Wybrane metody prognozowania tempa rozwoju dyscyplin naukowych (indeks citing half-life, metoda regresji nieliniowej, linearyzowanej i wielomianowej drugiego stopnia)

Łukasz Opaliński
ORCID 0000-0003-2797-2777
Oddział Informacji Naukowej, Biblioteka Politechniki Rzeszowskiej

Marcin Jaromin
ORCID 0000-0002-7256-7271
Zakład Biotechnologii i Bioinformatyki, Wydział Chemiczny, Politechnika Rzeszowska

Abstrakt

Cel/Teza: Dokonano przeglądu oraz porównano wybrane statystyczne metody prognozowania tempa ewolucji dyscyplin naukowych. Materiał empiryczny badania stanowiły cytowania uzyskane przez publikacje należące do wybranych dyscyplin. Jednocześnie zaakcentowano problem warunków możliwości uogólnienia wyników badań prób losowych na szerszą populację generalną dokumentów.
Koncepcja/Metody badań: Na przykładzie danych empirycznych, na które złożyło się prawie 25 tys. cytowań, zademonstrowano metodę tworzenia przedziałów ufności dla indeksu citing half-life oraz metody ukierunkowane na uogólnienie i prognozę zidentyfikowanych w badaniu trendów. Były to: metoda regresji nieliniowej, metoda regresji linearyzowanej i metoda regresji wielomianowej drugiego stopnia.
Wyniki i wnioski: Problemy, jakie napotkały metody regresyjne, to fakt niespełniania przez nie określonych warunków Gaussa-Markova. Dla przeanalizowanych danych wykluczyło to zastosowanie podstawowych form modeli regresji jako narzędzi prognostycznych. Wymagane są korekty lub wykorzystanie innego rodzaju modeli, co stanowi perspektywę dalszych badań.
Oryginalność/Wartość poznawcza: W artykule zestawiono metody ilościowe, które nie są powszechnie stosowane w celu ewaluacji tempa rozwoju nauki. Zademonstrowano ich potencjał i użyteczność w tym względzie oraz zaznaczono potrzebę dalszego ich doskonalenia i testowania metod bardziej wyrafinowanych.

Słowa kluczowe

Bibliometria. Dziedziny i dyscypliny naukowe. Komunikacja naukowa. Naukometria. Prognozowanie. Rozwój nauki. Statystyka w informatologii.

Otrzymany: 18 stycznia 2020. Zrecenzowany: 17 lutego 2020. Poprawiony: 22 lutego 2020. Zaakceptowany: 18 kwietnia 2020.

1. Wprowadzenie

Za pierwowzór dzisiejszej, stosowanej powszechnie do celów naukoznawczych, praktyki indeksowania przypisów bibliograficznych występujących w publikacjach naukowych najczęściej uważa się indeks cytowań amerykańskiego prawodawstwa federalnego, które

wykorzystywano do śledzenia precedensów sądowych – tzw. Shepard's Citations z 1873 r. (Sosińska-Kalata & Roszkowski, 2016, 319; zob. też: Shapiro, 1992, 338). Na aktualnej w danym roku cytowalności pewnego zbioru publikacji, przejawiającej się w postaci częstotliwości wystąpień odwołań bibliograficznych do konkretnych publikacji w bibliografiach załącznikowych, a mówiąc ściślej na ich cyklu życiowym, może natomiast opierać się bibliometryczne prognozowanie tempa przyszłego rozwoju dyscyplin naukowych. Potrzeba i praktyczne zastosowania tego rodzaju prognoz zostały omówione przez autorów w ramach odrębnych opracowań (zob. Opaliński, 2017a; 2017b; Opaliński & Jaromin, 2017). W ich treści wspomniano m.in., że obiecującą metodą znajdującą zastosowanie w tym zakresie jest statystyczna analiza szeregów czasowych, jak również zaakcentowano fakt, że przedstawiona w ostatnim z wymienionych opracowań metoda konstruowania prognoz jest ściśle związana z badaniem procesu starzenia się piśmiennictwa. Za szereg czasowy mogą bowiem zostać uznane uporządkowane według daty zacytowania (bądź daty opublikowania cytowanej pozycji literaturowej) liczby cytowań otrzymanych przez pewne wybrane publikacje bądź dokumenty funkcjonujące w ramach komunikacji naukowej.

Jednym z ważniejszych aspektów wykorzystania tej metody jest możliwość budowy prognoz na podstawie dopasowanej do danych doświadczalnych funkcji regresji w postaci, w jakiej zademonstrowano ją w jednym z kolejnych podrozdziałów. Ze względu na to, że celem badań przeprowadzonych przez autorów jest zastosowanie, przetestowanie i porównanie kilku odmiennych podejść do zasadniczego, poruszanego tutaj problemu (tj. generalizacji i przewidywania zjawisk ilościowych), jak również ze względu na szeroki zakres zgromadzonych i poddanych analizie danych empirycznych nt. cytowań publikacji naukowych, autorzy zdecydowali się podzielić omówienie wyników badań na dwie części. Część pierwsza, przedstawiona w niniejszym artykule, została skupiona przede wszystkim na podstawowej metodzie, jaką jest wspomniana wyżej regresja, która występuje w kilku odmianach. Funkcja regresji jest bowiem jednym z najważniejszych, chociaż z pewnością nie jedynym narzędziem, umożliwiającym z jednej strony analizę szeregu czasowego, a z drugiej zarówno uogólnianie zjawisk zaobserwowanych w ramach prób losowych, jak i prognozowanie przyszłego przebiegu bądź kształtu takich zjawisk (spodziewanych wartości, które wystąpią w szeregu).

Część druga, która zostanie opublikowana jako odrębny artykuł w kolejnym numerze ZIN, zawiera omówienie i próbę wykorzystania nieco bardziej wyrafinowanych postaci metody regresji oraz dwóch podejść alternatywnych (wobec metody regresji), które wymagają spełnienia mniej rygorystycznych warunków metodologicznych przez budowany w określonym środowisku empiryczno-teoretycznym model, a pomimo tego wciąż niosą z sobą pewien potencjał wyjaśniający i informacyjny.

Warto dodatkowo zaznaczyć, iż za pewnego rodzaju sposób prognozowania, bądź przynajmniej sposób odzwierciedlenia aktualnego w danym roku tempa rozwoju danej dyscypliny naukowej, można także uznać powszechnie stosowany w biblio- i naukometrii indeks citing half-life. W tym względzie wyjątkowo istotna jest pewna szczególna postać tego wskaźnika, która jest jego generalizacją, tj. która wykracza poza sam materiał empiryczny będący podstawą badania, czyli poza faktycznie zbadaną próbę. Postać tę można określić jako przedział ufności dla indeksu citing half-life.

2. Uogólniony citing half-life jako symptom tempa rozwoju dziedzin naukowych

Większość zbiorów danych empirycznych o cytowaniach jest na ogół uważana za pewną próbę losową, która została zaczerpnięta ze znacznie szerszej populacji generalnej. Na populację tę składają się pozycje nieobjęte danym badaniem – wydane w latach wcześniejszych i późniejszych niż przebadane roczniki określonych wydawnictw, pozycje wydane w językach innych niż język wyselekcjonowanych do badań źródeł lub pozycje opublikowane w postaci typów wydawniczych wyłączonych poza nawias zrealizowanego badania (np. dokumentów patentowych, prac dyplomowych, tzw. szarej literatury itp.).

Indeks citing half-life jest współczynnikiem pokazującym liczbowo, jak wiekowe, tj. jak dawno wydane (przeciętnie) publikacje są wciąż wykorzystywane w danej dziedzinie lub dyscyplinie nauki. Dyscypliny rozwijające się w powolnym tempie cechują się bowiem bazowaniem na publikacjach dawnych, powstałych w odległym czasie, na równi (bądź nawet w większym stopniu) z pracami najnowszymi, wnoszącymi ze sobą innowacje metodologiczne, koncepcyjne bądź dostarczające nowo pozyskanych danych obserwacyjnych i eksperymentalnych. W dziedzinach, w których rozwój postępuje szybko – przeciwnie, publikacje wiekowe w krótkim czasie wypadają z bieżącego obiegu, ponieważ są zastępowane (wypierane) przez dorobek najnowszy (o niskim wieku), eliminujący z obszaru zainteresowań naukowców wydawnictwa zdezaktualizowane i uznawane przez społeczność badaczy za nienadające się do dalszego wykorzystywania (zob. np.: Borgman & Furner, 2002, 26; Vinkler, 1996, 375–376, 382). Uczeni działający w takich obszarach mają więc do dyspozycji wiele stosunkowo niedawno wydanych publikacji, co skutkuje „odmładzaniem się" bibliografii załącznikowych (zob. np.: Jarić et al., 2014, 526). W tym sensie indeks citing half-life informuje o tym, w którym miejscu – w kontinuum wszystkich możliwych wartości wskaźnika tempa rozwoju nauki – mieści się dana jej specjalność, dziedzina czy obszar tematyczny. Im wyższy wskaźnik citing half-life (lub dostępny w bazach Thomson Reuters indeks aggergated citing half-life) tym wolniej przebiega tempo postępu dokonującego się w pewnej dziedzinie. Można również w tym miejscu dodatkowo nadmienić, że idea takiego znaczenia i roli wskaźnika half-life wzięła swój początek z klasycznej już dzisiaj pracy Roberta Burtona i Richarda Keblera, w której po raz pierwszy zdefiniowano go na gruncie bibliometrii i opisano jego naukoznawcze zastosowania (zob. Burton & Kebler, 1960). Według Burtona i Keblera dziedziny, w których starzenie się piśmiennictwa jest procesem powolnym, mają charakter bardziej teoretyczny niż eksperymentalny (np. matematyka, geologia) i cechuje je wysoka wartość indeksu half-life[1]. Natomiast dziedziny, w których piśmiennictwo starzeje się szybko, co odzwierciedla niższa wartość indeksu, charakteryzują się bujnym rozwojem technik eksperymentalnych, są obszarem rodzenia się nowych teorii, które zajmują miejsce dotychczas obowiązujących, lub też wspierają się na dynamicznym rozwoju nowych technologii (Burton & Kebler, 1960, 22).

Indeks ten jest w rzeczywistości medianą wieku publikacji, których opisy bibliograficzne figurują w bibliografiach załącznikowych rocznika (lub roczników) danego czasopisma (lub

[1] W pracy Burtona i Keblera nie pojawiło się jeszcze wyrażone *explicite* rozróżnienie wskaźników cited i citing half-life. Powstało ono dopiero później, za sprawą filadelfijskiego Institute for Scientific Information (ISI) (zob. Sen, 1999, 327).

grupy czasopism czy innego typu dokumentów), dla którego wskaźnik jest wyznaczany (Opaliński, 2013, 155). Mediana to inaczej tzw. wartość środkowa, która jest statystyczną miarą tendencji centralnej, bardzo podobną do średniej arytmetycznej. Różnica pomiędzy nimi polega na tym, że w przeciwieństwie do średniej mediana jest niewrażliwa na występowanie nawet największych odchyleń pojedynczych wyników od średniej w danej grupie (tzw. *outliers*), tj. pojedynczych wyników bardzo dużych lub bardzo małych, które odbiegają od przeciętnej wartości całościowego wyniku. Można powiedzieć, że w dowolnym, uszeregowanym malejąco zbiorze liczbowym (np. wyników jakiegoś doświadczenia) mediana jest liczbą dzielącą go na pół w tym znaczeniu, że jedna połowa wyników ma wartość większą od mediany, a druga połowa ma wartość od niej mniejszą (Carlberg, 2012, 61–63). Oznacza to, że wybierając (losując) przypadkową wartość z całej populacji generalnej mamy dokładnie 50% szans na to, że wartość ta będzie większa niż wartość mediany, oraz 50% szans na to, że będzie ona od niej mniejsza. Rozkład badanej w doświadczeniu cechy elementów jakiejś populacji jest wtedy tzw. rozkładem dwumianowym (ang. *binomial distribution*) (Sheskin, 2007, 289–290, 305–306; zob. też: Berk & Carey, 2010, 253–254; Dowdy et al., 2004, 77; Ott & Longnecker, 2010, 265). Dla dużych prób losowych rozkład dwumianowy przybliża się rozkładem normalnym (Larocque & Randles, 2008, 33; Sheskin, 2007, 294, 300–301; Ott & Longnecker, 2010, 267). Próba duża to próba, która składa się z przynajmniej 50 wartości pobranych z populacji generalnej („n > 50").

Do oceny tego, czy trendy obecne w tempie rozwoju nauki, widziane przez pryzmat cytowalności piśmiennictwa, wykryte i skwantyfikowane w ramach przeanalizowanych przez różnych autorów prób, dają się uogólnić poza same te próby, tj. czy trendy te obowiązują również w ramach populacji generalnej, potrzebna jest uogólniona wersja tego wskaźnika. Jest nią mianowicie tzw. przedział ufności dla indeksu half-life.

Przykładowe obliczenia służące do konstruowania przedziału ufności dla wskaźnika citing half-life cytowanych w pewnym korpusie literatury źródeł bibliograficznych zamieszczono w kolejnym akapicie (wszystkie wykorzystane w tym miejscu dane empiryczne zawarto w Aneksie 2). W pierwszej kolejności wybrane do analizy dane należy zorganizować w postaci tabelarycznej (Tab. 1).

Tab. 1. Zestaw danych, który posłużył do wyznaczenia przedziałów ufności dla mediany cytowanych w obrębie zbadanej próby polskojęzycznych wydawnictw zwartych

Numeracja jednostek czasu według Rousseau (2006): średni wiek źródła w momencie zacytowania	**Rok wydania publikacji cytowanych w 2015 r.**	**Liczba cytowań prac o danym wieku**	**Skumulowana suma cytowań**
0.25	2015	68	68
1	2014	170	68+170=238
2	2013	238	238+238=476
3	2012	249	476+249=725
...	...	...	...
486	1529	1	5374

Podstawowym punktem odniesienia w tabeli 1 jest bazowy rok badania, tj. rok wydania publikacji cytujących, czyli rok 2015, oraz liczba prac (cytowań), które wystąpiły w bibliografiach załącznikowych publikacji cytujących i które same zostały wydane pomiędzy rokiem 1529 a 2015. Podając średni wiek źródeł wydanych w pewnym roku i cytowanych w roku 2015 (kolumna nr 1 w Tab. 1), przyjęto sposób jego wyznaczania i numerowania zademonstrowany i uzasadniony przez Ronalda Rousseau (zob. Rousseau, 2006). Wiek cytowanych źródeł jest ułożony w porządku rosnącym, tzn. w obrębie zbadanej próby zacytowano 68 dokumentów o średnim wieku 0.25 lat, 170 dokumentów o średnim wieku 1 roku, 238 dokumentów o średnim wieku 2 lat, itd. W ostatnim wierszu odnotowano zacytowanie jednego dokumentu o wieku 486 lat. W sumie dało to 5374 cytowania, co jest zarazem wielkością próby (n = 5374). Gdyby zapisać cały szereg wartości w rozwiniętej formie przybrałby on postać: 0.25, 0.25, ... (68 wystąpień wartości 0.25), 1, 1, ... (170 wystąpień wartości 1), itd., aż do ostatniego, pojedynczego wystąpienia wartości 486. Stąd wiadomo, że np. piętnastym wyrazem szeregu jest wartość 0.25, dwieście piętnastym – wartość 1, a siedemset pierwszym – wartość 3 itd. Przedział ufności wyznacza się za pomocą następujących wyrażeń (dalszą notację i sposób wykonywania obliczeń zaczerpnięto z pracy: Ott & Longnecker, 2010, 265–270; por. też: Sheskin, 2007, 234–235, 300–301):

$$(M_L, M_U) = (y_{(L\alpha/2)}, y_{(U\alpha/2)})$$

M_L to dolna granica przedziału ufności mediany, M_U to górna granica przedziału ufności mediany, $y_{(...)}$ symbolizuje jedną z kolejnych wartości powyższego szeregu, której pozycję w tym szeregu określa wyrażenie w nawiasie (np. $y_{(10)}$ to dziesiąta pozycja w szeregu). Ujęte w nawiasy wyrażenia (czyli pozycja wartości w szeregu) muszą zostać obliczone przy pomocy ich definicji.

$$L_{(\alpha/2)} = C_{\alpha(2),n} + 1$$

$$U_{(\alpha/2)} = n - C_{\alpha(2),n}$$

Wartość $C_{(...)}$ jest stałą, do uzyskania której potrzebna jest odczytywana z tablic tzw. wartość krytyczna $z_{(...)}$. Wartość ta pełni funkcję parametru równania i wskazuje na odsetek przypadków jakiegoś zjawiska podlegającego rozkładowi normalnemu, poniżej której znajduje się 2,5% wszystkich możliwych przypadków (realizacji zmiennej losowej) (Sheskin, 2007, 59; Vaughan, 2001, 81–83). W ten sposób ustala się więc prawdopodobieństwo popełnienia pomyłki, która nie powinna zdarzyć się częściej niż 2–3 razy na 100 różnych prób, wyników czy zdarzeń (ogólniej – realizacji zmiennej losowej). Kiedy stosuje się przybliżenie rozkładu dwumianowego rozkładem normalnym, wartość $C_{(...)}$ określona jest następującym wyrażeniem:

$$C_{\alpha(2),n} \approx \frac{n}{2} - z_{\alpha/2}\sqrt{\frac{n}{4}}$$

Wartość $z_{\alpha/2}$ dla rozkładu normalnego i poziomu istotności równego 95% wynosi 1.96 (Ott & Longnecker, 2010, 267; Sheskin, 2007, 234, 301). Stosując powyższe zasady i wzory do danych z tabeli 1 (i Aneksu 2) otrzymuje się następujące zależności:

$$C_{\alpha(2),n} \approx \frac{5374}{2} - (1{,}96 \times \sqrt{\frac{5374}{4}}) = 2687 - 71.8 = 2615.2 \approx 2615$$

Stąd:

$$L_{(\alpha/2)} = 2615 + 1 = 2616$$

$$U_{(\alpha/2)} = 5374 - 2615 = 2759$$

$$(M_{L,} M_U) = (y_{(2616)}, y_{(2759)}) = <11, 12>,$$

ponieważ 2616-tym z kolei wyrazem w ciągu wartości średniego wieku publikacji cytowanych jest 11, a 2759-tym z kolei wyrazem tego ciągu jest 12. To znaczy, że wartość tego wskaźnika wynosi nie mniej niż 11 i nie więcej niż 12 lat (włącznie). Innymi słowy, prawdziwa wartość wskaźnika citing half-life dla całej populacji mieści się w tym przedziale. Może ona wynosić np. 11.01 lub 11.5 roku, albo też 11.99 czy nawet dokładnie 12 lat. Nie musi też być ona naturalnie dokładnie tą samą wartością, którą obliczono na podstawie danych empirycznych pochodzących z próby (w tym konkretnym przypadku była to wartość wynosząca 11.27 roku).

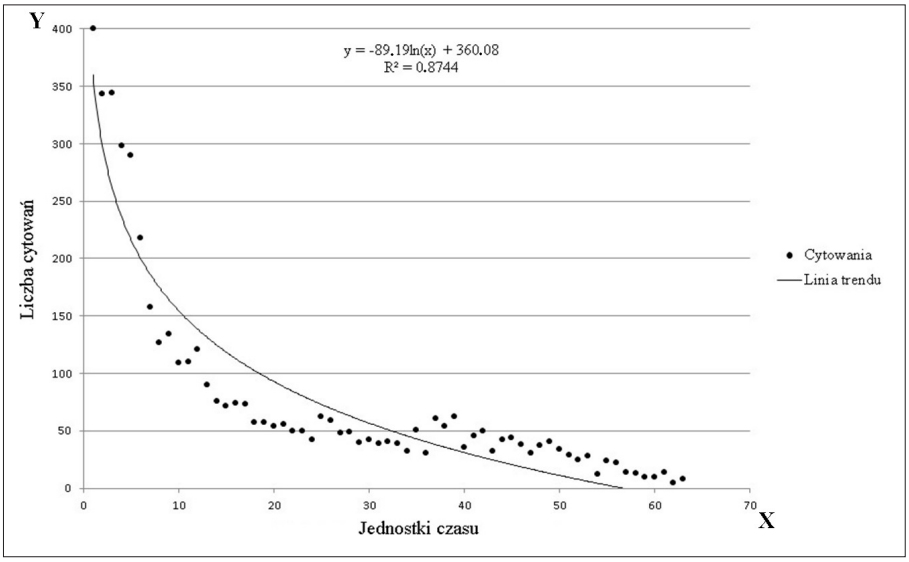
3. Istota metody regresji i warunki Gaussa-Markova

Indeks *citing half-life* daje jednak bardzo statyczny i ograniczony obraz wewnętrznej dynamiki danej dziedziny. Jedną z bardziej zaawansowanych metod znajdujących szerokie zastosowanie w modelowaniu długofalowych trendów i ich uogólnianiu jest tzw. metoda regresji. Pozwala ona znacznie wyraźniej uchwycić ekspansywność rozwoju zjawisk ilościowych. Nazwa tej metody została wprowadzona do statystyki przez Francisa Galtona, który badał zjawisko dziedziczenia przez ludzi cech ich przodków, takich jak np. wzrost (zob. np.: Aczel, 2007, 456; Allen, 1997, 1–3; Bensman, 2000, 821). Zgodnie z tą metodą w pierwszej kolejności należy przedstawić w określonej formie oryginalne dane zebrane przez badacza, a następnie zidentyfikować charakteryzujący je trend. Rysunek 1 prezentuje liczby cytowań artykułów przywoływanych w bibliografiach załącznikowych w polskojęzycznych czasopismach z zakresu nauk o Ziemi[2] w 2015 r. Dokładne liczby cytowań kolejnych, coraz starszych polskojęzycznych artykułów z czasopism (oraz pozostałych przeanalizowanych form wydawniczych), które pojawiły się w 2015 r., zamieszczono w Aneksie 2.

Równanie trendu opisującego widniejące na wykresie punkty symbolizujące cytowania może zostać ustalone metodą tzw. najmniejszych kwadratów przy użyciu programów komputerowych, takich jak np. Microsoft Excel. Należy również podkreślić, że trend opisany równaniem: $y = -89.19 \ln(x) + 360.08$ jest trendem nieliniowym. Wartości liczbowe zmiennej zależnej y odnoszą się do liczb cytowań (ilościowego poziomu cytowalności) uzyskanych przez pewne artykuły, podczas gdy wartości zmiennej niezależnej x odnoszą się do jednostek czasu, jaki upłynął od pewnego momentu początkowego. Zmienna x jest tutaj dodatkowo zlogarytmowana, tj. nie występuje ona w równaniu w „czystej postaci", ale

[2] Wykaz wszystkich czasopism, które stały się materiałem badawczym dla autorów, podaje Aneks 1.

jako ln(x). Nieliniowość trendu oznacza, że zmiana wartości y wywołana przez jednostkową zmianę wartości x (tj. zmianę o 1 przyjętą jednostkę, czyli np. o 1 rok) sama nie jest wartością stałą, ale zależy od tego, jak „duże" jest x. Im większe jest x (tj. im dalej położone na poziomej osi wykresu od jej początku) tym mniejsza – w omawianym przypadku – jest odpowiadająca jednostkowej zmianie tego x zmiana wartości y.



Rys. 1. Spadkowy trend cytowalności polskojęzycznych artykułów z czasopism o różnym wieku zarejestrowany w przebadanej próbie

Należy ponadto zauważyć, że funkcja (linia) trendu stanowi najlepsze z możliwych dopasowanie do zbioru danych empirycznych, ale nigdy nie będzie to dopasowanie idealne. Równanie trendu powinno zatem uwzględniać dodatkowy składnik losowy *e*, który byłby wyrazem (konsekwencją) istnienia losowych odchyleń danych doświadczalnych od przewidywań płynących z równania trendu. Równanie powinno więc w istocie przyjąć postać:

$$y = -89.19 \times \ln(x) + 360.08 + e$$

W ogólniejszej postaci równanie tego rodzaju miałoby następujący wygląd:

$$y = A \times x + B + e$$

równoważnie:

$$y = B + A \times x + e$$

Jest ono nazywane równaniem regresji liniowej, a wyrażenia A i B noszą nazwy współczynników lub parametrów regresji (zob. np.: Allen, 1997, 16–20; Dowdy et al., 2004, 201–208; Dunn & Clark, 1987, 261–264; Finkelstein & Levin, 2001, 358–359; McPherson,

2001, 519–522; Montgomery et al., 2008, 73–75; Ross, 2009, 353–354; Sen & Srivastava, 1990, 5–7; Sobczyk, 2015, 252–257; Stoodley et al., 1980, 36–37). Przedstawienie pewnej ilościowej zależności zaobserwowanej doświadczalnie w postaci równania regresji liniowej otwiera przed badaczami duże możliwości analityczne, gdzie jedną z nich jest analiza statystycznej istotności współczynników regresji (istotności trendu). Celem tej analizy jest ustalenie czy (i z jakim prawdopodobieństwem) relacja zaobserwowana i skwantyfikowana w ramach próby zachodzi także w ramach całości populacji, czy może jest ona tylko wynikiem przypadku i nie stanowi przejawu żadnego głębszego mechanizmu rządzącego jej strukturą lub cechami (zob. np.: Allen, 1997, 61–70, 106–112; Dowdy et al., 2004, 216–221, 224–226, 404–410; McClave & Benson, 1988, 506–509; McPherson, 2001, 536, 541–543; Oktaba, 1980, s. 330–333; Rawlings et al., 1998, 238–242; Ross, 2009, 363–377; Sen & Srivastava, 1990, 60–62; Sobczyk, 2015, 276–292). Jeżeli współczynniki równania regresji okażą się statystycznie istotne (np. na przyjmowanym często poziomie istotności równym 95%), równanie to może zostać wykorzystane do prognozowania (wnioskowania o zachodzeniu) wykrytych w ramach próby prawidłowości lub występowania w niej określonych cech bądź zjawisk, w obszarze całej szerokiej populacji generalnej.

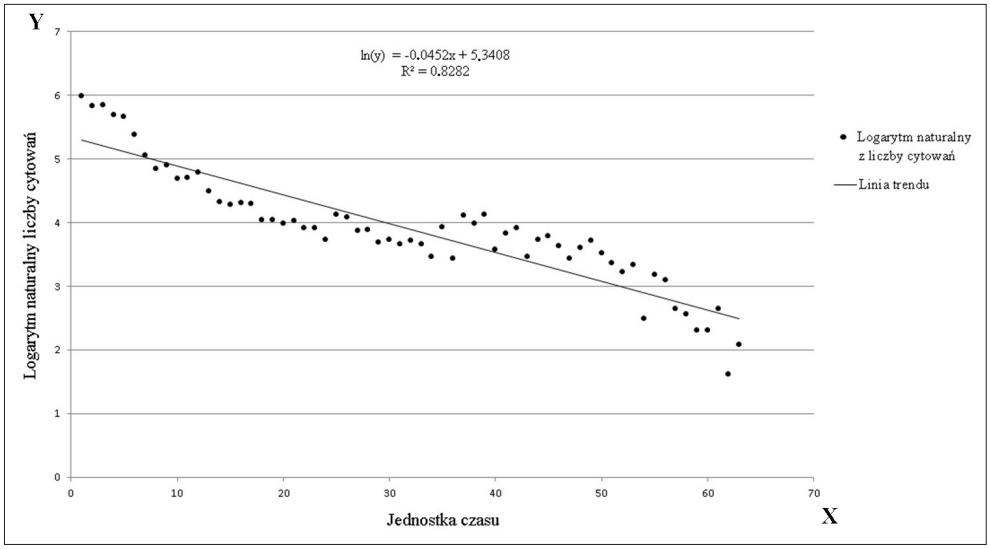
W przypadkach, w których ma się do czynienia z relacją nieliniową, podręczniki statystki zalecają niekiedy uprzednie wykonanie tzw. linearyzacji tej relacji (zob. np.: Allen, 1997, 123–127; Andersen et al., 1987, 303–310; Bingham & Fry, 2010, 168–173; Benoit, 2011; Haynes, 1982, 149; Ross, 2009, 383–386; Sachs, 1984, 453–455; Sheskin, 2007, 463–468; Stoodley et al., 1980, 42–46). Polega ona na przekształceniu trendu nieliniowego w liniowy poprzez odpowiednią transformację wartości liczbowych zmiennej zależnej, zmiennej niezależnej lub obu tych zmiennych W przypadku danych zilustrowanych przez rysunek 1 najkorzystniejsze jest zastosowanie transformacji logarytmicznej dla samej zmiennej zależnej y. Logarytmowanie zmiennej y oznacza, że każdą kolejną wartość tej zmiennej zależnej należy zamienić na jej logarytm, a następnie tak zmodyfikowane dane ponownie nanieść na wykres i dopasować do nich nowe równanie prostej (nową prostą regresji). Np. oryginalną wartością zmiennej y dla x równego 1 (tj. dla 2012 r.) było 400 cytowań. Logarytm z 400 wynosi ln(400) 5.991. Oryginalną wartością zmiennej y dla x równego 2 (2011 r.) było 343, a ln(343) 5.838 itd. Następstwem tego zabiegu jest niejako reorganizacja pierwotnej relacji i doprowadzenie jej do postaci liniowej, „wymuszenie" na niej postaci liniowej (zob. Rys. 2).

Po transformacji wyjściowe (nieliniowe) równanie trendu ($y = -89.19 \times \ln(x) + 360.08$) przybrało postać równania liniowego:

$$\ln(y) = -0.0452 \times x + 5.3408$$

Równanie to można poddać standardowej analizie regresji, którą umożliwia m.in. dodatek „Analiza danych" programu Microsoft Excel (zob. np.: Carlberg, 2012, 113–114, 117–129, 343–356; Snarska, 2011, 156–190; Winston, 2014, 590–598). Ocena istotności statystycznej równania regresji liniowej sprowadza się do oceny istotności parametrów regresji oraz wyznaczenia – najczęściej 95% – przedziałów ufności dla tych stałych. W tym celu należy podzielić wartość współczynnika regresji przez jego błąd standardowy. Otrzymany wynik ma wówczas tzw. rozkład t-Studenta dla $n - p - 1$ stopni swobody, gdzie n to liczba dostępnych obserwacji, a p – liczba predyktorów (współczynników regresji) zastosowanych w rozpatrywanym modelu (zob. np.: Aczel, 2007, 474–478, 533–539; Agarwal,

2009, 397–399; Allen, 1997, 66–70; Sobczyk, 2008, 170). Zbadanie tego aspektu równania regresji pozwoliłoby na wyciągnięcie lub odrzucenie wniosku, według którego zaobserwowana prawidłowość w zakresie zachowań informacyjnych i komunikacyjnych cytujących badaczy (a konkretniej mówiąc ich zachowań w zakresie cytowań – zob. np.: Opaliński et al., 2015) zachowuje ważność także poza granicami próby podanej analizie. Innymi słowy, gdyby analizie poddano całą zbiorowość (populację) generalną dokumentów, o której wspominano na wstępie niniejszego artykułu, można byłoby oczekiwać z 95% pewnością, że udałoby się zaobserwować wzorce pokrywające się z tymi, które zidentyfikowano w ramach badania próby losowej. W szczególności można byłoby w tej sytuacji oczekiwać, że wzorce te polegałyby na możliwości sformułowaniu analogicznego równania regresji liniowej, którego parametry A i B mieściłyby się w danym (zidentyfikowanym doświadczalnie) 95% przedziale ufności. Michael Allen (1997, 67) podaje, że dla stosunkowo dużych prób losowych rozkład t-Studenta wskazuje, że parametr regresji (liniowej) będzie statystycznie istotny na tym właśnie poziomie, jeżeli jego wartość przekroczy około dwukrotnie wartość jego błędu (odchylenia) standardowego.



Rys. 2. Przetransformowane liniowo dane z Rys. 1 z prostą regresji dopasowaną przez program Microsoft Excel 2010

Zanim jednak przejdziemy do właściwej analizy regresji, a w szczególności oceny jej statystycznej istotności, po której można przejść do analizy ukierunkowanej prognostycznie, należy upewnić się, że spełnione są pewne warunki, które stanowią podstawę wiarygodności (adekwatności) zaprojektowanego modelu do samej opisanej nim próby. Założenia te dotyczą rozkładu tzw. reszt modelu regresji liniowej i są określane mianem warunków Gaussa-Markova. Reszty modelu regresji liniowej są zdefiniowane jako różnice pomiędzy wartością empiryczną a oczekiwaną (wynikającą z przewidywań modelu) dla każdej z zarejestrowanych obserwacji. Symbolicznie resztę można zapisać jako:

$$e_i = y_i - \hat{y}_i,$$

gdzie y_i jest wartością zaobserwowaną, $\hat{y}_i$ jest wartością przewidywaną, a i = 1, 2, ..., n, przy czym n jest liczebnością przebadanej próby.

Pierwszy warunek Gaussa-Markova zawiera się w stwierdzeniu, że średnia arytmetyczna wartości reszt modelu jest równa zeru. Ponieważ niektóre z reszt przybierają wartości dodatnie (gdy wartość obserwowana jest większa od oczekiwanej), a inne ujemne (kiedy wartość obserwowana jest mniejsza od oczekiwanej) warunek ten wyraża przekonanie, że efekty wywołane przez wszystkie czynniki nieuwzględnione w równaniu regresji znoszą się wzajemnie, a co za tym idzie nie wpływają w istotny sposób na postać modelu.

Warunek drugi odnosi się do stabilności tzw. wariancji resztowej, tzn. niezależności jej skali od wartości przyjmowanych przez zmienną niezależną. Rozrzut wartości empirycznych (obserwacyjnych) wokół linii regresji powinien być w związku z tym podobny pod względem swojego poziomu ilościowego niezależnie od tego, jakie wartości przyjmuje zmienna niezależna. Pogwałcenie tego założenia oznaczałoby np., że wariancja systematycznie zwiększa się (bądź maleje) wraz ze wzrostem wartości zmiennej niezależnej. Jeżeli założenie o stabilności wariancji reszt modelu regresji jest spełnione określa się go wtedy jako model homoskedastyczny[3]. Gdy założenie to jest niespełnione – model jest tzw. modelem heteroskedastycznym.

Trzeci warunek mówi z kolei o tzw. wzajemnej niezależności reszt. Reszty są od siebie niezależne wtedy, gdy znajomość jednego z błędów obserwacji (tj. znajomość wartości pewnej konkretnej reszty) nie mówi nam nic na temat innych wartości reszt powiązanych z pozostałymi obserwacjami, czyli nie pozwala na wywnioskowanie wartości jakiejkolwiek innej reszty. Inaczej mówiąc wartości reszt są od siebie izolowane w tym sensie, że nie wpływają na siebie w żaden dający się wykryć sposób – skutki działania czynników przypadkowych wygasają, a nie są przenoszone na zasadzie echa na kolejne okresy realizacji pewnego procesu lub zjawiska. Warunek ten nazywa się również warunkiem braku zjawiska autokorelacji resztowej.

Czwarty i ostatni warunek dotyczy pewnej konkretnej postaci rozkładu reszt o którym zakłada się, że jest on tzw. rozkładem normalnym (nazywanym często rozkładem Gaussa – zob. np.: Taylor, 2011, 149–186). Zmienna (cecha elementów populacji) przyjmująca rozkład normalny charakteryzuje się symetrią rozkładu wokół swojej wartości średniej. Symetria ta oznacza, że najwięcej wyników pewnego doświadczenia losowego (w tym przypadku najwięcej poszczególnych, konkretnych wartości składnika resztowego modelu regresji) skupia się w okolicach wartości średniej arytmetycznej wszystkich wyników (reszt – tj. w okolicach wartości zerowej). Im bardziej wartości odbiegają od tej średniej tym mniejsza jest częstość ich występowania, przy czym dotyczy to w dokładnie takim samym stopniu wartości od średniej większych, jak i wartości od niej mniejszych. Zbadanie charakteru rozkładu reszt i zgodności tego rozkładu z rozkładem normalnym wymaga przeprowadzenia testu statystycznego. Może to być m.in. test chi-kwadrat (χ^2), test Kołmogorova-Smirnova, test Shapiro-Wilka lub test Lillieforsa (zob. np.: Allen, 1997, 181–185; Bingham & Fry, 2010, 163–168; Krzysztofiak & Luszniewicz, 1976, 390–394; McClave & Benson, 1988, 500–502, 601–613; McPherson, 2001, 523, 534–535; Montgomery et al., 2008, 98–100; Sheskin, 2007, 241–255, 279–280; Snarska, 2011, 248–252; Thode, 2002; Thode, 2011).

[3] Z gr. *skedastikos* – rozpraszający się.

W pierwszej kolejności, przed przystąpieniem do właściwej analizy statystycznej istotności parametrów regresji, należy zatem zweryfikować wspomniane cztery założenia w odniesieniu do konkretnego zbioru danych empirycznych, który ma być przedmiotem dalszego badania. W programie Microsoft Excel istnieje możliwość wyświetlenia składników resztowych modelu (regresji liniowej), który został skonstruowany przez funkcję dostępną w pakiecie „Analiza danych". Tabela 2 prezentuje składniki resztowe dla przedstawionego wcześniej przykładu.

Tab. 2. Składniki resztowe uzyskane dla równania regresji ln(y) = −0.0452 × x + 5.3408

Obser-wacja	Przewidywane „ln(y)"	Składniki resztowe	Obser-wacja	Przewidywane „ln(y)"	Składniki resztowe
1	5.295600465	0.695864082	33	3.848665966	-0.185104319
2	5.250383762	0.587346686	34	3.803449262	-0.33771336
3	5.205167058	0.635474599	35	3.758232559	0.173593073
4	5.159950355	0.537143131	36	3.713015856	-0.279028652
5	5.114733652	0.555147271	37	3.667799153	0.443074711
6	5.069516949	0.314978114	38	3.62258245	0.366401596
7	5.024300246	0.038294787	39	3.577365747	0.549768638
8	4.979083543	-0.134896457	40	3.532149044	0.051369895
9	4.93386684	-0.03602704	41	3.486932341	0.341709056
10	4.888650137	-0.197302255	42	3.441715638	0.470307368
11	4.843433434	-0.142953068	43	3.396498935	0.069236968
12	4.798216731	-0.002426185	44	3.351282232	0.386387387
13	4.753000028	-0.253190357	45	3.306065528	0.478124105
14	4.707783324	-0.377049984	46	3.260848825	0.376737334
15	4.662566621	-0.385900502	47	3.215632122	0.218355082
16	4.617349918	-0.313284825	48	3.170415419	0.440502494
17	4.572133215	-0.281673774	49	3.125198716	0.588373351
18	4.526916512	-0.483865244	50	3.079982013	0.446378512
19	4.481699809	-0.438648541	51	3.03476531	0.33253052
20	4.436483106	-0.447499059	52	2.989548607	0.229327218
21	4.391266403	-0.365914712	53	2.944331904	0.387872607
22	4.3460497	-0.434026694	54	2.899115201	-0.414208551
23	4.300832997	-0.388809991	55	2.853898497	0.324155333
24	4.255616293	-0.517946675	56	2.808681794	0.282360659
25	4.21039959	-0.083265205	57	2.763465091	-0.124407762
26	4.165182887	-0.087645443	58	2.718248388	-0.153299031
27	4.119966184	-0.248765173	59	2.673031685	-0.370446592
28	4.074749481	-0.182929183	60	2.627814982	-0.325229889
29	4.029532778	-0.340653324	61	2.582598279	0.056459051
30	3.984316075	-0.246646457	62	2.537381576	-0.927943663
31	3.939099372	-0.275537726	63	2.492164873	-0.412723331
32	3.893882669	-0.180310602			
Średnia arytmetyczna składników resztowych:		3.34829E−16 (= 3.34829×10^{-16} = $3.34829 \times \frac{1}{10^{16}}$ = 0.0000000000000000334829)			

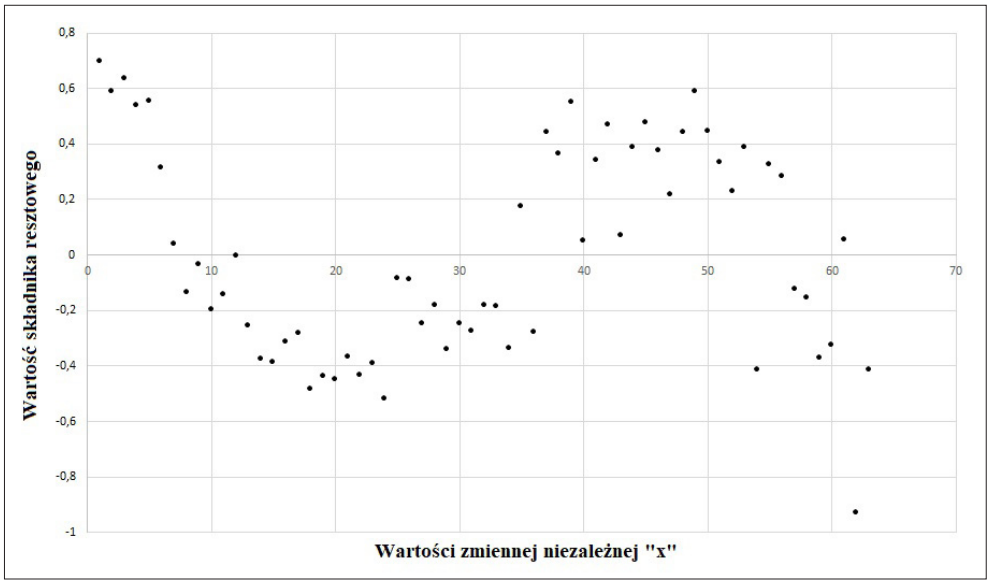
Dla przykładu, pierwsza reszta (0.695864082) została uzyskana w następujący sposób: w 2012 r. oznaczonym numerem 1 (tj. dla x = 1) odnotowano 400 cytowań; logarytm z liczby 400 jest równy w przybliżeniu 5.991464547 i jest to wartość zaobserwowana y_1; wartość $\hat{y}_1$ oblicza się z równania regresji jako: $\hat{y}_1 = -0{,}0452 \times 1 + 5.3408 = 5.2956$; reszta e_1 jest więc równa: $y_1 - \hat{y}_1 = 5.991464547 - 5.2956 = 0.695865$. Niewielka rozbieżność obu wartości pojawiająca się na szóstym miejscu po przecinku wynika z zaokrąglenia wartości ln(400), która jest w rzeczywistości wartością niewymierną i można przedstawić ją z dowolnie dużą liczbą miejsc po przecinku. Ponieważ wartość zaobserwowana była w tym przypadku większa niż wartość przewidywana składnik resztowy e_1 ma znak dodatni.

W odniesieniu do pierwszego założenia można stwierdzić, że wydaje się ono spełnione skoro średnia składników resztowych jest bardzo bliska zeru (zob. Tab. 2). Jest ona tak mała, że w rzeczywistości stanowi to przypuszczalnie wynik zaokrągleń wartości logarytmów naturalnych do dziewięciu miejsc po przecinku, które są konsekwencją wykorzystania w obliczeniach funkcji logarytmującej LN(...) programu Microsoft Excel. Ponadto nawet jeżeli przyjąć, że przyczyna, dla której wartość średniej reszt jest różna od zera, leży poza marginesem błędu obliczeń, niespełnienie tego założenia nie niesie ze sobą poważnych konsekwencji. Przejawiają się one bowiem jedynie w możliwym zniekształceniu wyrazu wolnego B równania regresji (Allen, 1997, 183). W przypadku cytowań ryzyko obejmowałoby więc tu metodologicznie nieuprawnione zaniżenie lub zawyżenie średniego, ilościowego poziomu cytowalności określonego zbioru publikacji. Zarazem samo tempo starzenia się zinterpretowane jako kąt, pod którym prosta regresji opada w kierunku poziomej osi wykresu, nie zostałoby w żaden sposób zdeformowane.

Wymogiem stawianym przed modelem regresji przez drugie założenie jest jego homoskedastyczność. Obecność tej cechy w przyjętym modelu regresji może być wstępnie oceniona przy pomocy wykresu pokazującego zależność wartości reszt od wartości zmiennej niezależnej. Rysunek 3 prezentuje wykres zależności wartości reszt modelu regresji od wartości zmiennej niezależnej dla omawianego przypadku. Rozkład ten i widoczny w jego ramach regularny, przypominający sinusoidę wzorzec wydaje się wskazywać, że model nie posiada cechy homoskedastyczności. W przypadku, w którym sam wykres nie daje wystarczająco jednoznacznej odpowiedzi, można dodatkowo posłużyć się tzw. testem Goldfelda-Quandta, polegającym na porównaniu wariancji składników resztowych w dwu „podpróbach" badanej próby (Snarska, 2011, 188–190). Przy jego użyciu testuje się hipotezę zerową, według której wariancje reszt w podpróbach są jednakowe (model jest homoskedastyczny) i przeciwstawną jej hipotezę alternatywną stwierdzającą, że wariancje te są różne, a model jest heteroskedastyczny. Metoda ta polega na porównaniu ilorazu wariancji dwóch podprób z wartościami krytycznymi tzw. rozkładu F na wybranym poziomie istotności. Podpróby te powinny przy tym łącznie zawierać około dwóch trzecich (66%) wszystkich elementów wchodzących w skład całej badanej zbiorowości (zob. np.: Crown, 1998, 84–85; Snarska, 2011, 188–190). W omawianym przypadku iloraz wariancji dwóch prób wyniósł w przybliżeniu 2.814. Fakt przekroczenia przez niego wartości krytycznej, która na poziomie istotności równym 95% wynosi w przybliżeniu 2.124, nakazał odrzucenie hipotezy zerowej na rzecz hipotezy alternatywnej. Model w rzeczywistości nie jest więc homoskedastyczny.

Trzeci warunek mówiący o braku zjawiska autokorelacji resztowej zazwyczaj bada się tzw. testem Durbina-Watsona. Polega on na wyznaczeniu statystyki testowej d, która opiera się na identyfikacji zjawiska korelacji pomiędzy sąsiednimi parami składników resztowych,

tj. pomiędzy składnikami resztowymi odpowiadającymi obserwacji nr 1 i obserwacji nr 2, obserwacji nr 2 i obserwacji nr 3 itd. (zob. np.: Krzysztofiak & Luszniewicz, 1976, 390–394; Sheskin, 2007, 1267–1268, 1273–1281; Snarska, 2011, 183–185). Statystykę d porównuje się następnie z dwiema podanymi w odpowiednich tablicach wartościami krytycznymi (tzw. górną i dolną), które pozwalają na podjęcie decyzji co do istnienia zjawiska autokorelacji lub jego braku. Dla danych zawartych w powyższym przykładzie wartość statystyki d wyniosła w przybliżeniu 0.599, co było wartością mniejszą niż dolna wartość krytyczna (równa w przybliżeniu 1.5) i wskazało na istnienie pozytywnej autokorelacji składników resztowych.

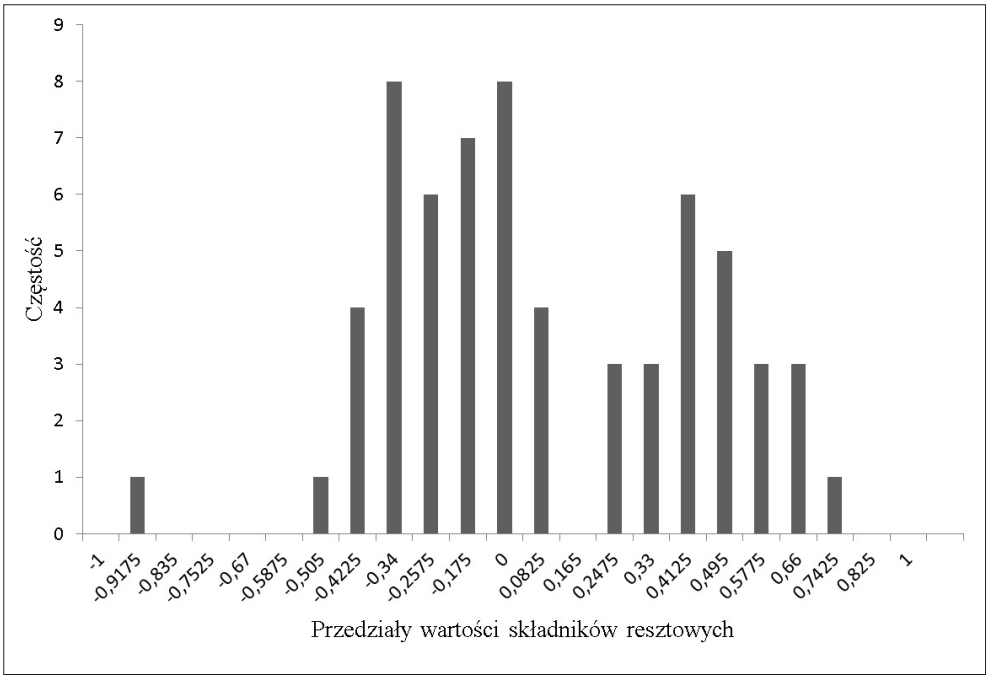


Rys. 3. Zależność wartości reszt modelu regresji od wartości zmiennej niezależnej

Ostatni warunek (normalność rozkładu składników resztowych ze średnią arytmetyczną równą 0) również może zostać wstępnie oceniony na podstawie graficznej prezentacji danych empirycznych. W omawianym przykładzie rozkład częstości składników resztowych, których wartości mieszczą się w określonych przedziałach, może zostać zilustrowany histogramem (zob. Rys. 4).

Histogram ten budzi pewne wątpliwości co do normalności zobrazowanego nim rozkładu ze względu na zauważalny brak symetrii wartości reszt usytuowanych wokół wartości średniej. Aby jednak ściślej zweryfikować tego rodzaju przypuszczenia, warto zastosować jeden z testów służących badaniu zgodności rozkładów empirycznych z teoretycznymi, taki jak np. test chi-kwadrat. Jego zasadą jest porównanie liczności wyników pewnego doświadczenia losowego, których wartości liczbowe mieszczą się w określonych przedziałach, z licznościami teoretycznymi tj. takimi, których należałoby w tych przedziałach oczekiwać w sytuacji, w której rzeczywiście mielibyśmy do czynienia z rozkładem normalnym. Wynikiem testu jest statystyka $\chi 2$, którą należy porównać z wartościami krytycznymi rozkładu chi-kwadrat na odpowiednim poziomie istotności. Dla rozpatrywanego przykładu statystyka ta przybrała wartość równą w przybliżeniu 32.47. Jest więc ona mniejsza niż

wynosząca 33.92 wartość krytyczna odczytana z tablic rozkładu chi-kwadrat dla 22 stopni swobody[4] i wartości $\alpha = 0.05$ (zob. np.: Sheskin, 2007, 1661). Dlatego oceniania w teście hipoteza zerowa mówiąca o braku istotnej różnicy między rozkładem zaobserwowanym (empirycznym), a rozkładem zakładanym (teoretycznym czyli rozkładem Gaussa) powinna zostać zaakceptowana na poziomie istotności równym 95%. Widoczne na histogramie odstępstwa od wzorca charakteryzującego rozkład normalny są więc wedle wszelkiego prawdopodobieństwa dziełem przypadku.



Rys. 4. Histogram demonstrujący rozkład składników resztowych w przyjętym modelu regresji

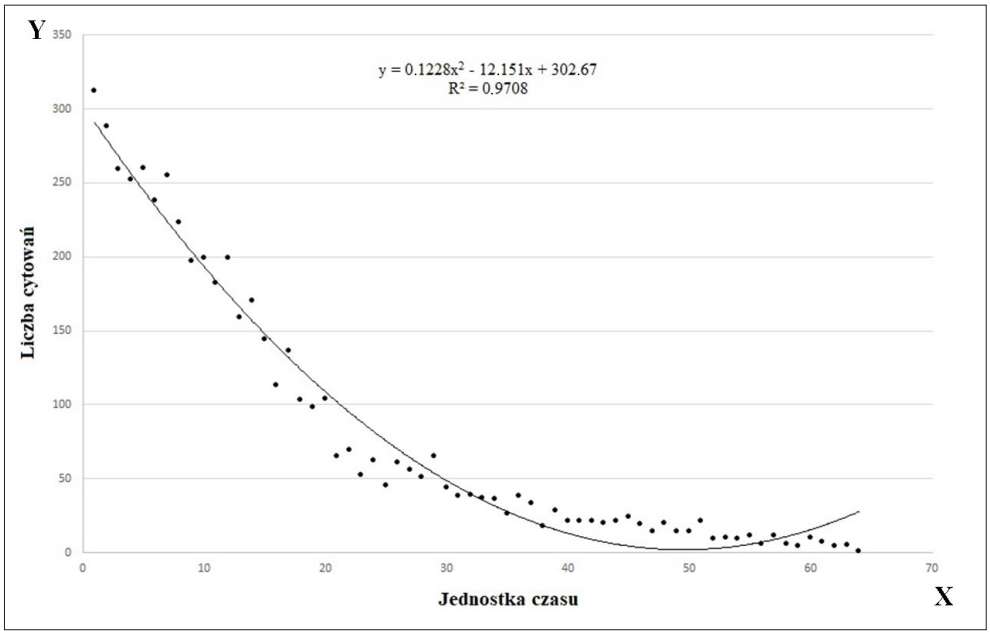
Ogółem więc dwa spośród czterech wymaganych założeń nie zostały spełnione w przypadku zaproponowanego powyżej modelu regresji liniowej. Sprawia to, że dalsza analiza regresji ukierunkowana na ocenę jej statystycznej istotności w ramach populacji generalnej staje się nieuzasadniona. Brakuje zatem podstaw nie tylko do tego, aby ekstrapolować zasięg obowiązywania wykrytej prawidłowości na całą populację generalną (w tym konkretnym przypadku – populację polskich publikacji z zakresu nauk o Ziemi), ale nawet do stwierdzenia, że model regresji wystarczająco wiernie opisuje rozkład wieku publikacji cytowanych w przebadanym korpusie literatury dziedzinowej.

[4] Liczba stopni swobody wynosi 22 ponieważ badany zakres danych podzielono na 25 przedziałów (szerzej na ten temat zob. np.: Opaliński et al., 2015, 75; Taylor, 2011, 316–319).

4. Regresja wielomianowa drugiego stopnia

Szczególnym przypadkiem tzw. wielokrotnej regresji liniowej, tj. takiej regresji liniowej, w której występuje więcej niż jedna zmienna niezależna (objaśniająca), jest regresja wielomianowa drugiego stopnia. W literaturze przedmiotu stwierdza się w związku z nią m.in., że zachowana zostaje tu ważność tych samych zasad konstrukcji i testowania istotności modeli, jak również warunków, które muszą być spełnione, aby można było zasadnie uogólnić model na całą populację, które wskazano wyżej podczas opisu metody analizy regresji liniowej (zob. np.: Allen, 1997, 127; Bingham & Fry, 2010, 99–103; Freud & Littell, 2000, 185; Haefner, 2012, 138; Huitema, 2011, 114–115; Oktaba, 1980, 325–326; Rawlings et al., 1998, 235–238; Ross, 2009, 393–396; Sachs, 1984, 447–449; Wetherill, 1981, 157–170). Może to być istotne w ramach niektórych analiz z uwagi na to, że niekiedy najlepszym dopasowaniem do danych empirycznych może okazać się właśnie parabola (tj. graficzna reprezentacja wielomianu drugiego stopnia). W badaniach własnych autorów niniejszego artykułu taką właśnie postać przybrała historia cytowań angielskojęzycznych źródeł czasopiśmienniczych. Zależność łącząca wiek i poziom cytowalności tego typu źródeł bibliograficznych wyglądała tak, jak pokazuje to rysunek 5. Dane doświadczalne będące podstawą konstrukcji tego rysunku zamieszczono w Aneksie 2.

W oparciu o możliwości obliczeniowe programu Microsoft Excel można było po pierwsze stwierdzić, że średnia arytmetyczna składników resztowych wyniosła tutaj –5.32907E–15 = $-5.32907 \times 10^{-15} = -5.32907 \times \frac{1}{10^{15}} = -0.00000000000000532907$. Jest to wielkość tak mała, że ponownie wydaje się, iż można przypisać ją niedokładności obliczeń wynikających z zaokrąglania parametrów modelu.



Rys. 5. Przebieg procesu starzenia się angielskojęzycznych źródeł czasopiśmienniczych w obrębie przebadanej próby

Otrzymana w ramach weryfikacji warunku nr 2 statystyka testowa Goldfelda-Quandta wyniosła w przybliżeniu 7.28 i przekroczyła wartość krytyczną równą w przybliżeniu 2.124 co świadczy o tym, że wariancje w dwu wyodrębnionych podpróbach różniły się, a model nie jest homoskedastyczny.

Wyznaczona w dalszej kolejności wartość statystyki d wyniosła w przybliżeniu 0.799, co ponownie było wartością mniejszą niż dolna wartość krytyczna dla testu Durbina-Watsona (równa w przybliżeniu 1.5). Wskazało to – analogicznie jak w przypadku funkcji regresji rozpatrywanej powyżej – na istnienie pozytywnej autokorelacji składników resztowych.

Statystyka testowa chi-kwadrat wykorzystania podczas badania normalności rozkładu składników resztowych przybrała wartość równą w przybliżeniu 35.93. Była ona większa niż wartość krytyczna rozkładu chi-kwadrat dla 22 stopni swobody i poziomu istotności 95% (tj. dla $\alpha = 0{,}05$), która jest równa 33.92. Hipoteza zerowa o braku istotnej różnicy między rozkładem reszt modelu a rozkładem normalnym nie mogła wobec tego zostać zaakceptowana, co oznacza, iż rozkład reszt modelu w istotnym stopniu różnił się od rozkładu normalnego.

5. Podsumowanie analiz przeprowadzonych dla wszystkich typów form wydawniczych publikacji cytowanych w obrębie przebadanej próby

W powyżej rozpatrzonym przypadku, podobnie jak miało to miejsce wcześniej, zabrakło dostatecznych podstaw do merytorycznego uzasadnienia generalizacji wielomianowej (parabolicznej) funkcji trendu wykrytej w obrębie przeanalizowanej próby. W dalszej kolejności przeanalizowano metodą regresji pozostałe, zidentyfikowane w badaniu własnym pierwszego z autorów niniejszego artykułu, cykle życiowe cytowanej literatury metodą nieróżniącą się od metody omówionej i zademonstrowanej na obu powyższych przykładach. Podsumowanie efektów wspomnianej analizy zawarto w tabeli 3. Dane empiryczne, które posłużyły za podstawę tej analizy, zostały wyszczególnione w Aneksie 2.

Jak można zauważyć, jedyne założenie, które konsekwentnie nie zostało spełnione w żadnym przeanalizowanym przypadku to założenie nr 2, dotyczące homoskedastyczności modelu regresji. Głównym skutkiem niespełnienia tego założenia jest zniekształcenie (niedoszacowanie lub „przeszacowanie") tzw. błędu standardowego (nazywanego też odchyleniem standardowym) oceny współczynnika regresji A występującego w równaniu regresji i wyznaczonego na podstawie próby losowej. Prowadzi to do ryzyka, że w trakcie statystycznych testów istotności tego współczynnika (oraz podczas wyznaczania przedziałów ufności) otrzyma się wynik wskazujący na to, że jest on w statystycznie istotny sposób różny od zera, podczas gdy w rzeczywistości będzie to nieprawdą. Ponadto w takiej sytuacji zaburzeniom ulegają typowe wskaźniki precyzji dopasowania linii (funkcji) regresji do danych empirycznych, takie jak np. współczynnik determinacji R2, które bazują na próbie losowej. Różne próby cechują się bowiem wtedy różną skalą wariancji swoich składników resztowych, przez co różne są także stopnie ich zgodności z modelem lub jego przewidywaniami. Za możliwe przyczyny heteroskedastyczności składników resztowych modeli regresji uważa się m.in. obecność obserwacji odstających (ang. *outliers*) w zbiorach danych tworzących próby losowe, szczególnie w przypadku małych prób, niewłaściwie dobrane parametry równania regresji, skośność rozkładu jednej lub więcej zmiennych objaśniających (niezależnych), błędną transformację danych empirycznych bądź też nieuwzględnienie

pewnych zmiennych niezależnych w równaniu regresji (Allen, 1997, 183; Andersen et al., 1987, 347; Bucevska, 2011, 631; Finkelstein & Levin, 2001, 404; Sen & Srivastava, 1990, 111–114, 122). Niektóre podręczniki zalecają tzw. standaryzację składników resztowych przed przystąpieniem do ich analizy ze względu na fakt, że reszty niestandaryzowane często wykazują cechę heteroskedastyczności (zob. np.: Christensen, 2011, 304, 361–370; Montgomery et al., 2008, 100–103), chociaż w przypadku danych omawianych w niniejszym artykule rozróżnienie to okazało się nieistotne – w przypadku wszystkich form wydawniczych reszty standaryzowane również wykazywały cechę heteroskedastyczności.

Tabela 3. Wyniki analizy regresji dla cykli życiowych cechujących rozpatrzone formy wydawnicze dokumentów cytowanych w obrębie zbadanej próby

Forma wydawnicza	Język publikacji	Równanie regresji	Równanie regresji po transformacji	Średnia składników resztowych	Wartość statystyki Goldfelda-Quandta	Wartość statystyki „d"	Wartość statystyki chi-kwadrat
Wydawnictwa zwarte (bez materiałów konferencyjnych)	pol.	y = –68.62 ln(x) + 273.85	ln(y) = –0.0669x +5.5154	6.08708 × 10^{-16}	13.2 (> 0.44)	0.637 (< 1.5)	313.05* (> 28.87)
Wydawnictwa zwarte (bez materiałów konferencyjnych)	ang.	y = –23.05 ln(x) + 93.228	ln(y) = –0.0621x + 4.3912	7.58652 × 10^{-16}	9.95 (> 0.43)	2.19 (> 1.6)**	32.46 (< 35.17)
Materiały konferencyjne	pol.	y = –16.09 ln(x) + 59.898	ln(y) = –0.1012x + 4.0513	–8.21565 × 10–15	26.73 (> 0.29)	1.93 (> 1.48)**	28.3 (> 22.36)***
Materiały konferencyjne	ang.	y = –13.98 ln(x) + 56.261	ln(y) = –0.0713x + 3.997	–1.07037 × 10^{-15}	12.1 (> 0.34)	1.97 (> 1.54)**	33.35 (> 30.14)***
Kategoria „inne" (tzw. zbiory specjalne)	pol.	y = –63.28 ln(x) + 247.74	ln(y) = –0.0694x + 5,2497	–1.42109 × 10^{-14}	5.3 (> 0.45)	1.686 (> 1.62)**	27.94 (< 30.14)
Kategoria „inne" (tzw. zbiory specjalne)	ang.	y = –17.4 ln(x) + 67.829	ln(y) = –0.0886x + 4.1954	–2.86321 × 10^{-16}	6.95 (> 0.39)	1.429 (< 1.43)	10.28 (< 30.14)

* Tak duża wartość statystyki testowej jest wynikiem wystąpienia w próbie dwóch wartości odstających (ang. *outliers*), tj. wartości znacznie odbiegających nie tylko od średniej arytmetycznej próby, ale i od całego zbioru wartości przyjmowanych przez elementy wchodzące w jej skład. Na poziomie danych o cytowaniach można stwierdzić, że wartości skrajne pojawiły się jako konsekwencja faktu, iż w zbadanej próbie artykułów cytujących powołano się na dokładnie 1 publikację wydaną w 1951 r. i 1 wydaną w 1952 r. Sprawiło to, ze logarytmy zmiennej zależnej y dla tych dwóch lat wyniosły 0. Po wykluczeniu ze zbioru danych wyjściowych obu wartości zerowych statystyka chi-kwadrat wyniosła 24.73 w związku z czym można przyjąć, że w tej sytuacji rozkład empiryczny jest w rzeczywistości zgodny z rozkładem normalnym.

** W tym przypadku zjawisko autokorelacji składników resztowych nie występuje.

*** Rozkład składników resztowych nie jest tu zgodny z rozkładem normalnym.

W statystyce istnieją wprawdzie sposoby usuwania tej cechy z modeli regresji liniowej (może to być m.in. wprowadzenie tzw. ważenia odchyleń wartości empirycznych od przewidywanych w trakcie wyznaczania parametrów modelu), wydaje się jednak, że są one z jednej strony nieco sztuczne (tj. wymagają znaczącej ingerencji w dane doświadczalne), z drugiej zaś – ich konsekwencją są kolejne transformacje modelu, takie jak np. pojawienie się kolejnej zmiennej niezależnej w modelu pierwotnie zawierającym tylko jedną zmienną niezależną (model przekształca się więc z modelu regresji liniowej w model regresji wieloliniowej). O ile zatem jedna zmienna niezależna, którą w powyższych przykładach zawsze jest czas (wiek źródeł cytowanych) posiada swoją naturalną interpretację i znaczenie, o tyle pojawienie się kolejnej zmiennej byłoby już trudniejsze do naukoznawczego uzasadnienia i zinterpretowania. Ponadto, po dokonaniu odpowiednich zmian w formule modelu, należałoby ponownie upewnić się, czy fakt spełnienia przez model wszystkich pozostałych założeń nie uległ w ich wyniku zmianie lub zaburzeniom. Jeszcze inną ewentualnością byłoby wykorzystanie faktu, że liczby cytowań odsyłających do publikacji liczących sobie około 20 i więcej lat stają się w ramach zebranego przez autorów zbioru danych coraz bardziej nieregularne i nieliczne. Przypuszczalnie to właśnie one są w dużej części odpowiedzialne za wzrost wariancji (niedokładności oszacowania linii trendu) występujący wraz ze wzrostem wieku publikacji cytowanych co nasuwa myśl, że skrócenie okresu obserwacji cytowań mogłoby stanowić remedium na problem niejednorodności wariancji modelu. Ponownie wydaje się jednak, że sztuczne skrócenie szeregu cytowań do np. wyłącznie tych, które zaobserwowano dla prac 20- lub 25-letnich i młodszych (wydanych wcześniej) byłoby niepożądane z punktu widzenia możliwości naukoznawczego zinterpretowania modelu. Inaczej mówiąc, trudno byłoby wtedy uzasadnić taki zabieg jakimikolwiek czynnikami niemającymi związku wyłącznie z dokonaną *ad hoc* modyfikacją danych, aby nadać im pożądany kształt i dostosować je do własnych, indywidualnych potrzeb.

6. Zakończenie

Ostatecznie należy więc stwierdzić, że kwestia identyfikacji i ewentualnej eliminacji przyczyn heteroskedastyczności modeli, jak również kwestia skonstruowania modeli o wyższym stopniu ogólności (tj. modeli dających się generalizować na całą populację generalną publikacji naukowych z zakresu wybranej dyscypliny) wymaga przyjęcia odmiennej perspektywy badawczej, w ramach której – jak się wydaje – należy wykorzystać bardziej zaawansowane techniki statystyki matematycznej. W omówionych dotychczas przypadkach wysuwanie jakichkolwiek prognoz w oparciu o modele typu regresyjnego byłoby ryzykowne i nieuprawnione z wskazanych powyżej przyczyn. Ocena powodów takiego stanu rzeczy mogłaby odwoływać się np. do niedostatków metodologicznych zaproponowanego aparatu statystycznego, w rodzaju hipotetycznie niewłaściwych technik gromadzenia danych empirycznych. Przykładowo, w przypadku celowego doboru elementów próby – w przeciwieństwie do doboru całkowicie losowego – mamy zawsze do czynienia z jakimiś kryteriami, które mogą być postrzegane jako zmienne wpływające na wyniki badania i jego dalszą ekstrapolację (i tak powinny one być traktowane podczas ich jakościowej interpretacji). Pomimo że w przedstawionych powyżej rozważaniach mamy rzeczywiście do czynienia z pewnymi konkretnymi kryteriami selekcji elementów przebadanej próby cytowań i publikacji

naukowych, co może poddawać w wątpliwość jej faktyczną losowość, zdaniem autorów artykułu w dalszym ciągu możliwe jest jednak traktowanie jej jako próby losowej (lub co najmniej quasi-losowej) z uwagi na fakt, że kryteria te zostały zakreślone bardzo ściśle i wpłynęły znacząco na pominięcie bardzo dużej części dziedzinowego piśmiennictwa. Ponadto wydaje się, że dla celów demonstracyjnych oraz ze względu na potrzebę wykazania możliwości zastosowania prezentowanych powyżej technik w znacznie szerszej gamie przypadków badań empirycznych, niż miało to miejsce w przypadku badania faktycznie zrealizowanego przez autorów niniejszej pracy, warto (nawet gdyby miało to zostać uznane za posunięcie w pewnej mierze arbitralne) potraktować zebrane dane jako dużą próbę losową, tj. próbę, w której zastosowane kryteria selekcji źródeł piśmienniczych mogłyby zostać w określony sposób zmodyfikowane, rozszerzone, bądź nawet pominięte, przez co nie przekreślają one całkowicie jej zakładanej losowości. Naturalnie, rolę mogły odegrać tu również inne, trudniejsze do zidentyfikowania czynniki psychologiczne, socjologiczne lub – ogólniej – związane ze specyfiką, niepowtarzalnością czy swoistością zachowań w zakresie cytowań społeczności naukowców, których publikacje złożyły się na przebadaną przez autorów próbę losową. Należy bowiem pamiętać, że w ostatecznym rozrachunku rozkłady cytowań i ich wszelkie jakościowe i ilościowe cechy są determinowane przez praktyki wielu indywidualnych naukowców, którzy powołują się w swoich pracach na publikacje swoich poprzedników. W sferze wyznaczonej przez zachowania, postawy, praktyki i wzorce postępowania człowieka, metody ilościowe nie zawsze są w stanie zapewnić dostateczny i w pełni uzasadniony opis efektów tych praktyk i – jak wskazuje się w literaturze przedmiotu – zastosowanie w tym zakresie znajdują często metody jakościowe (zob. np.: Stefaniak et al., 2016, 117–121). Z drugiej strony, warto też z pewnością przetestować inne, pominięte w niniejszym opracowaniu techniki statystyczne, które być może byłyby w stanie dostarczyć wyników ilościowych o akceptowalnym stopniu precyzji i mogących, w związku z tym, stanowić solidniejszy fundament dla naukoznawczego prognozowania zjawiska rozwoju nauki i jej poszczególnych obszarów. Zadanie to i próba jego realizacji jest przedmiotem drugiej części omówienia, które zawarte jest w kolejnym artykule, stanowiącym rozwinięcie i kontynuację niniejszego.

Bibliografia

Aczel, A. D. (2007). *Statystyka w zarządzaniu: pełny wykład*. Warszawa: PWN.

Agarwal, B. L. (2009). *Basic Statistics*. New Delhi: New Age International.

Allen, M. P. (1997). *Understanding Regression Analysis*. New York: Plenum Press, https://doi.org/10.1007/b102242

Andersen, E. B., Jensen, N. E., Kousgaard, N. (1987). *Statistics for Economics, Business Administration, and the Social Sciences*. Berlin: Springer, https://doi.org/10.1007/978–3-642–95528-0

Benoit, K. (2011). *Linear Regression Models with Logarithmic Transformations* [online]. London School of Economics, [19.11.2019], https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

Bensman, S. J. (2000). Probability Distributions in Library and Information Science: A Historical and Practicioner Viewpoint. *Journal of the American Society for Information Science*, 51(9), 816–833, https://doi.org/10.1002/(SICI)1097-4571(2000)51:9<816::AID-ASI50>3.0.CO;2-6

Berk, K. N., Carey, P. (2010). *Data Analysis with Microsoft Excel. Updated for Office 2007*. Boston: Brooks/Cole: Cengage Learning.

Bingham, N., Fry, J. (2010). *Regression: Linear Models in Statistics*. New York, London: Springer, https://doi.org/10.1007/978-1-84882-969-5

Borgman, Ch. L., Furner, J. (2002). Scholarly Communication and Bibliometrics. *Annual Review of Information Science & Technology*, 36(1), 3–72, https://doi.org/10.1002/aris.1440360102

Bucevska, V. (2011). Heteroscedasticity. In: M. Lovric (ed.). *International Encyclopedia of Statistical Science* (630–633). Berlin: Heidelberg: Springer, https://doi.org/10.1007/978-3-642-04898-2_628

Burton, R. E., Kebler, R. W. (1960). The 'Half-Life' of Some Scientific and Technical Literatures. *American Documentation*, 11(1), 18–22, https://doi.org/10.1002/asi.5090110105

Carlberg, C. (2012). *Analiza statystyczna. Microsoft Excel 2010 PL*. Gliwice: Helion.

Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models*. New York: Springer, https://doi.org/10.1007/978-1-4419-9816-3

Crown, W. H. (1998). *Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-dependent Variable Models*. Westport, Conn.: Praeger.

Dowdy, S., Wearden, S., Chilko, D. (2004). *Statistics for Research*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Dunn, O. J., Clark, V. (1987). *Applied Statistics: Analysis of Variance and Regression*. New York: Chichester [etc.]: John Wiley and Sons.

Finkelstein, M. O., Levin, B. (2001). *Statistics for Lawyers*. New York: Springer, https://doi.org/10.1007/b97319

Freud, R. J., Littell, R. C. (2000). *SAS System for Regression*. Cary (North Carolina): SAS Institute.

Haefner, J. W. (2012). *Modeling Biological Systems: Principles and Applications*. Dordrecht: Springer Science & Business Media, https://doi.org/10.1007/b106568

Haynes, R. M. (1982). *Environmental Science Methods*. London: New York: Chapman and Hall.

Huitema, B. E. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-case Studies*. Hoboken, NJ: Wiley & Sons, https://doi.org/10.1002/9781118067475

Jarić, I., Knežević-Jarić, J., Lenhardt, M. (2014). Relative Age of References as a Tool to Identify Emerging Research Fields With an Application to the Field of Ecology and Environmental Sciences. *Scientometrics*, 100(2), 519–529, https://doi.org/10.1007/s11192-014-1268-9

Krzysztofiak, M., Luszniewicz, A. (1976). *Statystyka*. Warszawa: Polskie Wydaw. Ekonomiczne.

Larocque, D., Randles, R. (2008). Confidence Intervals for a Discrete Population Median. *American Statistician*, 62(1), 32–39, https://doi.org/10.1198/000313008X269738

McClave, J. T., Benson, G. (1988). *Statistics for Business and Economics*. San Francisco: Dellen Pub. Co., London: Collier Macmillan.

McPherson, G. (2001). *Applying and Interpreting Statistics: A Comprehensive Guide*. New York: Springer, https://doi.org/10.1007/978-1-4757-3435-5

Montgomery, D. C., Jennings, Ch., Kulahci, M. (2008). *Forecasting and Time Series Analysis*. New York: Wiley.

Oktaba, W. (1980). *Metody statystyki matematycznej w doświadczalnictwie*. Warszawa: PWN.

Opaliński, Ł. (2013). Wybrane aspekty metodologii badań cyklu życiowego publikacji naukowych. *Przegląd Biblioteczny*, 81(2), 152–171.

Opaliński, Ł. (2017a). Bibliometryczna metodologia prognozowania i oceny rozwoju dyscyplin naukowych. Analiza piśmiennictwa. Część I. Publikacje pionierskie, metoda powiązań bibliograficznych, metoda współcytowań i metoda współwystępowania specjalistycznej terminologii naukowej. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 55(1), 34–65.

Opaliński, Ł. (2017b). Bibliometryczna metodologia prognozowania i oceny rozwoju dyscyplin naukowych. Analiza piśmiennictwa. Część 2. Badania porównawcze, hybrydowe, statystyczne, analizy dokumentów patentowych, ścieżek rozwoju dyscyplin oraz pozostałe oryginalne podejścia metodologiczne. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 55(2), 73–105.

Opaliński, Ł., Jaromin, M. (2017). Zastosowanie statystycznej analizy szeregów czasowych do krótkoterminowego prognozowania rozwoju dyscyplin naukowych. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 55(2), 106–125.

Opaliński, Ł., Jaromin, M., Wikiera, J. (2015). Problem stabilności zachowań naukowców w zakresie cytowań w kontekście metodologii badań starzenia się publikacji naukowych i możliwość jego ujęcia ilościowego. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 53(2), 65–83.

Ott, L., Longnecker, M. (2010). *An Introduction to Statistical Methods and Data Analysis*. Belmont, CA: Brooks/Cole: Cengage Learning.

Rawlings, J. O., Pantula, S. G., Dickey, D. A. (1998) *Applied Regression Analysis: A Research Tool*. Berlin: Springer, https://doi.org/10.1007/b98890

Ross, S. M. (2009). *Introduction to Probability and Statistics for Engineers and Scientists*. Amsterdam: Elsevier Academic Press, https://doi.org/10.1016/B978-0-12-370483-2.X0001-X

Rousseau, R. (2006). Timelines in Citation Research. *Journal of the American Society for Information Science and Technology*, 57(10), 1404–1405, https://doi.org/10.1002/asi.20397

Sachs, L. (1984). *Applied Statistics: A Handbook of Techniques*. Berlin: Springer, https://doi.org/10.1007/978-1-4612-5246-7

Sen, B. K. (1999). Symbols and Formulas for a Few Bibliometric Concepts. *Journal of Documentation*, 55(3), 325–334, https://doi.org/10.1108/EUM0000000007149

Sen, A., Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications*. Berlin: Heidelberg: Springer, https://doi.org/10.1007/978-1-4612-4470-7

Shapiro, F. R. (1992). Origins of Bibliometrics, Citation Indexing, and Citation Analysis: The Neglected Legal Literature. *Journal of the American Society for Information Science*, 43(5), 337–339, https://doi.org/10.1002/(SICI)1097-4571(199206)43:5<337::AID-ASI2>3.0.CO;2-T

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures. Fourth Edition*. Boca Raton: London: New York: Chapman & Hall/CRC, Taylor & Francis Group.

Snarska, A. (2011). *Statystyka, ekonometria, prognozowanie: ćwiczenia z Excelem 2007*. Warszawa: Wydaw. Placet.

Sobczyk, M. (2008). *Prognozowanie. Teoria, przykłady, zadania*. Warszawa: Wydaw. Placet.

Sobczyk, M. (2015). *Statystyka*. Warszawa: PWN.

Sosińska-Kalata, B., Roszkowski, M. (2016). Organizacja informacji i wiedzy. W: W. Babik (red.). *Nauka o informacji* (305–357). Warszawa: Wydaw. SBP.

Stefaniak, B., Skalska-Zlat, M., Cisek, S. (2016). Metody badań w nauce o informacji. W: W. Babik (red.). *Nauka o informacji* (89–122). Warszawa: Wydaw. SBP.

Stoodley, K. D. C., Lewis, T., Stainton, C. L. S. (1980). *Applied Statistical Techniques*. Chichester: Ellis Horwood.

Taylor, J. R. (2011). *Wstęp do analizy błędu pomiarowego*. Warszawa: PWN.

Thode, H. C. (2002). *Testing for Normality*. New York: Marcel Dekker.

Thode, H. C. (2011). Normality Tests. In: M. Lovric (ed*.). International Encyclopedia of Statistical Science* (1000–1002). Berlin: Heidelberg: Springer, https://doi.org/10.1007/978-3-642-04898-2_423

Vaughan, L. (2003). *Statistical Methods for the Information Professional: A Practical, Painless Approach to Understanding, Using, and Interpreting Statistics*. Medford, New Yersey: Information Today, Inc.

Vinkler, P. (1996). Relationships Between the Rate of Scientific Development and Citations. The Chance for Citedness Model. *Scientometrics*, 35(3), 375–386, https://doi.org/10.1007/BF02016908

Wetherill, G. B. (1981). *Intermediate Statistical Methods*. London: New York: Springer Netherlands, https://doi.org/10.1007/978-94-009-5836-4

Winston, W. L. (2014). *Microsoft Excel 2013: analiza i modelowanie danych biznesowych*. Warszawa: APN Promise.

Aneksy

Aneks 1. *Wykaz czasopism stanowiących materiał badawczy i źródło zarejestrowanych w badaniu przypisów bibliograficznych* [online]. Figshare repository, [03.06.2020], https://doi.org/10.6084/m9.figshare.11188274.v1

Aneks 2. *Liczby cytowań zarejestrowane w badaniu empirycznym z rozróżnieniem na poszczególne typy wydawnicze* [online]. Figshare repository, [03.06.2020], https://doi.org/10.6084/m9.figshare.11189228.v1

Selected Methods of Forecasting the Rate of Scientific Disciplines' Development (Citing Half-Life Index, Nonlinear Regression Method, Linearized Regression Method and Second-Degree Polynomial Regression)

Abstract

Purpose/Thesis: The study presents an overview and a comparison of several selected statistical methods basing on citations obtained from publications belonging to the selected disciplines. Furthermore, the study analyzed the requirements for and obstacles to generalizing quantitative results yielded on the basis on the random samples.

Approach/Methods: On the basis of the data set comprising of almost 25 thousands of citations, the authors have shown a method of establishing confidence intervals for the *citing half-life* indexes; then the data was subjected to a statistical study using nonlinear regression method, linearized regression method, and a second-degree polynomial regression.

Results and conclusions: The central difficulty of applying these methods was their failure to fulfill some of the Gauss-Markov conditions. It is necessary to correct the models applied, or to make use of statistical techniques of a different kind, which suggests future research directions.

Originality/Value: The originality of the presented overview lies in the novel juxtaposition of the quantitative methods, which, although well-known, are not commonly used to forecast the rate of the development of scientific disciplines The study highlighted their potential and expected usefulness in this regard, as well as the need of further improvement or testing other, more sophisticated methods.

Keywords

Bibliometrics. Development of science. Forecasting methods. Scientific communication. Scientific domains and areas. Scientometrics. Statistics in information science.

Dr ŁUKASZ OPALIŃSKI uzyskał tytuł doktora w zakresie nauk humanistycznych w dyscyplinie bibliologia i informatologia, nadany w grudniu 2018 r. przez Radę Wydziału Zarządzania i Komunikacji Społecznej Uniwersytetu Jagiellońskiego, na podstawie rozprawy pt.: „Starzenie się publikacji naukowych w języku polskim i angielskim w perspektywie zachowań warunkujących proces cytowania w naukach o Ziemi". Pracuje w Oddziale Informacji Naukowej Biblioteki Politechniki Rzeszowskiej na stanowisku kustosza. Najważniejsze publikacje: (1) Opaliński, Ł. (2019). Cytowanie narzędziem zarządzania informacją: teoria zachowań informacyjnych. W: W. Babik (red.) Zarządzanie informacją *(210–248). Warszawa: Wydaw. SBP. (2) Opaliński, Ł., Jaromin, M. (2017). Zastosowanie statystycznej analizy szeregów czasowych do krótkoterminowego prognozowania rozwoju dyscyplin naukowych.* Zagadnienia Informacji Naukowej – Studia Informacyjne*, 55(2), 106–125.*

Rola w przygotowaniu artykułu: opracowanie koncepcji artykułu, części teoretycznej, analiza literatury przedmiotu, opracowanie wykresów, tabel i aneksów, zebranie danych empirycznych i interpretacja wyników badania. Udział: 50%.

Kontakt z autorem:
lopa@prz.edu.pl
Biblioteka Politechniki Rzeszowskiej
Oddział Informacji Naukowej
al. Powstańców Warszawy 12
35-959 Rzeszów

MARCIN JAROMIN pracuje na stanowisku asystenta w grupie pracowników badawczo-dydaktycznych w Zakładzie Biotechnologii i Bioinformatyki Politechniki Rzeszowskiej. Tytuł magistra inżyniera uzyskał w 2004 r. na Wydziale Chemicznym Politechniki Rzeszowskiej oraz, równolegle, w 2005 r. na Wydziale Elektrotechniki i Informatyki Politechniki Rzeszowskiej. Specjalizuje się w dziedzinie bioinformatyki i statystyki matematycznej. Najważniejsze publikacje: Bocian A., Buczkowicz, J., Jaromin, M., Hus, K. K., Legáth, J. (2019). An Effective Method of Isolating Honey Proteins. Molecules*, 24(13), 2399; Ciura, J., Bocian, A., Kononiuk, A., Szeliga, M., Jaromin, M., Tyrka, M. (2017). Proteomic Signature of Fenugreek Treated by Methyl Jasmonate and Cholesterol.* Acta Physiologiae Plantarum*, 39, 112.*

Rola w przygotowaniu artykułu: analiza statystyczna danych empirycznych. Udział: 50%.

Kontakt z autorem:
mjaromin@prz.edu.pl
Zakład Biotechnologii i Bioinformatyki
Wydział Chemiczny
Politechnika Rzeszowska
al. Powstańców Warszawy 6
35-959 Rzeszów