów z ok. **1.500** różnych czasopism,
- dane dotyczące tytułu, liczby autorów, zawartości streszczenia oraz dyscypliny naukowej,
- **liczba cytowań** na dzień 31 grudnia 2014

UŻYTE ZMIENNE

cecha	tytuł	streszczenie
długość	liczba znaków	liczba słów
złożoność	— wskaźnik z C Herdana	indeks czytelności FOG F wskaźnik z C Herdana
emocje	walencja pobudzenie	walencja pobudzenie
liczba autorów		

1. indeks FOG: $F = \left(\frac{\#slow}{\#zdan} + 100 \frac{\#slow\ \ zlozonych}{\#slow} \right)$
2. miara C Herdana: $C = \frac{\log N}{\log M}$ $\left[\begin{array}{c} \mathrm{M - dlugosc\ tekstu} \\ \mathrm{N - liczba\ unikalnych\ slow} \end{array} \right]$
3. wskaźnik z: $z_{M,N} = \frac{N - \mu(M)}{\sigma(M)}$
4. Walencja — emocjonalny znak (ładunek) tekstu (dodatni - 9, obojętny - 5, ujemny - 1)
5. Pobudzenie — poziom reakcji emocjonalnej (niski - 1, średni - 5, wysoki - 9)

REGRESJA KWANTYLOWA (QUANTILE REGRESSION - QR)

ZAŁOŻENIE

Znaleźć współczynniki α i β prostej

$$Y = \alpha(\tau) + \beta(\tau)X,$$

która dzieli zbiór tak, aby ułamek τ punktów leżało poniżej linii a $(1 - \tau)$ poniżej.

KORZYŚCI PODEJŚCIA

- możemy rozpatrywać różne przedziały zmiennej Y,
- logarytm p-ego kwantyla jest równy p-emu kwantylowi zlogarytmowanej zmiennej Y

REGRESJA KWANTYLOWA (QUANTILE REGRESSION - QR)

ZAŁOŻENIE

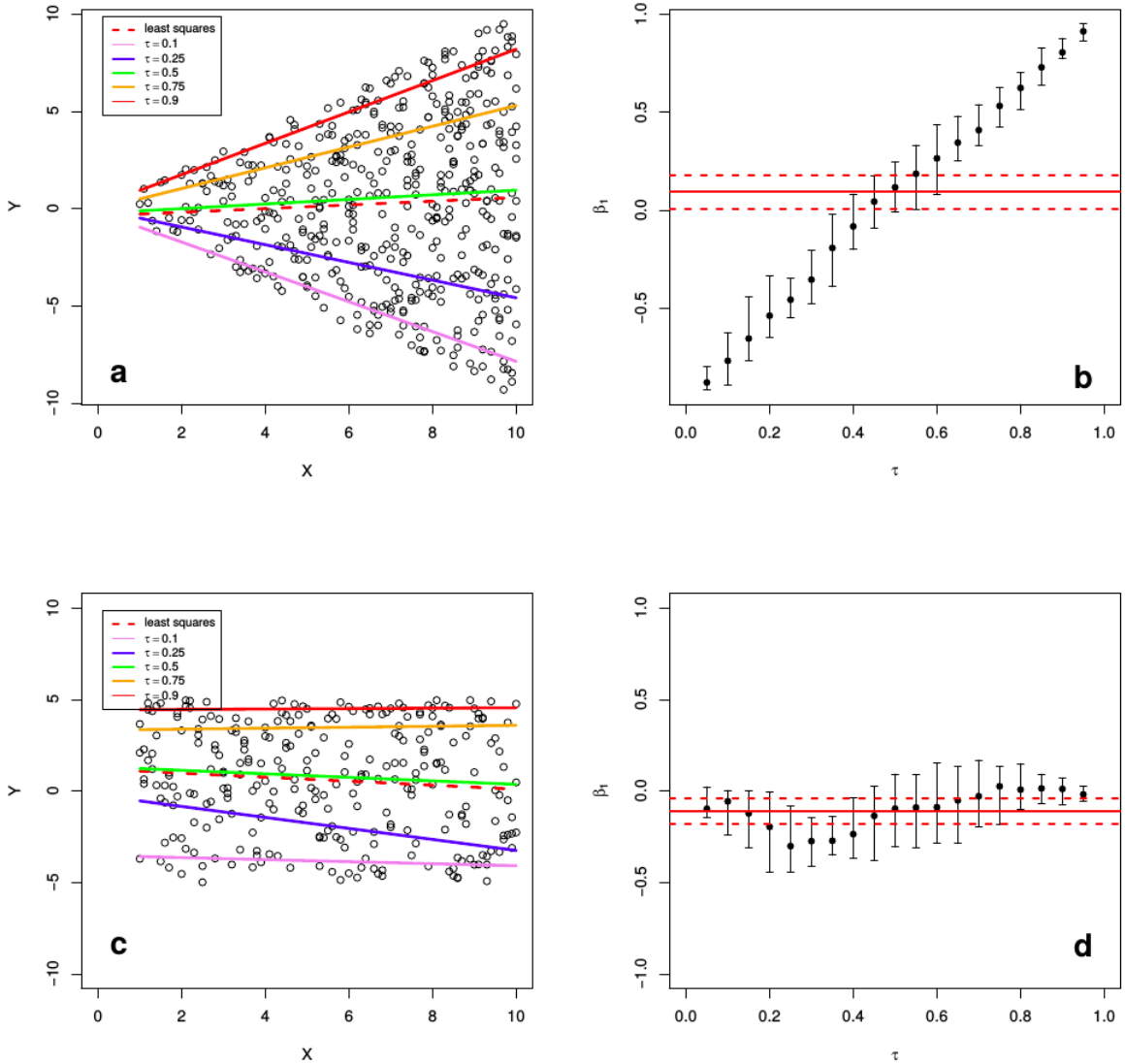
Znaleźć współczynniki α i β prostej

$$Y = \alpha(\tau) + \beta(\tau)X,$$

która dzieli zbiór tak, aby ułamek τ punktów leżało poniżej linii a $(1 - \tau)$ poniżej.

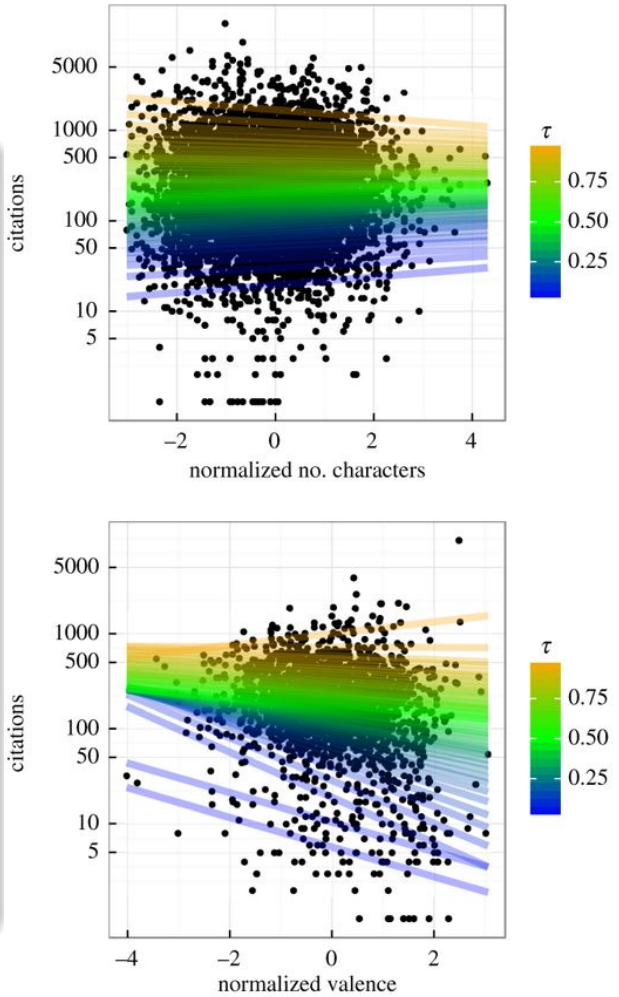
KORZYŚCI PODEJŚCIA

- możemy rozpatrywać różne przedziały zmiennej *Y*,
- logarytm *p*-ego kwantyla jest równy *p*-emu kwantylowi zlogarytmowanej zmiennej *Y*

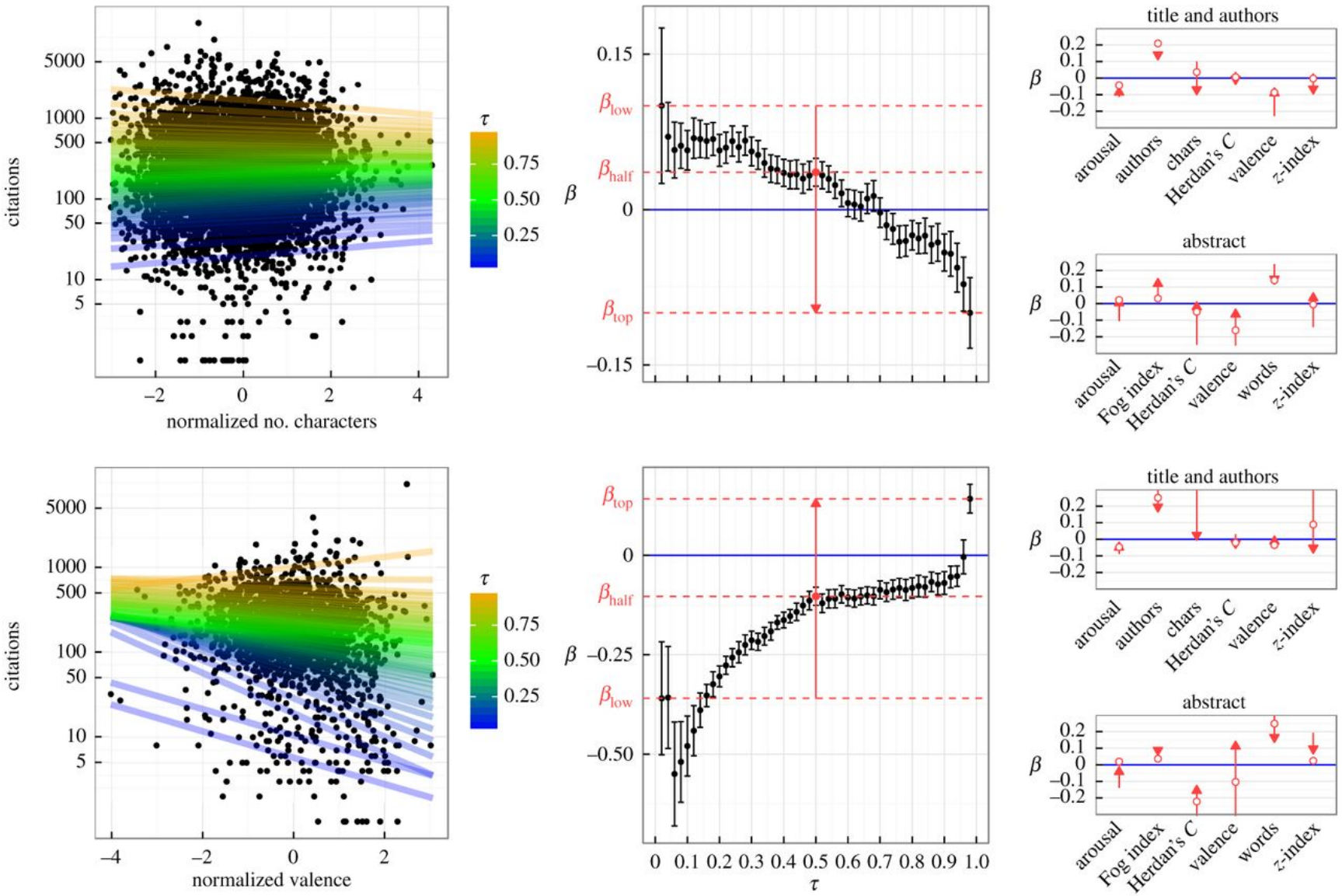


WYNIKI QR

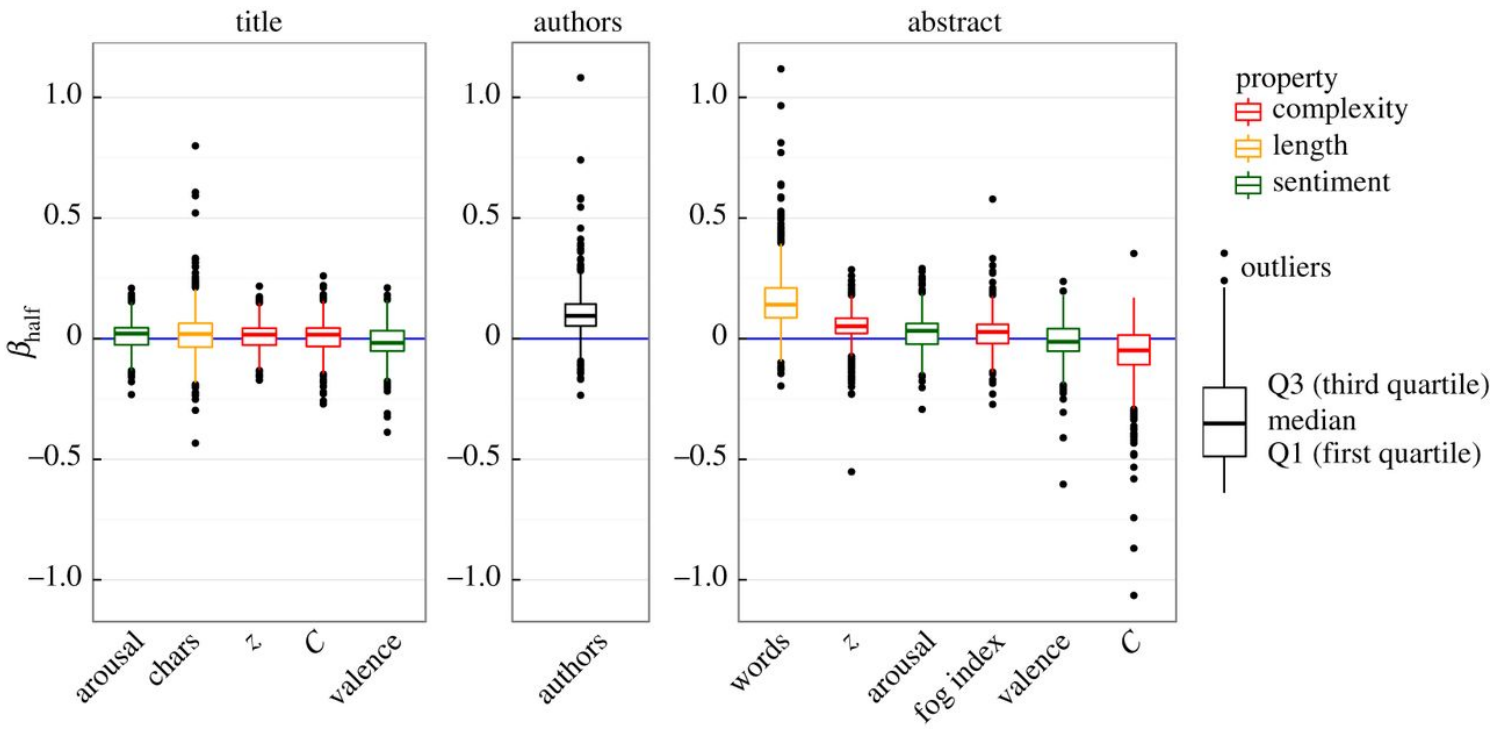
- duży rozrzut punktów — nie można rozróżnić, czy relacja pomiędzy X oraz Y jest rosnąca czy malejąca,
- wartość współczynnika korelacji Pearsona r wynosi: $r = 0.02 \pm 0.01$ dla długości tytułu (Science) oraz $r = -0.21 \pm 0.03$ dla walencji (Nature Genetics),
- jednoznaczna sugestia, iż analiza nie może opierać się na wykorzystaniu liniowych narzędzi, bazujących na założeniu homoskedastyczności (jednakowe odchyłki dla różnych wartości X).



WYNIKI - QR


citations
normalized no. characters
τ
0.75
0.50
0.25
β
β_low
β_half
β_top
title and authors
abstract
arousal
authors
chars
Herdan's C
valence
z-index
Fog index
words
normalized valence
τ

WYNIKI - PORÓWNANIE CZYNNIKÓW

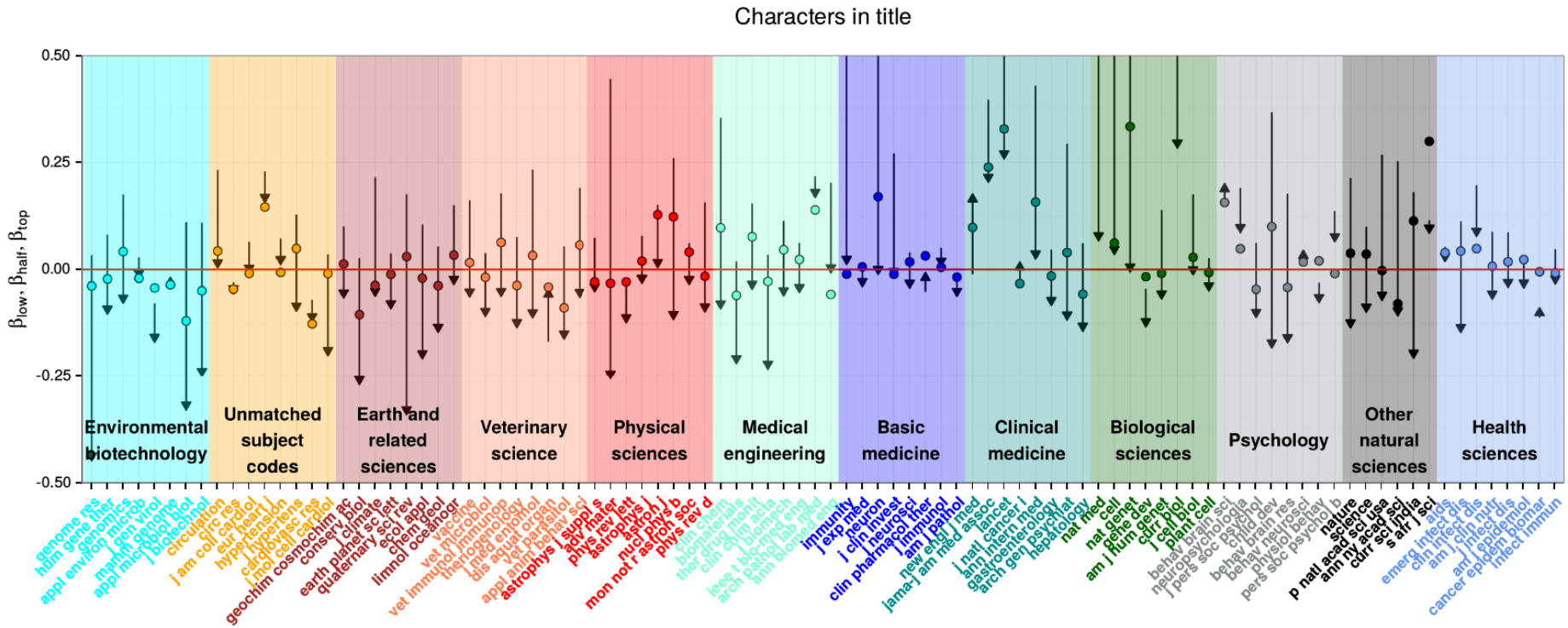


- Wpływ poszczególnych czynników jest dość słaby - $|\beta| < 0.5$ ($\beta = \ln 2$ oznacza, że liczba cytowań Y podwaja się przy przesunięciu X o jedno odch. stand.).
- Najsilniejszymi czynnikami są (i) liczba słów w streszczeniu, (ii) liczba autorów, oraz (iii) wskaźnik z w streszczeniu (ponad 75% of czasopism — czyli całe "skrzynia" znajduje się powyżej zera).
- Czynniki streszczenia są bardziej widoczne niż te dotyczące tytułu.

WYNIKI - RÓŻNICA POMIĘDZY TYPOWYMI ORAZ NAJLEPSZYMI

cecha	czynnik	$\beta_{top} > \beta_{half}$	$\beta_{top} < \beta_{half}$	$\beta_{top} \neq \beta_{half}$
długość	liczba znaków (tytuł)	2.6%	44.4%	47.0%
	liczba słów (streszczenie)	8.3%	29.4%	36.7%
			średnia	41.9%
złożoność	*C* Herdana (tytuł)	18.7%	8.5%	27.2%
	C Herdana (streszczenie)	34.9%	6.5%	41.4%
	wskaźnik *z* (tytuł)	8.3%	16.7%	25.0%
	wskaźnik *z* (streszczenie)	24.6%	7.7%	32.3%
	indeks FOG (streszczenie)	26.4%	8.0%	34.4%
			średnia	32.0%
emocje	pobudzenie (tytuł)	11.0%	13.5%	24.5%
	pobudzenie (streszczenie)	15.7%	13.7%	29.4%
	walencja (tytuł)	16.1%	11.3%	27.4%
	walencja (streszczenie)	29.2%	5.7%	34.9%
			średnia	29.1%
	liczba autorów	4.0%	39.6%	43.6%
			ogólna średnia	33.7%

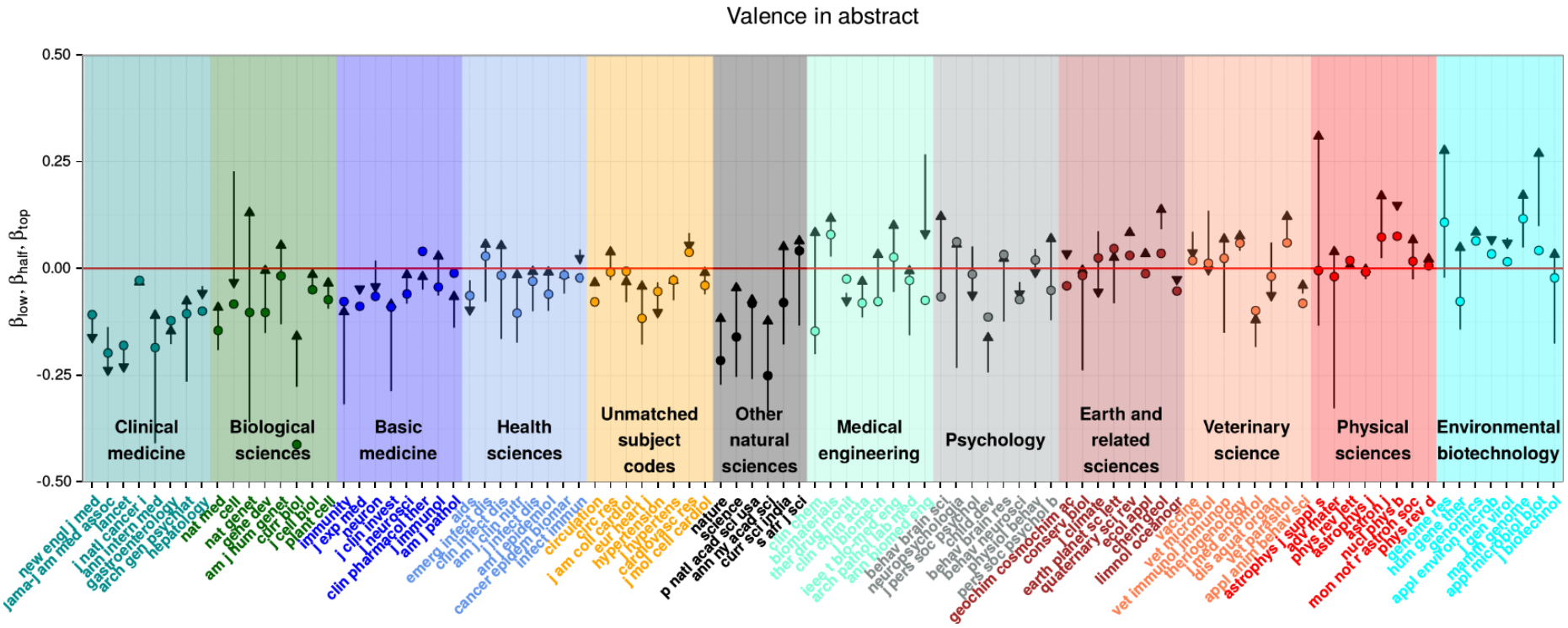
WYNIKI - PORÓWNANIE CZASOPISM (ZNAKI W TYTULE)



Wyznaczając $\exp(\beta \Delta X)$ można łatwo porównać siły czynników — w ten sposób bezpośrednio mierzymy ile średnio cytowań zyskuje się (lub traci) przesuwając się o ΔX odchylenia standardowego na zmiennej X):

- dla czasopisma *Lancet* $\beta_{\text{half}} = 0.33$, więc zwiększając liczbę znaków o 1 odch. stand. daje prawie 40% zysku w cytowaniach
- podobna operacja dla *Nature* $\beta_{\text{half}} = 0.038$ co odpowiada ok. 4% zyskowi.

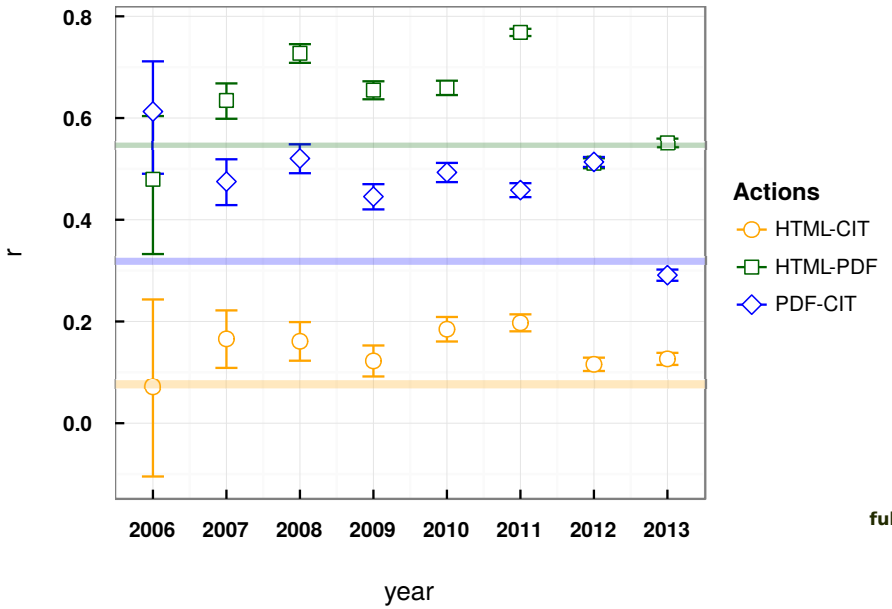
WYNIKI - PORÓWNANIE CZASOPISM (WALENCJA W STRESZCZENIU)

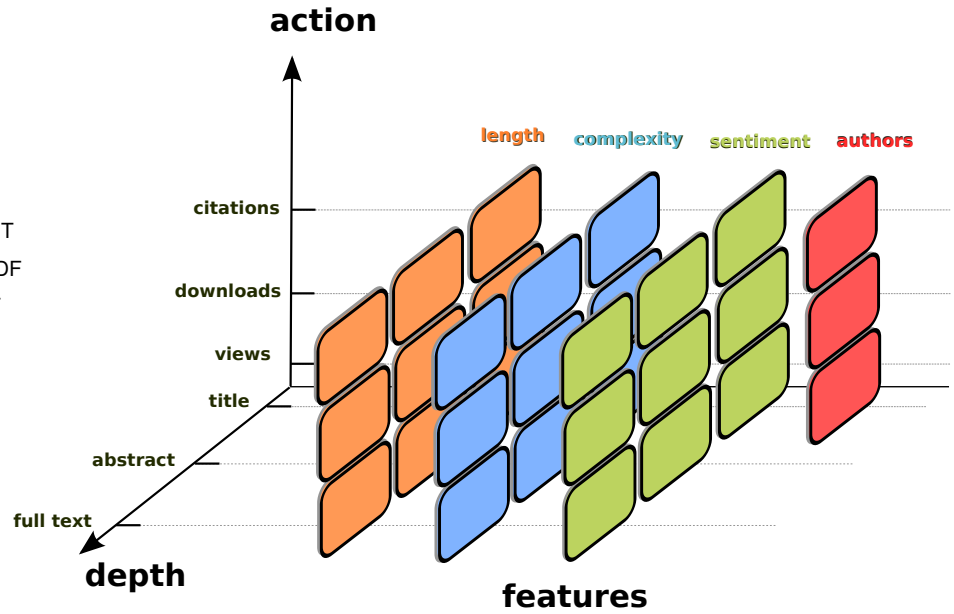


Zróżnicowanie wśród czasopism daje się częściowo wytłumaczyć poprzez przynależność do odrębnych dyscyplin naukowych, np. dla *medycyny klinicznej* wszystkie wartości współczynnika β są poniżej zera, podczas gdy dla *nauk fizycznych* większość jest dodatnia.

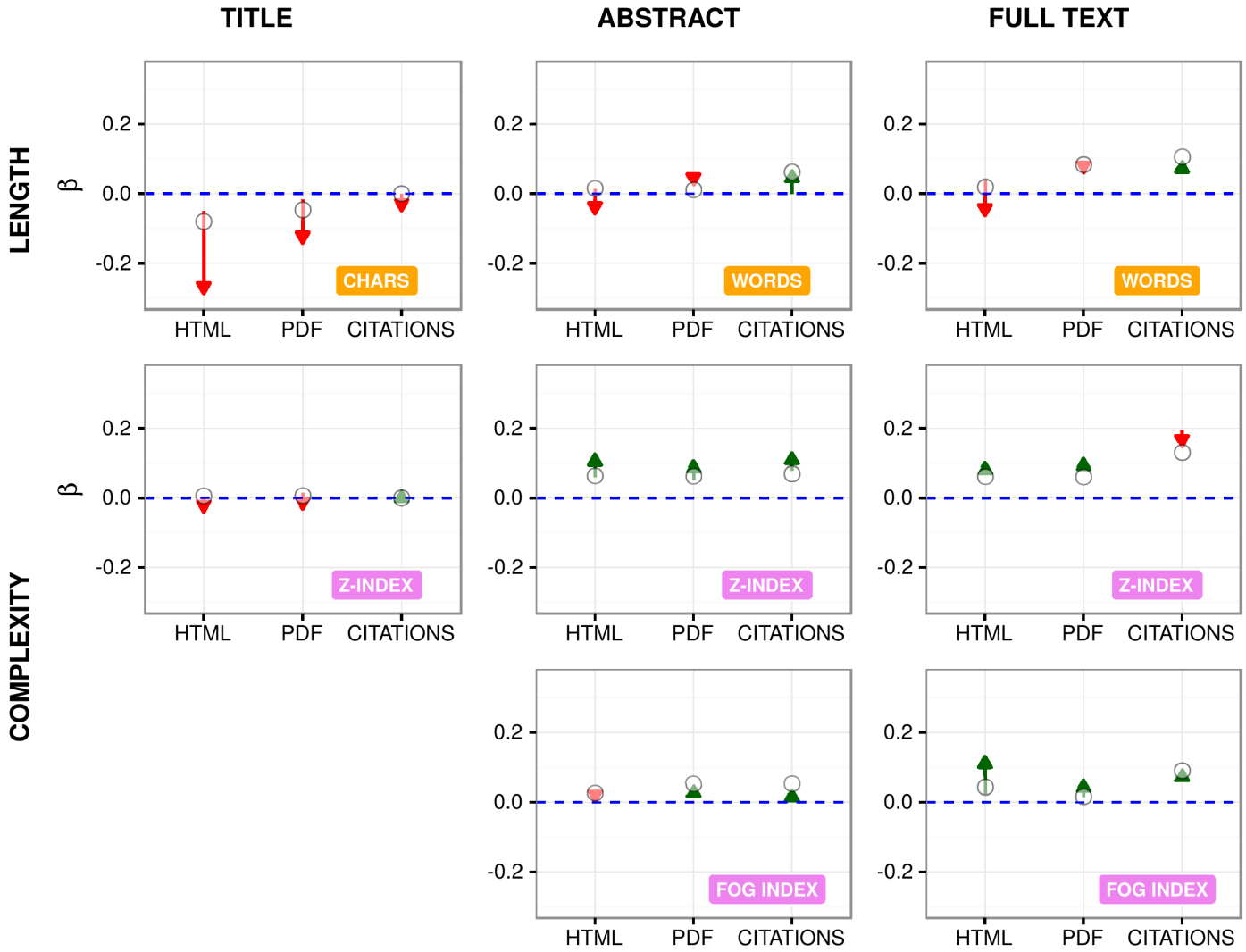
DALSZE BADANIA

1. Jakie zależność istnieją pomiędzy różnymi miarami działania (tj. czytaniem, pobraniem dokumentu etc)?
2. Jak odrębne części strktury dokumentu (tytuł, streszczenie etc) wpływają na działanie?
3. Świetny “poligon doświadczalny”: baza PLOS (pełny tekst).

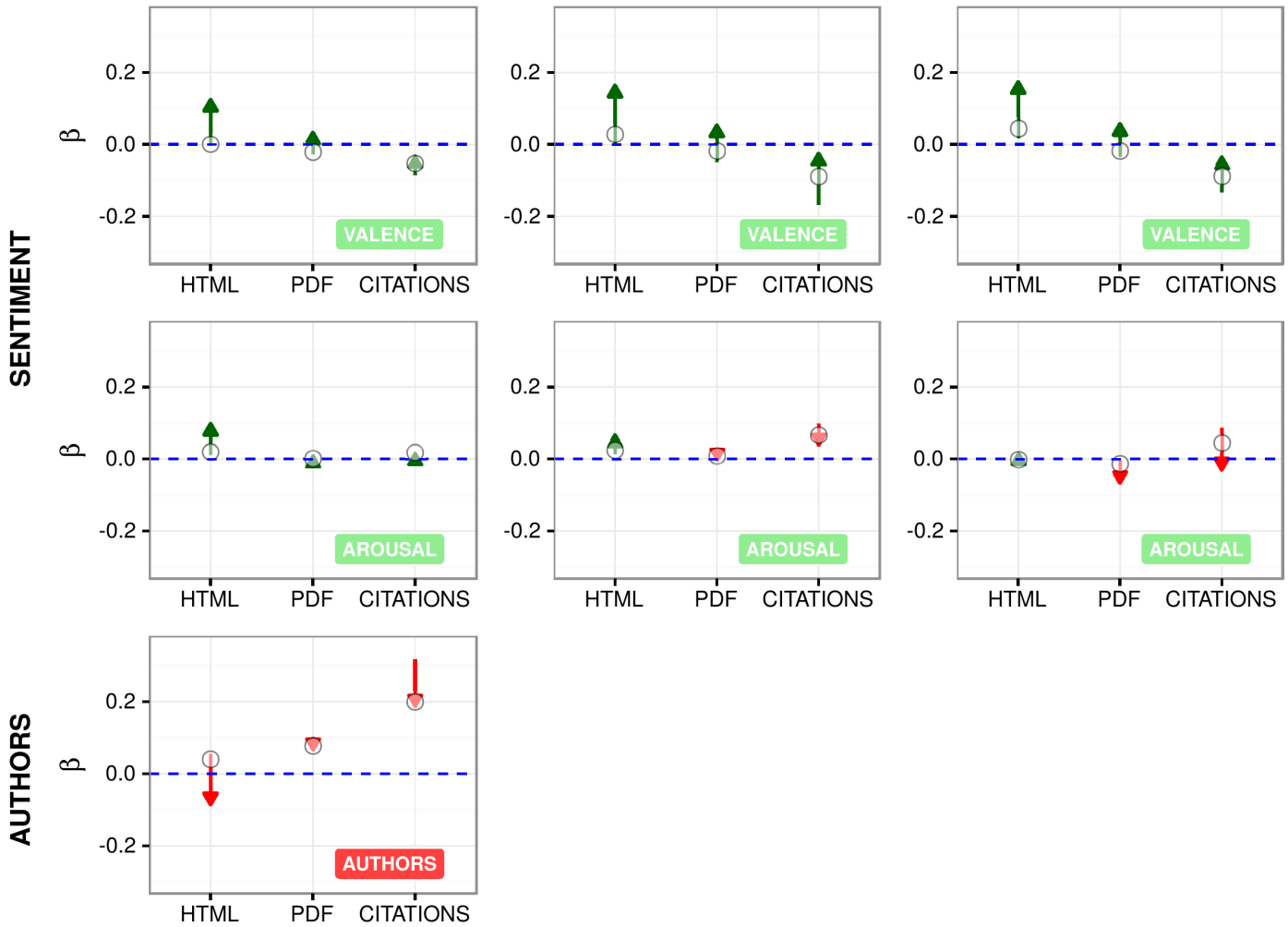




PRZYKŁADOWE WYNIKI



PRZYKŁADOWE WYNIKI



PODSUMOWANIE

1. Badanie zależności pomiędzy właściwościami tekstu publikacji naukowych oraz liczbą cytowan, którą otrzymują,
2. Główne wnioski: korelacje są **nieliniowe** i w różny sposób ujawniają się w przypadku **najlepiej cytowanych** i **typowych** prac,
3. W przypadku większości czasopism **krótkie tytuły** są **dodatnio** skorelowane z liczbą cytowań jedynie dla **najpoczytniejszych** prac,
4. Korelecje są widoczone dla większości badanych czynników lecz efekt zwykle jest dość **słaby** ($|\beta| < 0.5$),
5. duży rozrzut wśród czasopism.

szczegóły oraz niektóre dane:
R Soc Open Sci 3, 160140 (2016)

PODZIĘKOWANIA

Eduardo G. Altmann
@ Max Planck Institute for the Physics of Complex Systems, Dresden, Germany (obecnie Univ. Sydney, Australia)