Wybrane metody prognozowania tempa rozwoju dyscyplin naukowych (metoda regresji wielomianowej trzeciego stopnia, metoda autoregresji oraz wygładzania wykładniczego)

Łukasz Opaliński
ORCID 0000-0003-2797-2777
Oddział Informacji Naukowej Biblioteki Politechniki Rzeszowskiej
Marcin Jaromin
ORCID 0000-0002-7256-7271
Zakład Biotechnologii i Bioinformatyki Politechniki Rzeszowskiej

Abstrakt

Cel/Teza: W artykule porównano statystyczne metody prognozowania tempa ewolucji dyscyplin naukowych. Materiałem empirycznym były cytowania uzyskiwane przez publikacje. Zaakcentowano możliwość uogólnienia wyników badań prób losowych na szerszą populację generalną. Wskazano problemy, na jakie napotyka każda z wybranych metod i zaproponowano szkicowo potencjalne sposoby ich przezwyciężenia.
Koncepcja/Metody badań: Do zbioru danych empirycznych, na który złożyło się prawie 25 tys. cytowań, zastosowano metody inspirowane modelami ekonometrycznymi, tj. metodę regresji wielomianowej, metodę regresji z poprawką ze względu na autokorelację składników resztowych, autoregresję, autoregresję z korektą niestacjonarności modelowanego procesu oraz adaptacyjny model wygładzania wykładniczego Holta. Dla metod regresji zbadano fakt spełniania przez nie warunków Gaussa-Markova. Sprawdzono także statystyczne wskaźniki precyzji dopasowania modeli do danych doświadczalnych, jak również współczynniki dokładności skonstruowanych prognoz.
Wyniki i wnioski: Za najdokładniejszą metodę prognostyczną należy uznać, w świetle dostępnych dla autorów danych, metodę regresji wielomianowej z poprawką ze względu na autokorelację składników resztowych. Metody autoregresyjne wydają się porównywalne z metodami regresyjnymi, metoda adaptacyjna dała natomiast wyniki niejednoznaczne. Fakt ten stanowi zarazem perspektywę dalszych badań.
Ograniczenia badań: Podstawowym ograniczeniem jest dostępny autorom zakres danych empirycznych, które objęły tylko jedną dziedzinę nauki, a dodatkowo zostały zawężone do jej polskojęzycznej sfery oraz do źródeł czasopiśmienniczych.
Oryginalność/Wartość poznawcza: Zestawiono z sobą metody ilościowe, które nie są powszechnie stosowane w celu ewaluacji tempa rozwoju nauki. Zademonstrowano ich potencjał w tym względzie, oraz zaznaczono potrzebę dalszego ich doskonalenia. Wytypowanie najbardziej obiecującej metodologii powinno przyczynić się do lepszego zrozumienia wewnętrznej dynamiki nauki.

Słowa kluczowe

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

Otrzymany: 2 lutego 2020. Zrecenzowany: 19 lutego 2020. Poprawiony: 22 lutego 2020. Zaakceptowany: 3 czerwca 2020.

ZIN 2020, 58(2), 7–26

DOI: 10.36702/zin.731

1. Wprowadzenie

Niniejsze opracowanie stanowi drugą część analizy metod dopasowywania (identyfikowania) i kwantyfikacji opisu trendów rozwojowych charakteryzujących naukowe dyscypliny, w oparciu o ilościowe cechy zjawiska cytowania przez badaczy publikacji naukowych. Zademonstrowane poprzednio w części pierwszej artykułu metody, a w szczególności metoda regresji w każdym z rozpatrzonych przez autorów wariancie, okazała się posiadać na tyle istotne mankamenty, że w jego zakończeniu wskazano na potrzebę przetestowania kolejnych, odmiennych metod prognozowania trendów, które miałyby szansę na przyniesienie w efekcie swojego zastosowania wyników o wyższej dokładności i miarodajności. Na przykładzie tych samych danych ilościowych, które zostały wykorzystane w poprzednim opracowaniu, choć nieco inaczej zorganizowanych (tj. z wykorzystaniem rozróżnienia danych o cytowaniach na dane należące do sześciu różnych dyscyplin funkcjonujących w obrębie dziedziny nauk o Ziemi – zob. Aneks 1), zaproponowano i oceniono kolejne metody prognostyczne, które w ocenie autorów niniejszej pracy wykazały się wyższym potencjałem w zakresie swojej praktycznej stosowalności, niż metody zweryfikowane uprzednio.

2. Regresja wielomianowa trzeciego stopnia

Jak już wspominano w pierwszej części artykułu, aby prognozowanie z użyciem modeli regresji wielomianowej (dowolnego stopnia) było wiarygodne i metodologicznie uzasadnione, muszą być spełnione te same cztery warunki Gaussa-Markova, o których stwierdzono wcześniej, iż są one koniecznym wymogiem możliwości generalizacji wyników modeli skonstruowanych przy wykorzystaniu regresji liniowej (zob. Dowdy et al., 2004, 213–220; Wojna, 2007, 159–161; Zeliaś, 1997, 53–60, 190–191). Tak więc w przypadku regresji wielomianowej (tj. nie tylko wielomianu stopnia drugiego, ale i wielomianów wyższych stopni) punktem wyjścia dla opartej na niej predykcji rozwoju danej dziedziny nauki musi być sprawdzenie, czy warunki te są spełnione dla danych opisujących historie cytowań przeanalizowanych publikacji (niezależnie od badanego typu wydawniczego). W ramach niniejszego podrozdziału wszystkie przykłady będą również wykorzystywać dane empiryczne, które zebrano dla potrzeb pracy doktorskiej pierwszego z autorów artykułu, i w obrębie których zastosowano podział na dyscypliny funkcjonujące w dziedzinie nauk o Ziemi[1] (zob. Aneks 1).

W tym miejscu należy również podkreślić, że istotną z punktu widzenia celów niniejszego artykułu zaletą trendów wielomianowych stopnia drugiego, trzeciego i ewentualnie stopni

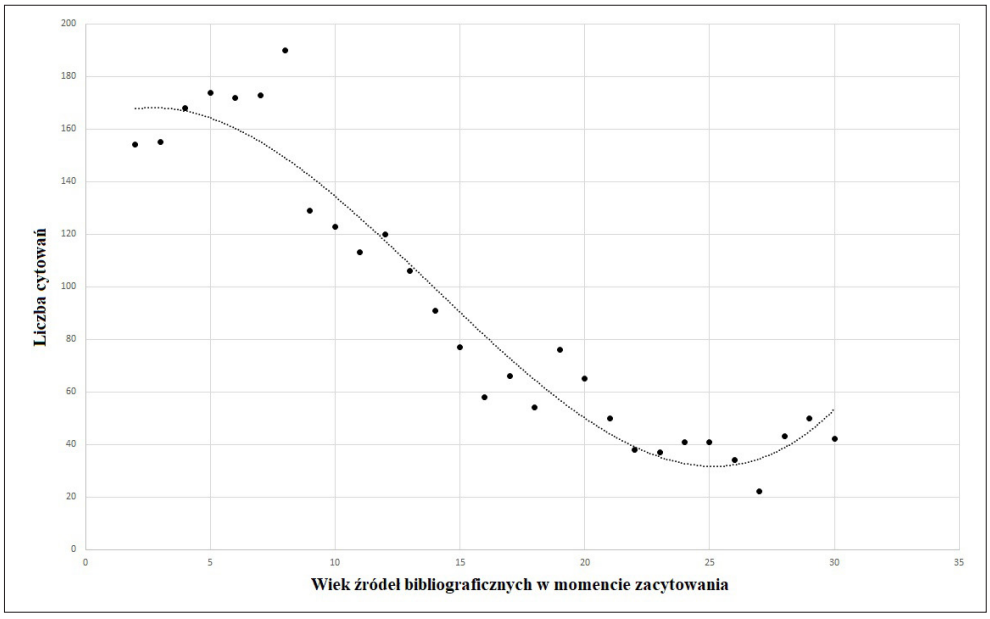
[1] Dziedzina nauk o Ziemi oraz cztery podstawowe wchodzące w jej skład dyscypliny badawcze są tutaj rozumiane w myśl Rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 8 sierpnia 2011 r., w sprawie obszarów nauki, dziedzin nauki i sztuki oraz dyscyplin naukowych i artystycznych (Dz.U. 2011 nr 179 poz. 1065). Decyzja o wyszczególnieniu dwóch dodatkowych dyscyplin („pogranicza nauk o Ziemi i nauk technicznych" oraz „pogranicza nauk o Ziemi i nauk biologicznych") została natomiast podyktowana oceną dokonaną w oparciu o deklaracje redakcji czasopism, treść (i abstrakty) publikowanych na ich łamach artykułów, afiliacje autorów artykułów i kategorie tematyczne przyporządkowane każdemu z czasopism w *Polskiej Bibliografii Naukowej* (https://pbn.nauka.gov.pl/), jak również w odniesieniu do przywołanego wcześniej rozporządzenia MNiSW.

wyższych, jako jednych z nielicznych spośród najczęściej wykorzystywanych w analizach statystycznych postaci funkcji trendu, jest posiadanie przez nie zdolności najlepszego odzwierciedlenia rosnąco-malejącego (w zależności od wybranego przedziału dziedziny funkcji) charakteru krzywej cytowań, w przeciwieństwie do funkcji takich jak funkcja liniowa, wykładnicza, potęgowa, hiperboliczna, logistyczna i logarytmiczna, których specyfika (tj. zależność postaci graficznej od doboru parametrów liczbowych) sprawia, że są one do tego celu mniej przydatne (zob. Krzysztofiak & Luszniewicz, 1976, 364–365; Opaliński & Jaromin, 2017, 111; Sobczyk, 2015, 345; Sobczyk, 2008, 57; Zeliaś et al., 2013, 80–86). Zaleta ta ma przy tym pierwszorzędne znaczenie w kontekście potrzeby uwzględnienia etapu narastania cytowalności, etapu i lokalizacji na osi czasu momentu osiągnięcia cytowalności maksymalnej oraz kształtu, czasowej rozpiętości i tempa zachodzenia etapu spadkowego w historii cytowań podczas konstrukcji równania (modelu regresji) służącego budowie prognozy. Inaczej mówiąc, modele wielomianowe umożliwiają uwzględnienie założenia, że wszystkie te trzy etapy (ich ilościowe i jakościowe charakterystyki) w równym stopniu przyczyniają się do wyłonienia określonej postaci, a zarazem możliwości uchwycenia dynamiki rozwoju dyscypliny. Prognozę determinuje zatem zarówno to, jak literatura naukowa starzeje się, jak i to, kiedy osiąga punkt cytowalności maksymalnej oraz w jaki sposób przebiega proces jej „dojrzewania". Dlatego wszystkie zamieszczone w tabeli 1 równania regresji są równaniami wielomianów. Ważne wydaje się w tym kontekście ponadto tzw. twierdzenie Weierstrassa głoszące, że każdą funkcję (czy każdy rozkład empiryczny) można w pewnym ograniczonym przedziale (fragmencie) dziedziny funkcji, tj. – w tym przypadku – w przedziale czasowym, aproksymować z dowolną dokładnością za pomocą wielomianu stopnia skończonego[2] (Oktaba, 1980, 325; Zeliaś, 1997, 267). Zarazem w literaturze przedmiotu uznaje się, że w zastosowaniach praktycznych lub w opisie pewnych zjawisk fizycznych nie należy stosować wielomianów stopnia wyższego niż stopień trzeci lub czwarty z uwagi na to, że wielomiany wyższych stopni tracą niejako zdolność odzwierciedlania rzeczywistych zjawisk i procesów stając się relacjami zbyt abstrakcyjnymi, jak również relacjami o nikłych zdolnościach predykcyjnych (zob. Bingham & Fry, 2010, 99–100; Ross, 2009, 393; Sen & Srivastava, 1990, 181; Sheskin, 2007, 1268–1269).

Celem uzyskania jak najlepszego stopnia dopasowania trendu do danych empirycznych, co w założeniu ma podnieść zdolności prognostyczne modelu, zastosowano dodatkowo dwa niestandardowe zabiegi metodologiczne, których uzasadnienie tkwi w naturze samych danych o cytowaniach i ich typowych charakterystykach. Po pierwsze zadecydowano o wykluczeniu z części zebranych zbiorów danych tych liczb cytowań, które uzyskały publikacje wydane w 2015 r., będącym bazowym rokiem badania (tj. rokiem, w którym ukazały się wszystkie prace będące źródłem zamieszczonych w nich i zliczonych przypisów bibliograficznych, które odnotowano i zamieszczono w aneksie). Liczby te są bowiem na ogół niedoszacowane (bardzo wyraźnie niższe niż liczby opisujące cytowania prac z 2014 r.) – nie wszystkie prace wydane w tym roku były dostępne dla autorów cytujących w trakcie przygotowywania przez nich ich własnych artykułów. Jeżeli np. artykuł powstawał w pierwszych miesiącach 2015 r., dokumenty wydane pod koniec tego roku nie miały już szans na wejście

[2] Stopień wielomianu jest równy najwyższej potędze zmiennej niezależnej „x", która występuje w jego równaniu. Np. wielomian stopnia 5 miałby ogólną formę: „y = a + bx + cx2 + dx3 + ex4 + fx5", gdzie „a", „b", „c", „d", „e" i „f" to pewne stałe (pewne konkretne liczby), przy czym stała „f" musi być różna od zera.

w skład jego bibliografii załącznikowej. Cytowania odwołujące się do prac z 2015 r. mogą zostać zamiast tego wykorzystane raczej jako swego rodzaju (wprawdzie jedynie bardzo przybliżony) miernik dokładności prognozy. Po drugie zastosowano ograniczenie zakresu danych wejściowych tworzących empiryczną podstawę dla sformułowania równania trendu (równania regresji). Główną przyczyną tego posunięcia był fakt, że przy uwzględnieniu pełnego zakresu danych empirycznych następowała najczęściej utrata zdolności wielomianu (trzeciego i czwartego stopnia) do odzwierciedlenia etapu narastania cytowań trwającego od punktu „zerowego" na osi czasu (tj. od momentu publikacji jakiejś pracy) do momentu osiągnięcia maksymalnej cytowalności. Przykładowo, w przypadku dyscypliny „geografia fizyczna" wykorzystano dane za lata 1986–2014, uzyskując dopasowanie o następującym kształcie (Rys. 1).

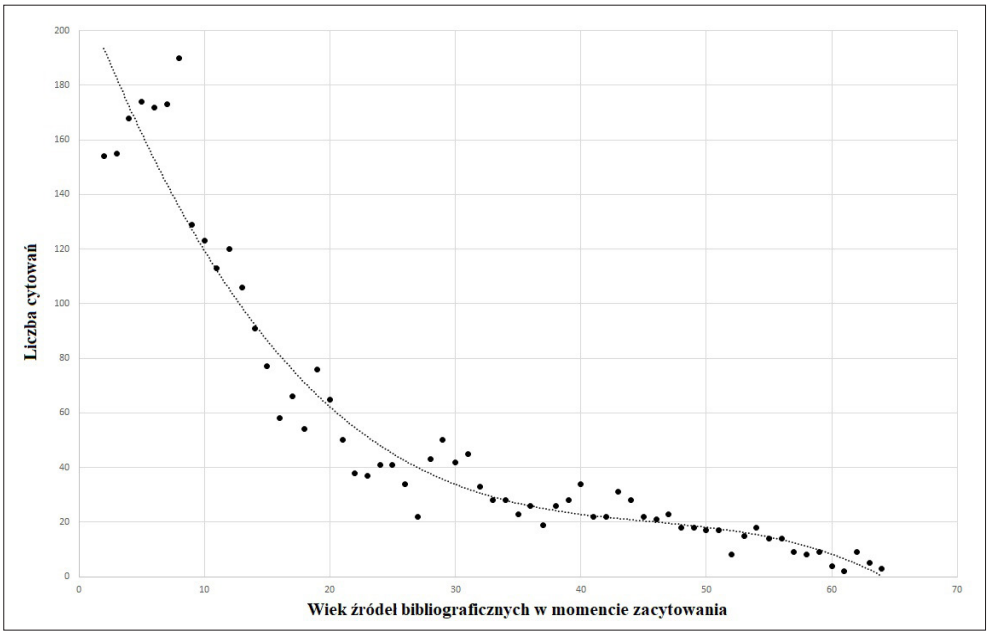


Rys. 1. Wielomian trzeciego stopnia dopasowany do danych o cytowaniach publikacji z lat 1986–2014 w obrębie dyscypliny geografia fizyczna

Przy wykorzystaniu znacznie szerszego zakresu danych, obejmującego mianowicie lata 1950–2014, dopasowanie wielomianu trzeciego stopnia miało natomiast postać zilustrowaną na rysunku 2.

Jak widać, w tym drugim przypadku (Rys. 2), w którym nie zastosowano ograniczenia czasowego dla zakresu danych o cytowaniach, linia regresji nie posiada żadnego widocznego momentu maksimum, po którym zaczynałaby opadać, odzwierciedlając w ten sposób osiągnięcie momentu najwyższej cytowalności przez zbadane źródła piśmiennicze i następujący po nim spadek cytowalności. Zamiast tego cytowalność zobrazowana trendem z rysunku 2 wydaje się nieprzerwanie wzrastać, wraz z przesuwaniem się na osi czasu ku najnowszym – wydanym najbliżej momentu „0" – źródłom piśmienniczym, co należy uznać

za wadliwe z punktu widzenia naukoznawczego sensu rzeczywistych danych empirycznych, jak również z punktu widzenia ogólniejszej wiedzy na temat zachowań w zakresie cytowań naukowców oraz konwencjonalnych (tj. dominujących w paradygmacie „nauki normalnej" w sensie teorii Thomasa Kuhna – zob. Kuhn, 2001, 53–85) wzorców rozwoju i postępu dokonującego się w większości współczesnych dziedzin i dyscyplin naukowych.



Rys. 2. Wielomian trzeciego stopnia dopasowany do danych o cytowaniach publikacji z lat 1950–2014 w obrębie dyscypliny geografia fizyczna

W przypadku każdej z dyscyplin będących przedmiotem badania starano się wobec tego dobrać okres służący za podstawę sformułowania równania trendu tak, aby był on mimo wszystko możliwie długi. Dążeniem autorów było też, by spełniał on założenie dotyczące krzywoliniowego przebiegu początkowego etapu historii cytowalności przywoływanych w publikacjach źródeł literaturowych.

Wyniki sprawdzianu zasadności wszystkich czterech omówionych w części pierwszej artykułu założeń Gaussa-Markova w odniesieniu do sześciu wyszczególnionych w ramach zebranych danych empirycznych dyscyplin, należących do dziedziny nauk o Ziemi, przedstawia tabela 1.

W odniesieniu do danych zamieszczonych w tabeli 1 warto nadmienić, że wartość statystyki chi-kwadrat dla dyscypliny z pogranicza nauk o Ziemi i nauk biologicznych przekroczyła wartość krytyczną wyłącznie z uwagi na wystąpienie jednej wartości odstającej (ang. *outlier*) wśród całego zbioru składników resztowych. Po jej ewentualnym wyeliminowaniu statystyka przybrałaby znacznie mniejszą wartość, która pozwoliłaby na wyciągnięcie wniosku o zgodności rozkładu reszt z rozkładem normalnym.

Tab. 1. Wyniki analizy założeń dotyczących zjawisk autokorelacji, homoskedastyczności oraz typu rozkładu składników resztowych modeli regresji dla wyróżnionych w badaniu dyscyplin

Dyscyplina	Równanie regresji	Wartość „R2”	Zakres czasowy uwzględnionych danych empirycznych	Średnia składników resztowych	Wartość statystyki Goldfelda-Quandta	Wartość statystyki „d”	Wartość statystyki chi-kwadrat
Geografia fizyczna	$y = 0.0241x^3 - 0.937x^2 + 3.012x + 165.62$	0.935	1986–2014	$3.96924 \times \frac{1}{10^{14}}$	0.14 (< 3.44)*	1.24 (< 1.34)**	38.889 (< 38.89)*****
Geografia społeczno-ekonomiczna	$y = 0.0818x^3 - 2.2493x^2 + 0.5211x + 300.29$	0.946	1995–2014	$6.25278 \times \frac{1}{10^{13}}$	0.017 (< 0.198)*	1.47 (> 1.41)****	10.99 (< 27.59)*****
Geologia	$y = 0.0055x^3 - 0.3232x^2 + 0.6154x + 188.44$	0.896	1971–2015	$-1.01055 \times \frac{1}{10^{14}}$	0.17 (< 0.4)*	2.1 (> 1.57)****	40.91 (< 55.76)*****
Geofizyka i geochemia	$y = 0.0129x^3 - 0.558x^2 + 0.073x + 192.97$	0.91	1983–2014	$-1.28786 \times \frac{1}{10^{14}}$	0.218 (< 0.336)*	1.99 (> 1.5)****	8.52 (< 42.56)*****
Pogranicze nauk o Ziemi i nauk technicznych	$y = 0.0523x^3 - 1.5064x^2 - 3.7643x + 295.58$	0.85	1993–2015	$1.91538 \times \frac{1}{10^{14}}$	0.012 (< 0.26)*	1.37 (> 1.26 i < 1.44)******	34.12 (> 31.41)***
Pogranicze nauk o Ziemi i nauk biologicznych	$y = 0.0186x^3 - 0.9827x^2 + 7.9607x + 161.19$	0.973	1981–2014	$-6.68746 \times \frac{1}{10^{15}}$	0.164 (< 0.336)*	1.23 (< 1.39)**	82.7 (> 43.77)***

* Rozkład reszt dla tego modelu jest homoskedastyczny (wariancja składników resztowych nie zmienia się widocznie wraz ze wzrostem wartości zmiennej niezależnej).
** Wynik ten oznacza, że występuje tu zjawisko autokorelacji składników resztowych.
*** Rozkład reszt jest niezgodny z rozkładem normalnym.
**** Zjawisko autokorelacji składników resztowych nie występuje.
***** Rozkład reszt jest zgodny z rozkładem normalnym.
****** Wartość ta znajduje się w tzw. obszarze braku decyzji, tzn. ani nie można na jej podstawie uznać, że występuje tu cecha autokorelacji, ani też, że cecha ta nie występuje.

W przypadkach, w których jedno lub więcej spośród czterech wskazanych powyżej założeń modelu regresji nie zostało spełnione, istnieją wielorakie inne możliwości skonstruowania prognozy punktowej, spośród których trudno jest jednoznacznie wskazać tę pod pewnym względem najlepszą czy najwłaściwszą. Możliwości te obejmują m.in. prognozowanie z wykorzystaniem adaptacyjnych modeli wygładzania wykładniczego, całej klasy tzw. modeli autoregresyjnych, modeli wykładniczo-autoregresyjnych, modeli uwzględniających poprawkę ze względu na autokorelację składników resztowych lub ze względu na heteroskedastyczność wariancji składnika losowego, modeli średnich ruchomych, modeli mieszanych (tj. łączących składnik autoregresyjny ze średnią ruchomą) i innych (zob. Chatfield, 1975, 82–108; Guzik et al., 2004, 101–136, 231–252; Hyndman et al., 2008; Krawiec, 2014; Montgomery et al., 2008, 171–287; Pawłowski, 1981, 208–245; Shumway & Stoffer, 2011, 70–162; Snarska, 2011, 212–262; Sobczyk, 2008, 97–152; Sobczyk, 2015, 309–356; Witkowska, 2005, 122–176; Zaiontz, 2017; Zeliaś, 1997, 189–288; Zeliaś et al., 2013, 140–172, 233–261). W przypadkach, w których założenia zostały spełnione dopuszczalne jest natomiast budowanie prognozy przy użyciu metody analitycznej i równania trendu, którą zademonstrowano w innym opracowaniu autorów niniejszego artykułu (Opaliński & Jaromin, 2017). Jest to podejście uzasadnione pod warunkiem wcześniejszego przetestowania i potwierdzenia możliwości uogólnienia równania trendu, tj. zbadania statystycznej istotności współczynników regresji wielomianowej za pomocą odpowiedniego testu statystycznego (zob. Aczel & Sounderpandian, 2018, 691–698). Z danych przedstawionych w tabeli 1 wynika, że do przypadków pierwszego rodzaju (pewne założenia nie są spełnione) należą dyscypliny takie jak: geografia fizyczna, pogranicze nauk o Ziemi i nauk technicznych oraz pogranicze nauk o Ziemi i nauk biologicznych. Do grupy przypadków drugiego rodzaju (wszystkie założenia są spełnione) należą pozostałe dyscypliny: geografia społeczno-ekonomiczna, geologia oraz geofizyka i geochemia.

3. Ocena statystycznej istotności modeli spełniających wszystkie założenia Gaussa-Markova

W nawiązaniu do wymienionych wyżej trzech przypadków dyscyplin należących do grupy pierwszego rodzaju, tj. grupy, w której wszystkie założenia analizy regresji zostały spełnione, wykonano statystyczny test „t", sprawdzający istotność współczynników regresji stojących przy kolejnych zmiennych niezależnych (tj. „x^3", „x^2" itd.). W teście tym weryfikuje się kolejne hipotezy zerowe mówiące, że w rzeczywistości poszczególne z oszacowanych dla danego modelu parametrów są równe zeru, poprzez porównanie obliczonej wartości statystyki testowej (którą w przypadku testu „t" jest prosty iloraz wartości konkretnego współczynnika regresji i jego błędu standardowego), z wartością krytyczną rozkładu statystyki „t" (zob. Aczel & Sounderpandian, 2018, 691–695; Allen, 1997, 66–70; Sobczyk, 2008, 47). W dalszej części niniejszego opracowania wykorzystano zapis w klamrze („{...}"), w którą ujęto element (tj. pewien konkretny parametr bądź parametry regresji) nieistotny statystycznie na pięcioprocentowym poziomie ufności, wobec czego należało wykluczyć go z równania w trakcie budowania prognozy. Inaczej mówiąc elementy tego rodzaju występują w równaniu opisującym badaną próbę losową, ale nie należy traktować ich jako składników równania, kiedy rozpatruje się je jako model opisujący pewne cechy szerszej populacji generalnej.

W odniesieniu do omawianej grupy wydaje się, że po zrealizowanym przez autorów przetestowaniu modeli z tabeli 1 i ich ocenie, można w jej ramach podać prognozy wynikające wprost z wielomianowych równań regresji oraz ocenić zdolność prognostyczną modelu poprzez wskazanie wartości współczynnika „V_e" oraz wartości tzw. średniego kwadratowego błędu prognozy *ex post* występującego w przedziale weryfikacji (liczącym – w omawianych przypadkach sześć kolejnych wartości zmiennej zależnej „y"). Dla dyscypliny geografii społeczno-ekonomicznej uzyskano następujący rezultat[3] (zapis „y_{2015}" oznacza wartość prognozowanej liczby cytowań w dyscyplinie dla roku 2015 itd.):

$$y_{2015} = 0.0818\ x^3 - \{2.2493x^2 + 0.5211x\} + 300.29 = 0.0818\ (0)^3 + 300.29 = 300.29 \approx 300$$

Dla kolejnych lat analogicznie uzyskano kolejne wyniki: $y_{2016} \approx 300$, $y_{2017} \approx 300$, $y_{2018} \approx 298$, $y_{2019} \approx 295$, $y_{2020} \approx 290$.

Współczynnik "V_e" dla tego zbioru danych jest równy „0.14" (14%), podczas gdy średni kwadratowy błąd prognozy *ex post* w przedziale 2014–2009 wyniósł 32.9 cytowania, co jest wartością większą niż wartość odchylenia standardowego składników resztowych modelu („21.28"). Ze względu na ostatni z wymienionych faktów prognozę należy określić jako nie w pełni zadowalającą. Warto jednak dodatkowo zaznaczyć, że zwiększając szerokość przedziału uzyskujemy poprawę dokładności prognozy. Przykładowo dla przedziału 2006–2014 błąd prognozy *ex post* jest równy 27.4 cytowaniom, a dla całego rozpatrywanego okresu 1995–2014 błąd ten jest równy 20.74 cytowaniom i jest on wtedy mniejszy od błędu standardowego składników resztowych.

Dla dyscypliny geologii otrzymano dalsze rezultaty postaci:

$$y_{2016} = 0.0055\ x^3 - 0.3232\ x^2 + \{0.6154\ x\} + 188.44 =$$
$$0.0055 \times (0)^3 - 0.3232 \times (0)^2 + 188.44 = 188.44 \approx 188$$

W dalszej kolejności: $y_{2017} \approx 188$, $y_{2018} \approx 187$, $y_{2019} \approx 185$, $y_{2020} \approx 182$

Współczynnik "V_e" dla tego zbioru danych jest równy „0.17" (17%), podczas gdy średni kwadratowy błąd prognozy *ex post* w przedziale 2009–2015 wyniósł „23.4" cytowania, co ponownie jest wartością większą niż wartość odchylenia standardowego składników resztowych modelu wynosząca „17.58". Dla całego zakresu 1971–2015 błąd prognozy jest natomiast równy 17.38 cytowaniom[4].

[3] Pewnym mankamentem przedstawionego rozumowania jest konieczność „wstecznej" numeracji jednostek czasu, tj. przy przyjętej postaci modelu rok 2015 opisuje wartość „t" równa „0", rok 2016 – wartość „t" równa „–1", rok 2017 – wartość „t" równa „–2" itd. Problem ten nie wpływa wprawdzie w negatywny sposób na same wartości prognozowanych liczb cytowań dla kolejnych lat, utrudnia jednak (a nawet uniemożliwia) wyznaczenie sensownych wartości błędów *ex ante* (np. dla roku 2015 błąd ten wyniósł ponad 700%). Z kolei próby zmiany numeracji jednostek czasu na zgodne z propozycją Mieczysława Sobczyka (Sobczyk, 2008, 59–63), gdzie numery wszystkich jednostek czasu sumują się do zera, wywoływały problemy innego rodzaju, które były związane przede wszystkim z możliwościami spełnienia założeń modelu regresji. Z tego też powodu w ramach tej części niniejszego opracowania wykorzystano alternatywną miarę błędu *ex post* przewidywań modelu, która również może służyć za podstawę oceny dopuszczalności predykcji.

[4] Warto tu także nadmienić, że sytuacja ta wskazuje na to, że tak w tym, jak i w poprzednim przypadku odchylenia od prognoz są większe dla wartości bliskich współczesności (tj. dla lat 2015, 2014, 2013 itd.) niż dla znacznie wcześniejszych lat wydania źródeł bibliograficznych (np. 1980 itd.). Z pozoru może to świadczyć

Dla dyscypliny geofizyki i geochemii mamy:

$$y_{2015} = 0.0129\ x^3 - 0.558\ x^2 + \{0.073\ x\} + 192.97 =$$
$$0.0129 \times (0)^3 - 0.558 \times (0)^2 + 192.97 = 192.97 \approx 193$$

Następnie: $y_{2016} \approx 192$, $y_{2017} \approx 191$, $y_{2018} \approx 188$, $y_{2019} \approx 183$ i $y_{2020} \approx 177$

Współczynnik "V_e" dla tego zbioru danych jest równy „0.17" (17%), podczas gdy średni kwadratowy błąd prognozy *ex post* w przedziale 2009–2014 wyniósł „23.4" cytowania, co jest wartością większą niż wartość odchylenia standardowego składników resztowych modelu wynoszącego „17.6". Dla całego zakresu 1983–2014 błąd prognozy jest natomiast równy 17.33 cytowaniom.

Z uwagi na fakt, że w każdym z trzech powyższych przypadków błąd prognozy *ex post* maleje wraz ze wzrostem przedziału czasowego uwzględnionych przez model danych wejściowych należy stwierdzić, że zabieg korygowania (skracania) tego przedziału, streszczony na wstępie podrozdziału 2, okazał się w tej perspektywie niekorzystny. Rozwiązanie lub minimalizacja zasięgu tego problemu metodologicznego wydaje się obecnie pozostawać jedną z dalszych perspektyw badawczych.

4. Wybrane alternatywne metody prognozowania przyszłej cytowalności publikacji w oparciu o dotychczasową historię cytowań

W związku z padającymi powyżej stwierdzeniami i dotychczas osiągniętymi wynikami wydaje się, że do danych o dyscyplinach pierwszego rodzaju, tj. takich, dla których pewne założenia Gaussa-Markova nie zostały spełnione, warto w pierwszej kolejności wykorzystać możliwość zastosowania pewnych statystycznie sformalizowanych korekt do metod już wcześniej przez autorów wykorzystanych, a niewskazanych w tego rodzaju przypadkach właśnie ze względu na pogwałcenie któregoś z warunków Gaussa-Markova. W szczególności mowa tu o geografii fizycznej ponieważ jest to jedyna dyscyplina, dla której nie zostało spełnione wyłącznie założenie o braku autokorelacji składnika losowego. W tej sytuacji celowe wydaje się więc zastosowanie poprawki ze względu na tę autokorelację, która teoretycznie powinna wyeliminować występujący tu problem. Wprowadzenie takiej poprawki polega na ustaleniu istnienia i postaci zależności liniowej zachodzącej pomiędzy kolejnymi składnikami losowymi, tj. pomiędzy „e_t" i „e_{t-1}", (np. „e_{2014}" i „e_{2013}"), „e_{t-1}" i „e_{t-2}", (np. „e_{2013}" i „e_{2012}") itd. i dodaniu jej do wyjściowego równania regresji, co wpłynie na rozkład składników losowych (reszt modelu). W tym kontekście ustalono, że składniki losowe łączy (słaba) korelacja liniowa o postaci „$e_t = 0.36\ e_{t-1} + 0.35$" wobec czego model wyjściowy (zob. Tab. 1) przekształcono do formy:

$$y = 0.0241 \times x^3 - 0.937 \times x^2 + \{3.012 \times x\} + 165.62 + [0.36 \times e_{t-1} + 0.35]$$

o niejednorodności wariancji składnika losowego czyli o heteroskedastyczności modelu, co zostało jednak wykluczone przez odpowiednią wartość statystyki Goldfelda-Quandta (zob. Tab. 1). Zgodnie z tym wynikiem pojawiające się tu różnice należy przypisać przypadkowi wynikłemu z takiego akurat, a nie innego doboru próby (tj. charakterystykom pobranej próbki, a nie charakterystykom samej populacji generalnej).

Po ponownym wyznaczeniu statystyki „d" dla nowo powstałych składników resztowych okazało się, że jej wartość wyniosła „1.24", co przesunęło ją z obszaru nakazującego odrzucenie hipotezy zerowej (o braku zjawiska autokorelacji) do tzw. obszaru braku decyzji. Nie można zatem było jednoznacznie stwierdzić, że wprowadzenie poprawki wyeliminowało całkowicie zjawisko autokorelacji, tak samo jak nie można było stwierdzić, że wprowadzenie poprawki nie odniosło żadnego skutku. W związku z tym autorzy uznali, że warto mimo wszystko wyprowadzić w tym momencie prognozę punktową, podać jej błąd *ex post* i w dalszej kolejności porównać go z błędem alternatywnej prognozy dla tej samej dyscypliny, uzyskanej tym razem metodą autoregresji. Prognoza powstaje poprzez podstawienie do równania trendu (równania regresji) wartości „0" dla (bazowego) 2015 r., wartości „–1" dla 2016 r. itd., aż do roku 2020 (wartość „–5"). I tak, dla roku 2015 otrzymuje się:

$$y_{2015} = 0.0241 \times x^3 - 0.937 \times x^2 + 165.62 + [0.36 \times (-13.09) + 0.35] =$$
$$0.0241 \times (0)^3 - 0.937 \times (0)^2 + 165.62 - 4.33 = 161.29 \approx 161$$

Analogicznie obliczone wyniki dla kolejnych lat wyniosły: $y_{2016} \approx 165$, $y_{2017} \approx 162$, $y_{2018} \approx 157$, $y_{2019} \approx 149$ i $y_{2020} \approx 139$. Współczynnik zgodności dopasowania „R2" okazał się bardzo wysoki („0.99"), a współczynnik zmienności resztowej „V_e" był równy „0.053". Wartość tego współczynnika oznacza, że tylko około 5.3% zmienności badanego zjawiska stanowią przypadkowe (niewyjaśnione przez przyjęty model) odchylenia danych od teoretycznej funkcji trendu. Średni kwadratowy błąd prognoz *ex post* w wybranym przez autorów przedziale weryfikacji (tj. w pięcioletnim przedziale 2010–2014) wyniósł „3.91" cytowania i był mniejszy od wartości odchylenia standardowego reszt modelu (tj. od „4.88") co pozwala uznać prognozy za zadowalające (dopuszczalne).

Dalsze alternatywne metody prognozowania zjawisk ilościowych, które pojawiły się w obszarze zainteresowania autorów niniejszej pracy i które ich zdaniem warto przetestować na dostępnym dla nich zbiorze danych empirycznych, to przede wszystkim bądź jeden z modeli adaptacyjnych, bądź model autoregresyjny[5]. Podstawową charakterystyką tego ostatniego jest wyprowadzenie prognozy punktowej ze związku pomiędzy wartościami zmiennej zależnej, które przybrała ona w pewnym czasie „t" (tj. „y_t"), a wartościami tej samej zmiennej zależnej, które przybrała ona we wcześniejszych momentach czasu „t-1", „t-2", ..., „t-p" (tj. „y_{t-1}", „y_{t-2}", ..., „y_{t-p}"). Liczba „p" oznaczająca to, jak daleko należy „cofnąć się w czasie" z zamiarem odczytania wartości zmiennej „yt-p", nazywanej też zmienną opóźnioną, określa tzw. rząd autoregresji. Model tego rodzaju symbolicznie zapisuje się jako „AR(p)", czyli model autoregresyjny rzędu „p" (zob. Zeliaś, 1997, 247–288; Zeliaś et al., 2013, 243–246, 258–262). Można powiedzieć, że jest to zasadniczo model regresji liniowej (lub nieliniowej), z jedną lub

[5] Przedstawione w dalszej części niniejszego artykułu koncepcje, obliczenia i notacja są wzorowane głównie na cytowanych wyżej podręcznikach, a w szczególności na pracach Aleksandra Zeliasia (Zeliaś, 1997), Aleksandra Zeliasia i in. (Zeliaś et al., 2013), Stanisława Krawca (Krawiec, 2014) oraz Agnieszki Snarskiej (Snarska, 2011). Obliczenia wykonano z użyciem programu Microsoft Excel w wersjach 2010 i 2013, jak również z użyciem kalkulatora macierzy online dostępnego pod adresem: https://matrixcalc.org/pl/. Zakres uwzględnionych w obliczeniach danych empirycznych pokrywa się z zakresem wskazanym w tabeli 1 (np. dla geografii fizycznej są to lata 1986–2014). Można dodatkowo podkreślić, że technika obliczeń oraz metody wyznaczania dokładności (dopuszczalności) prognoz punktowych nie różnią się od techniki wykorzystanej we wcześniejszym opracowaniu autorów (Opaliński & Jaromin, 2017).

więcej niż jedną zmienną niezależną (tj. zmienną „x"), w którym rolę tej zmiennej niezależnej pełnią opóźnione wartości zmiennej zależnej. Z naukoznawczego punktu widzenia jako uzasadnienie przyjęcia tego typu modelu można przywołać tzw. zasadę skumulowanej korzyści Price'a, według której przyszłą cytowalność publikacji naukowych niemal w pełni determinuje dotychczasowa historia ich cytowalności (ang. *cumulative advantage distribution* – zob. Price, 1976, 292–293). Zasadę tę można określić też w pewnym uproszczeniu mianem idei, według której „sukces rodzi sukces" (ang. *success breeds success* – zob. np. Huber, 1998). Ponieważ wartość zmiennej zależnej „y_t" należy interpretować jako poziom cytowalności (liczbę cytowań) pewnych publikacji w czasie „t" (np. w 2015 r.), a wartości zmiennych opóźnionych („y_{t-1}" itd.) jako cytowalność tych samych publikacji w latach wcześniejszych (czyli w roku 2014, 2013 itd.) ustalenie tego, ile takich zmiennych opóźnionych należy uwzględnić w modelu oraz w jakim stopniu oddziałują one na wartości zmiennej nieopóźnionej jest w zasadzie dokładną realizacją (konkretyzacją) oryginalnej idei Price'a z 1976 r.

W przypadku analizy autoregresji podstawowym problemem jest tzw. stacjonarność modelowanego procesu. Procesem stacjonarnym jest mianowicie taki proces, który kształtuje się wokół wartości średniej arytmetycznej, wykazując jedynie losowe odchylenia od tej wartości. Jakikolwiek proces charakteryzujący się trendem jest więc procesem niestacjonarnym, a do takich właśnie procesów należy ujmowana w ramach metodologii synchronicznej cytowalność wszystkich rozpatrywanych w niniejszym artykule dyscyplin naukowych. Wydaje się w związku z tym, że do celu jednorazowego sprawdzenia wiarygodności i porównania wartości uzyskanych dwoma alternatywnymi metodami prognoz dozwolone będzie chwilowe „zawieszenie" tego szczególnego wymogu. Mając na uwadze to zastrzeżenie i decydując się na podjęcie analizy autoregresji, w pierwszej kolejności należy ustalić odpowiedni dla zebranego zbioru danych jej rząd. Rząd ten (tj. liczbę uwzględnionych w równaniu zmiennych opóźnionych) dobiera się w oparciu o wartości funkcji „SR(k)", preferując tę z nich, która okazuje się wartością najmniejszą. Należy dodatkowo mieć na uwadze zasadę, według której w praktyce nie należy stosować modeli o zbyt wysokim rzędzie autoregresji. Wartości funkcji SR(k) wyznacza się ze wzoru:

$$SR(k) = \ln s_e^2(k) + \frac{k}{n} \times \ln n;\ k = 0, 1, ..., K$$

gdzie:

ln – symbol logarytmu naturalnego,

k – wybrany rząd autoregresji,

n – liczba elementów badanego szeregu czasowego,

$s_e^2(k)$ – wariancja składnika losowego modelu autoregresji rzędu „k",

K – maksymalny rząd autoregresji (zob. Zeliaś et al., 2013, 245).

Wartości wyrażeń: „$s_e^2(k)$" i „SR(k)" dla „k = 1, 2, 3, 4, 5, 6, 7", w przypadku danych obejmujących dyscyplinę geografii fizycznej, prezentuje tabela 2.

Tab. 2. Wartości wyrażeń „$s_e^2(k)$" i „SR(k)" przy „k = 1, 2, 3, 4, 5, 6, 7" dla geografii fizycznej

Geografia fizyczna							
k	**1**	**2**	**3**	**4**	**5**	**6**	**7**
s_e^2	255.974	270.659	272.108	271.178	285.12	279.05	115.55
SR(k)	5.66	5.83	5.95	6.07	6.23	6.33	5.56

Stąd, że najmniejsza wartość „SR(k)" wystąpiła dla „k" równego „7" wynika, iż dla powyższych danych należy przyjąć rząd autoregresji równy „7". Najlepszym modelem autoregresji (zgodnie z oszacowaniami wykonanymi przy pomocy programu Microsoft Excel 2013) jest zatem:

$$y_t = 7.51 + 0.324\, y_{t-1} + 0.109\, y_{t-2} - 0.176\, y_{t-3} + 0.218\, y_{t-4} + 0.249\, y_{t-5} - 0.00096\, y_{t-6} - 0.023\, y_{t-7} + e_t$$

Dla tego modelu[6] współczynnik determinacji „R^2" równał się „0.928", a współczynnik zmienności resztowej „V_e" był równy „0.16". W jego ramach można następnie wyprowadzić prognozowane liczby cytowań populacji generalnej publikacji należących do geografii fizycznej. Prognozy punktowe na lata 2015–2020 kształtują się na następującym poziomie (w nawiasach na końcu obliczeń dla każdego rozpatrzonego roku podano wartości względnego błędu prognozy *ex ante*):

$$y_{2015} = 7.51 + 0.324\, y_{2014} + 0.109\, y_{2013} - 0.176\, y_{2012} + 0.218\, y_{2011} + 0.249\, y_{2010} - 0.00096\, y_{2009} - 0.023\, y_{2008} + e_t = 7.51 + 0.324 \times 154 + 0.109 \times 155 - 0.176 \times 168 + 0.218 \times 174 + 0.249 \times 172 - 0.00096 \times 173 - 0.023 \times 190 + e_t = 120.957 \approx 121\ (10.18\%)$$

W analogiczny sposób wyznaczono wartości: $y_{2016} \approx 112$ (10.63%), $y_{2017} \approx 101$ (11.8%), $y_{2018} \approx 99$ (11.99%), $y_{2019} \approx 92$ (12.86%) i $y_{2020} \approx 81$ (14.53%).

Porównanie wartości wskaźników „R^2", „V_e" oraz oszacowanych wartości błędów wydaje się wskazywać, że prognozy na lata 2015–2020 zbudowane w oparciu o model z poprawką ze względu na autokorelację składników resztowych cechują się wyższym stopniem dokładności i większą zgodnością z danymi empirycznymi od prognoz uzyskanych przy pomocy (wykorzystanej tu uproszczonej wersji) modelu autoregresyjnego.

Ponieważ wyniki zastosowania metody uproszczonej (tj. metody, w której zawieszono założenie o stacjonarności modelowanego procesu) okazały się nie w pełni zadowalające, można dodatkowo wykorzystać fakt, że w statystyce istnieją techniki eliminowania niestacjonarności modelowanych procesów. Najprostszą z nich jest tzw. różnicowanie szeregu czasowego, które polega na zamianie jego kolejnych wyrazów na różnice wyrazów sąsiednich i przeprowadzeniu procedury autoregresji na tak przekształconych danych. Następnie należy odwrócić proces różnicowania, aby otrzymać prognozy dotyczące szeregu

[6] Należy w związku z nim dodatkowo podkreślić, że – jak już wspomniano – model autoregresyjny traktuje się jako model regresji wieloliniowej (tj. z wieloma zmiennymi objaśniającymi i jedną zmienną objaśnianą). W związku z tym oceny statystycznej istotności wszystkich parametrów modelu z osobna można dokonywać nie tylko za pomocą testu „t" (jak miało to miejsce np. podczas wcześniejszej analizy standardowych modeli regresji wielomianowej oraz modelu z poprawką ze względu na autokorelację), ale też za pomocą testu „F". Fakt ten ma istotne znaczenie w świetle tego, że wszystkie elementy powyższego równania autoregresyjnego z osobna są statystycznie nieistotne, ale wzięte łącznie i zbadane testem „F" są statystycznie istotne. Innymi słowy test „F" pozwala na pozytywne lub negatywne zweryfikowanie hipotezy zerowej, według której ze statystycznego punktu widzenia wszystkie współczynniki (auto)regresji są sobie równe i mają (jednakową) wartość „0" (Allen, 1997, 109–112). W omawianym przypadku hipotezę tę należy odrzucić. Szczegóły testu, a przede wszystkim wartości statystyk testowych otrzymano w pakiecie „Analiza danych" programu Microsoft Excel 2013.

wyjściowego (nieprzekształconego). W przypadku szeregów cechujących się trendem kwadratowym (tj. parabolicznym) należy zastosować różnicowanie drugiego rzędu, tj. obliczenie różnic szeregu złożonego z różnic wartości oryginalnych (tj. obliczenie różnic w szeregu już wcześniej zróżnicowanym). Odwracanie różnicowania poziomu drugiego odbywa się na podstawie wzoru:

$$X_i = \Delta X_i + 2X_{i-1} - X_{i-2}$$

gdzie:

X_i – wartość szeregu nr „i" (np. pierwsza wartość szeregu, druga wartość szeregu itd.),

ΔX_i – różnica wartości nr „i" oraz wartości „i-1" szeregu (np. wartości nr 5 i nr 4),

X_{i-1} – wartość szeregu poprzedzająca bezpośrednio wartość nr „i",

X_{i-2} – wartość stojąca w szeregu o dwa „miejsca" wcześniej niż wartość nr „i".

Do celu prognozowania tą metodą można wykorzystać np. prosty program komputerowy stworzony przez Piotra Chudzika[7]. Wyniki uzyskane w tym programie wskazały, że w ramach modelu autoregresji 7 rzędu z różnicowaniem rzędu 2, wartości prognozowane na kolejne lata począwszy od roku 2015, w dalszym ciągu uzyskane w odniesieniu do dyscypliny geografii fizycznej, to: $y_{2015} \approx 178$, $y_{2016} \approx 153$ $y_{2017} \approx 153$, $y_{2018} \approx 146$, $y_{2019} \approx 147$ i $y_{2020} \approx 137$.

Program wyznacza również wartości pewnych standardowych błędów prognostycznych, m.in. tzw. średni błąd bezwzględny (ang. *mean absolute error*) informujący o tym, o ile średnio w okresie prognoz będzie wynosić odchylenie predykcji od wartości rzeczywistej. W przypadku powyższych danych jest on równy „ 13" cytowań. Można więc powiedzieć, że prognozy uzyskane tą metodą okazały się sytuować na poziomie dopuszczalności zbliżonym do poziomu dopuszczalności prognoz uzyskanych metodą uproszczoną (np. dla 2015 r. błąd był równy 10.2% ze 121 cytowań, co daje około 12.3 cytowań). Wydaje się więc, że w dalszym ciągu należy preferować prognozy otrzymane przy pomocy metody z poprawką ze względu na zjawisko autokorelacji. Testowanie innych poziomów różnicowania oraz innych rzędów autoregresji pozostaje natomiast poza nawiasem niniejszego opracowania i stanowi zarazem jeszcze jedną możliwą perspektywę przyszłych badań.

W przypadku obszaru nazwanego „pograniczem nauk o Ziemi i nauk technicznych" oraz „pograniczem nauk o Ziemi i nauk biologicznych" niespełnione są dwa założenia (tj. założenie o braku autokorelacji składników resztowych oraz o normalności ich rozkładu). Z uwagi na to najbezpieczniejszym (najbardziej uzasadnionym i racjonalnym) wyjściem wydaje się wykorzystanie jednego z modeli adaptacyjnych. Jedynym założeniem, na którym spoczywają metody wykorzystywane do tworzenia tego typu modeli jest bowiem teza, że w najbliższej przyszłości nie nastąpią zmiany w dotychczasowym sposobie oddziaływania całokształtu czynników determinujących wartości przybierane przez zmienne prognozowane (Krawiec, 2014, 11). Ponieważ wejściowe zakresy danych empirycznych nie obejmują całości przedziału 1950–2015, a jedynie pewne jego wycinki (zob. Tab. 1) uzasadnione wydaje się wykorzystanie tzw. kwadratowego (trójparametrowego) adaptacyjnego modelu Holta, który nadaje się do prognozowania szeregów wykazujących obecność tendencji rozwojowej o postaci parabolicznej (Krawiec, 2014, 87). Kiedy bowiem rozpatruje się wspomniane przedziały

[7] Opis działania oraz link do pobrania dostępne są pod adresem: http://visualmonsters.cba.pl/index.php/prognozowanie/autoregresja-wyzszego-poziomu/

cytowalności zamiast całości cyklu, uwidaczniająca się w nich tendencja ma formę najbliższą właśnie wielomianowi stopnia drugiego (którego wykres ma postać paraboli). Dokładną formułę pozwalającą na obliczenie prognoz (oraz błędów *ex post*) podano w podręczniku Stanisława Krawca (Krawiec, 2014, 88–89, 103–104), a wyniki jej zastosowania do danych nieznacznie skróconych w stosunku do przedziałów podanych w tabeli 1 (co ma na celu silniejsze uwydatnienie podobieństwa kształtu badanego fragmentu historii cytowań do paraboli, zamiast wykresu wielomianu stopnia trzeciego) prezentuje tabela 3.

Figurujące w tabeli parametry „α", „β" i „Φ" są konieczne do obliczenia prognoz, a ich wartości wyznacza się eksperymentalnie dla poszczególnych zbiorów danych z osobna wykorzystując w tym celu wartość średniego względnego błędu prognoz wygasłych, który różni się w zależności od konkretnych zestawów parametrów. Prognoza jest naturalnie tym dokładniejsza, im błąd ten jest mniejszy.

Tab. 3. Zaokrąglone wyniki zastosowania kwadratowego modelu Holta do prognozowania cytowalności dyscypliny usytuowanej na pograniczu nauk o Ziemi i nauk technicznych oraz dyscypliny z pogranicza nauk o Ziemi i nauk biologicznych

Dyscyplina z pogranicza nauk o Ziemi i nauk technicznych				
Parametr	**Wartość parametru**	**Zaokrąglona wartość prognozy**		**Średni względny błąd prognoz wygasłych (*ex post*)**
α	0.1	2016	354	0.104
β	0.2	2017	355	
Φ	0.9	2018	346	
		2019	327	
		2020	297	
Dyscyplina z pogranicza nauk o Ziemi i nauk biologicznych				
Parametr	**Wartość parametru**	**Zaokrąglona wartość prognozy**		**Średni względny błąd prognoz wygasłych (*ex post*)**
α	0.6	2015	138	0.233
β	0.7	2016	128	
Φ	0.7	2017	120	
		2018	113	
		2019	109	
		2020	107	

5. Podsumowanie i wnioski

Podsumowując, prognozy dalszego rozwoju widzianego przez pryzmat cytowalności ogółu dyscyplinarnej literatury naukowej dla sześciu wybranych dyscyplin, na lata 2015–2020, przedstawiają się tak, jak prezentuje to tabela 4. Prognozy te zostały uzyskane z użyciem różnych metod, w oparciu o dużą próbę losową (w 43 czasopismach wskazanych w znajdującym się

w pierwszej części niniejszego opracowania aneksie) opublikowano łącznie 1088 artykułów w języku polskim, w których zacytowano łącznie 24 582 pozycje literaturowe). Wykorzystane przez autorów metody prognostyczne inspirowane są w dużej mierze metodami ekonometrycznymi, a za najdokładniejszą z nich i jednocześnie najbardziej perspektywiczną, przynajmniej w świetle otrzymanych rezultatów i mierników dokładności prognoz, można uznać metodę regresji wielomianowej z poprawką ze względu na autokorelację składników resztowych.

Należy przy tym podkreślić, że natura zebranych danych uniemożliwiła zarazem dokładniejsze przetestowanie skuteczności metody regresji nieliniowej i linearyzowanej, do czego nawiązanie i wyjaśnienie znajduje się w części pierwszej niniejszego artykułu.

Podobnie rzecz miała się z testowaniem prognoz skonstruowanych z użyciem wielomianu stopnia drugiego, natomiast wielomian stopnia trzeciego pozwolił na wyprowadzenie przewidywań o umiarkowanie zadowalającym stopniu precyzji. Oceniając je według współczynnika „V_e" były one porównywalne z rezultatami prognozowania uproszczoną metodą autoregresji, a spoglądając na jakość prognoz przez pryzmat wielkości błędów *ex post* oraz *ex ante* były one z kolei nieco gorsze od metody autoregresji uproszczonej i autoregresji z poprawką ze względu na niestacjonarność modelowanego procesu. Prognozy dostarczone przez model wygładzania adaptacyjnego Holta usytuowały się natomiast w niejednoznacznej pozycji. Prognoza dla dyscypliny z pogranicza nauk o Ziemi i nauk technicznych cechuje się podobnym poziomem dokładności jak prognoza wywnioskowana z modelu regresji wielomianowej trzeciego stopnia, natomiast w przypadku dyscypliny z pogranicza nauk o Ziemi i nauk biologicznych prognoza ta stoi na widocznie niższej pozycji zarówno od pierwszej z dyscyplin granicznych, jak i prognoz zbudowanych z użyciem pozostałych metod.

Metoda adaptacyjna posiada również pewne inne mankamenty, które polegają po pierwsze na konieczności niejako sztucznego dostosowania zakresu badanych danych do wymogów wpisanych w istotę tej zastosowania, jak również – po drugie – na istnieniu pewnej dowolności w doborze konkretnych wartości parametrów „α", „β" i „Φ", które wpływają na ostateczny wynik i zarazem dokładność prognozy. Wyeliminowanie tej dowolności wymagałoby przetestowania wszystkich możliwych wartości wspomnianych parametrów, co wykracza znacznie poza zakres i cel niniejszej pracy i jest zadaniem o charakterze czysto obliczeniowym, do którego należałoby wykorzystać specjalistyczne narzędzia informatyczne.

Tab. 4. Podsumowanie prognoz dalszego rozwoju (jako przyszłej cytowalności) wszystkich wyróżnionych w badaniu dyscyplin

	Dyscyplina					
Rok	**Geografia fizyczna**	**Geografia społeczno-ekonomiczna**	**Geologia**	**Geofizyka i geochemia**	**Pogranicze nauk o Ziemi i nauk technicznych**	**Pogranicze nauk o Ziemi i nauk biologicznych**
2015	161	300	-	193	-	138
2016	165	300	188	192	354	128
2017	162	300	188	191	355	120
2018	157	298	187	188	346	113
2019	149	295	185	183	327	109
2020	139	290	182	177	297	107

Jeszcze jeden problem i niedostatek metodologiczny, który był obecny w przedstawionych powyżej analizach, to konieczność wykorzystania kilku różnych miar adekwatności i dokładności prognoz zbudowanych z użyciem różnych metod. Konieczność tę narzuciła autorom specyfika dostępnego im materiału badawczego, jak również ograniczenia i właściwości samych wykorzystanych przez nich metod, tym niemniej wydaje się, że korzystniej byłoby porównywać wartość uzyskiwanych wyników przy użyciu jednego i tego samego miernika. Zapewniłoby to swego rodzaju standaryzację pomiarów i poskutkowałoby z pewnością podniesieniem poziomu ich wiarygodności.

Na tym końcowym etapie analizy powstaje ponadto jeszcze jedno pytanie – w jaki sposób można przekształcić otrzymane prognozy punktowe w możliwie jednoznaczny i syntetyczny opis prognozy dalszego przebiegu zjawiska rozwoju dyscyplin, pozwalający na porównanie ich ze sobą i uszeregowanie ich w kolejności malejącej (lub rosnącej). Wydaje się, że najprostszym i zarazem czytelnym indeksem tego rodzaju byłoby np. określenie średniego procentowego tempa spadku wysokości prognozy dla kolejnych lat z przedziału 2015–2020. Ujęcie procentowe pozwoliłoby w tym kontekście na wyeliminowanie czynnika ilościowego, tj. ogólnego poziomu cytowalności (całkowitych liczb cytowań), który może wywoływać wrażenie, iż dyscypliny, w których cytowalność utrzymuje się na wyższym poziomie (jak np. geografia społeczno-ekonomiczna) rozwijają się szybciej od dyscyplin o niższym przeciętnym poziomie cytowalności, co naturalnie może, choć nie musi, być prawdą. Dla geografii fizycznej tego rodzaju wskaźnik byłby zdefiniowany na podstawie następujących faktów.

Liczba „161" stanowi w przybliżeniu 97.6% z liczby „165", co oznacza, że pomiędzy rokiem 2016 a 2015 nastąpił spadek cytowalności (jako spadek stopnia wykorzystania literatury dziedzinowej/dyscyplinarnej) o „100 – 97.6 = 2.4%". Liczba „162" stanowi 98.18% z liczby „165", liczba „157" stanowi 96.91% z liczby „162" itd. Średnia arytmetyczna wszystkich tak wyznaczonych spadków wynosi w tym przypadku: 3.8%. O tyle więc – średnio – spada prognozowana cytowalność (co można zinterpretować jako spowolnienie rozwoju[8]) w dyscyplinie geografii fizycznej z roku na rok, pomiędzy rokiem 2015 a 2020. W przypadku pozostałych dyscyplin ten prognostyczny „indeks natychmiastowości" (którego bardziej szczegółową interpretację podano we wcześniejszym opracowaniu autorów – Opaliński & Jaromin, 2017) przyjął wartości: 1.13% (geografia społeczno-ekonomiczna), 1.08% (geologia), 1.7% (geofizyka i geochemia), 4.3% (pogranicze nauk o Ziemi i nauk technicznych) i 4.9% (pogranicze nauk o Ziemi i nauk biologicznych). Ostatecznie można stwierdzić, że w perspektywie metodologii przyjętej w niniejszej pracy i mając na uwadze zastrzeżenia związane z wykorzystaniem różnych metod do konstrukcji prognoz zademonstrowanych w tabeli 4, dyscypliną rozwijającą się w najszybszym tempie (według stanu podyktowanego aktywnym wykorzystaniem dyscyplinarnej literatury przez badaczy, którzy publikowali swoje artykuły w roku 2015) była geologia. W ślad za nią podążają kolejno: geografia społeczno-ekonomiczna, geofizyka i geochemia, geografia fizyczna, pogranicze nauk o Ziemi i nauk technicznych i pogranicze nauk o ziemi i nauk biologicznych, którego dynamika rozwoju była najsłabsza. Można tu dodatkowo zaznaczyć, że prosta średnia arytmetyczna nie jest jedynym rodzajem średniej i istnieją jej alternatywne rodzaje. Przykładem może być tzw. średnia ważona, której przyjęcie pozwoliłoby na np. nadanie większego znaczenia

[8] Gdyby w tej perspektywie cytowalność wzrastała zamiast spadać (tj. publikacje najnowsze byłyby cytowane coraz szybciej i częściej), można byłoby mówić o przyspieszeniu tempa rozwoju dyscypliny.

(uwzględnienie silniejszego wpływu na ostateczny wynik) wartościom prognozowanym leżącym na osi czasu bliżej ostatniej znanej i wyznaczonej doświadczalnie wartości bądź np. wartościom prognozowanym, dla których stopień dokładności prognozy jest najwyższy.

Podsumowując, rozważania zamieszczone w obu częściach niniejszej pracy wydają się stwarzać szerokie pole dla dalszego namysłu, rozwijania, testowania i dopracowywania statystycznej metodologii prognostycznej znajdującej zastosowanie w naukoznawczych badaniach tempa rozwoju dyscyplin. Otrzymane jak dotąd rezultaty sugerują, że więcej uwagi należałoby poświęcić przede wszystkim różnym odmianom metody regresji wielomianowej, ale także metodom autoregresyjnym. Uwaga ta powinna skupić się na poprowadzeniu dalszych prób poprawy dokładności prognoz, na wykorzystaniu innego materiału empirycznego, na testowaniu innych rzędów modeli autoregresji, jak również na szerzej zakrojonym eksperymentalnym doborze wartości parametrów modeli wygładzania wykładniczego, czy też na próbach wykorzystania podejść pominiętych przez autorów niniejszego artykułu, np. metody konstruowania prognoz przedziałowych (tj. wyznaczenia prognozowanej liczby cytowań w postaci pewnego przedziału wartości, zamiast w postaci pojedynczej liczby), zamiast metody budowy prognoz punktowych. Postać przedziałów mogłaby również zostać nadana wyznaczonym doświadczalnie (i statystycznie istotnym) współczynnikom regresji, co wpłynęłoby na pewne „rozmycie" wartości prognozy, choć z drugiej strony powinno teoretycznie podnieść poziom jej wiarygodności (zmniejszyć szacunki jej błędu). Ponadto wydaje się, że warto również skupić uwagę na dążeniach do ujednolicenia stosowanych miar dokładności wyprowadzonych z przyjętych modeli prognoz. Mając na uwadze to szerokie pole przyszłych analiz należy stwierdzić, że obie przedstawione przez autorów części niniejszego artykułu powinny być potraktowane jedynie jako wstępne zarysowanie samej problematyki wykorzystania statystyki matematycznej w prognozowaniu naukoznawczym (i naukometrycznym) oraz jako pokazanie istniejących w tym obszarze wielorakich ewentualności, nie mniej licznych problemów i przede wszystkim tkwiącego w nim potencjału.

Bibliografia

Aczel, A. D., Sounderpandian, J. (2018). *Statystyka w zarządzaniu*. Warszawa: PWN.

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

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

Chatfield, Ch. (1975). *The Analysis of Time Series: Theory and Practice*. London: Chapman and Hall, https://dx.doi.org/10.1007/978-1-4899-2925-9

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

Guzik, B., Appenzeller, D., Jurek, W. (2007). *Prognozowanie i symulacje. Wybrane zagadnienia*. Poznań: Wydaw. Akademii Ekonomicznej w Poznaniu.

Huber, J. C. (1998). Cumulative Advantage and Success-Breeds-Success: The Value of Time Pattern Analysis. *Journal of the American Society for Information Science*, 49(5), 471–476, https://dx.doi.org/10.1002/(SICI)1097-4571(19980415)49:5<471::AID-ASI8>3.0.CO;2-T

Hyndman, R., Koehler, A. B., Ord, J. K., Snyder, R. D. (2008). *Forecasting with Exponential Smoothing : The State Space Approach*. Berlin: Heidelberg: Springer, https://dx.doi.org/10.1007/978-3-540-71918-2

Krawiec, S. (2014). *Adaptacyjne modele wygładzania wykładniczego jako instrumenty prognozowania krótkoterminowego zjawisk ilościowych*. Gliwice: Wydaw. Politechniki Śląskiej.
Krzysztofiak, M., Luszniewicz, A. (1976). *Statystyka*. Warszawa: Polskie Wydawnictwo Ekonomiczne.
Kuhn, T. S. (2001). *Struktura rewolucji naukowych*. Warszawa: Fundacja Aletheia.
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., Jaromin, M. (2017). Zastosowanie statystycznej analizy szeregów czasowych do krótkoterminowego prognozowania rozwoju dyscyplin naukowych. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 55(2), 106–125, https://doi.org/10.36702/zin.368
Pawłowski, Z. (1981). *Elementy ekonometrii: podręcznik*. Warszawa: PWN.
Price, D. de Solla (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, 27(5), 292–306, https://dx.doi.org/10.1002/asi.4630270505
Ross, S. M. (2009). *Introduction to Probability and Statistics for Engineers and Scientists*. Amsterdam: Elsevier Academic Press, https://dx.doi.org/10.1016/B978-0-12-370483-2.X0001-X
Sen, A., Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications*. Berlin: Heidelberg: Springer, https://dx.doi.org/10.1007/978-1-4612-4470-7
Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures. Fourth Edition.* Boca Raton: London: New York: Chapman & Hall/CRC, Taylor & Francis Group.
Shumway, R. H., Stoffer, D. S. (2011). *Time Series Analysis and Its Applications: With R Examples*. Berlin: Springer International Publishing, https://dx.doi.org/10.1007/978-3-319-52452-8
Snarska, A. (2011). *Statystyka, ekonometria, prognozowanie: ćwiczenia z Excelem 2007*. Warszawa: Wydawnictwo Placet.
Sobczyk, M. (2008). *Prognozowanie: teoria, przykłady, zadania*. Warszawa: Wydawnictwo Placet.
Sobczyk, M. (2015). *Statystyka*. Warszawa: PWN.
Witkowska, D. (2005). *Podstawy ekonometrii i teorii prognozowania: podręcznik z przykładami i zadaniami*. Kraków: Oficyna Ekonomiczna.
Wojna, A. (2007). *Predykcja ekonometryczna oraz modelowanie stochastyczne.* Cz.1. Koszalin: Wydawnictwo Politechniki Koszalińskiej.
Zaiontz, Ch. (2017). *Time Series Analysis* [online]. Real Statistics Using Excel, [21.11.2019], https://www.real-statistics.com/time-series-analysis/
Zeliaś, A. (1997). *Teoria prognozy*. Warszawa: Polskie Wydawnictwo Ekonomiczne.
Zeliaś, A., Pawełek, B., Wanat, S. (2013). *Prognozowanie ekonomiczne: teoria, przykłady, zadania*. Warszawa: PWN.

Aneksy

Aneks 1. *Wyniki badania liczby i wieku źródeł piśmienniczych przywoływanych w bibliografiach załącznikowych artykułów opublikowanych w czasopismach stanowiących przedmiot badania, z podziałem na dyscypliny, do których należały artykuły cytujące* [online]. Figshare database, [27.01.2020], https://figshare.com/s/a1d147ee9baae900b789

Selected Statistical Methods of Trend Analysis and Predicting the Rate of the Development of Scientific Disciplines (Third Degree Polynomial Regression Method, Autoregression Method and Exponential Smoothing)

Abstract

Purpose/Thesis: The article compares several statistical methods, which can be used to forecast the rate of the evolution of scientific disciplines. The data sample comprised the citations of specific scientific publications. The article emphasized the possibility of generalization of the results yielded by the analysis of this random sample. It also highlighted the limitations of each forecasting method, and proposed rough solutions to these problems.
Approach/Methods: The authors used a data set comprising almost 25 thousands of citations. They applied several distinct statistical methods inspired by econometric models. These were polynomial regression method, regression method with a correction for the autocorrelation in the residual components, an autoregression method, an autoregression method with the correction for non-stationarity of modelled process and Holt's adaptive model of exponential smoothing. The regression methods were tested for the fulfillment of the Gauss-Markov conditions. Moreover, common accuracy measures, as well as the prognosis errors coefficients were calculated and compared for all the methods applied.
Results and conclusions: The analysis showed that the most precise method was the polynomial regression method with a correction for the autocorrelation in residual components. The reliability of the autoregression method is comparable with that of the regression methods. The adaptive exponential smoothing method yielded ambiguous results. This suggests directions for further research.
Research limitations: The basic limitation of this study was the range of empirical data available to the authors, which was restricted to a single scientific discipline, and, further, limited to Polish-language texts published in journals (periodicals).
Originality/Value: The originality of the study lies in the innovative juxtaposition of the well-known quantitative methods, which have not been used to predict the rate of the development of scientific disciplines before. Secondly, the study shows their potential in this field of inquiry, and makes clear the need for further improvement. The study identifies the most promising methodology, which may open the way for the better understanding of science's internal dynamics.

Keywords

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

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 *napisanej pod kierunkiem dr hab. Remigiusza Sapy z Instytutu Informacji Naukowej i Bibliotekoznawstwa UJ. Pracuje w Oddziale Informacji Naukowej Biblioteki Politechniki Rzeszowskiej na stanowisku kustosza. Najważniejsze publikacje: Opaliński, Ł. (2019). Cytowanie narzędziem zarządzania informacją: teoria zachowań informacyjnych. W: W. Babik (red.)* Zarządzanie informacją *(210–248). Warszawa: Wydawnictwo SBP; 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 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
Oddział Informacji Naukowej Biblioteki Politechniki Rzeszowskiej
Al. Powstańców Warszawy 12, bud. V-B, pok. V-B 105
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 biotechnologii, 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
Wydział Chemiczny, Politechnika Rzeszowska
al. Powstańców Warszawy 6, bud. H, pok. H-242
35-959 Rzeszów