Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Instytut Automatyki i Informatyki Stosowanej

Praca dyplomowa magisterska

na kierunku Informatyka
w specjalności Systemy informacyjno-decyzyjne

Analiza motywów doboru cytowań w publikacjach naukowych

Jacob Tarasiewicz
Numer albumu 221845

promotor
dr inż. Mariusz Kamola

Warszawa 2017

Temat: Analiza motywów doboru cytowań w publikacjach naukowych

Streszczenie:

Celem pracy było wykonanie analizy doboru cytowań w publikacjach naukowych oraz weryfikacja postawionych hipotez dotyczących wpływu statystyk autorów na publikacje artykułu w renomowanych czasopismach.

W celu dokonania analizy cytowań w publikacjach naukowych konieczne było posiadanie danych o artykułach naukowych, czasopismach oraz samych autorach. Należało znaleźć i wybrać odpowiednie źródło danych. Spośród znalezionych baz danych wybrana została baza Scopus utrzymywana przez wydawnictwo Elsevier. Zostały stworzone specjalne moduły, aby móc pobierać dane. W celu przechowywania pobranych już danych została zaprojektowana i stworzona baza danych, do której moduły zbierające zapisują dane. Następnie została przeprowadzona analiza zebranych danych.

W pracy pokazano kolejne etapy powstawania pracy, opisano napotkane problemy oraz ich rozwiązania. Pokazano przebieg analizy oraz wynikające z niej wnioski. Zostały zweryfikowane postawione hipotezy.

Słowa kluczowe: cytowania, wpływ na publikację, impact factor, h-indeks

Title: Analysis of motives of selection of citations in scientific publications

Abstract:

The aim of the thesis was to perform analysis of the selection of citations in scientific publications and verification of stated hypotheses concerning the impact of authors statistics on the publishing article in prestigious journals.

In order to analyze the citations in scientific publications, it was necessary to have information about scientific articles, journals and authors. It was necessary to find appropriate data source. Among the found databases Scopus database has been selected, which is maintained by Elsevier. To download data special modules has been created. To store downloaded data database has been designed and created, to which downloading modules writes data. Than analysis of collected data has been performed.

In the thesis has been shown next steps of preparing the study. Encountered problems has been described with solutions. There has been shown the process of analysis and it conclusions. Stated hypothesis has been verified.

Keywords: citation, the impact of the publication, impact factor, h-index

Spis treści

1 Cel i motywacja pracy

Głównym celem pracy było przeprowadzenie analizy motywów doboru cytowań w publikacjach naukowych. W szczególności sprawdzenie czy dobór cytowanych publikacji ma wpływ na ukazanie się artykułu w lepiej punktowanych czasopismach. Analiza została przeprowadzona na publicznie dostępnych danych o artykułach naukowych oraz ich twórcach, które zostały pobrane przez specjalnie zaimplementowane moduły zbierające oraz zapisane w lokalnej bazie danych. Poza głównym celem zostały również przeprowadzone badania w poszukiwaniu także innych zależności związanych z autorami i ich przeszłymi publikacjami mających wpływ na publikację w bardziej renomowanych czasopismach.

Motywacją powstania pracy był między innymi artykuł [1], w którym poruszana zostaje przez autorów kwestia liczby cytowań danego artykułu rozłożona w czasie. Zostaje stwierdzone, że na podstawie danych o liczbie cytowań pracy w ciągu pierwszych 5 lat można stwierdzić ile cytowań będzie w kolejnych latach, w swoich badaniach pokazują poprawne przewidywanie dla kolejnych 20 lat. Chciałem to odnieść do swojej pracy i sprawdzić czy na podstawie historii publikacji autora możliwa jest estymacja jego przyszłych publikacji.

Kolejnym powodem powstania pracy był także artykuł [2] dotyczący oszustw naukowców w publikowaniu. Opisuje on działania pewnego francuskiego badacza Labbé, który stworzył narzędzie do wyszukiwania wygenerowanych przez program Scigen publikacji. W ten sposób znalazł kilkadziesiąt komputerowo wygenerowanych publikacji z różnych konferencji naukowych, które zostały wydane, z czego ponad 100 zostało opublikowane przez IEEE. Proceder ten miał się odbywać od 2008 do 2013 roku, w którym to Labbé wykrył oszustwo i zgłosił je wydawnictwom. Wydawnictwa poinformowane przez badacza miały usunąć sfałszowane publikacje. Większość konferencji, z których materiały zostały opublikowane odbywały się w Chinach i ich autorami w większości byli również chińscy naukowcy. Aczkolwiek zastanawiające jest to, że taki proceder jest w ogóle możliwy. Przytaczając powyższy artykuł należy również wspomnieć o publikacji [3] autorstwa Labbé, w której opisuje jak oszukał serwis Google Scholar za pomocą fikcyjnego autora „Ike Antkare", który uzyskał w ciągu roku indeks Hirscha wynoszący 94.

Powyższe artykuły mogą sugerować, że są czynniki niewynikające z samej treści artykułu, które mają wpływ na publikacje. Utwierdził mnie w tym przekonaniu kolejny artykuł [4], który traktuje o słabościach recenzji w czasopismach naukowych. W 2013 roku autor wymienionej publikacji stworzył artykuł o domniemanym lekarstwie na raka, który według autora nie powinien przejść żadnej recenzji pozytywnie. Uważał, że sposób przeprowadzenia badań był tak niedbały, że jego wyniki nie mają znaczenia. Artykuł wysłał do ponad 300 redakcji, z czego ponad połowa przyjęła artykuł do publikacji po jego recenzji.

Artykuł został zaakceptowany m.in. przez wydawnictwa Elsevier oraz Sage. Zaledwie około 100 odpowiedzi od redakcji było z uwagami recenzentów, z czego większość miała uwagi jedynie do samej formy artykułu, a nie treści. Tylko 36 odpowiedzi na 306 wysłanych artykułów zawierało naukowe komentarze. Skoro tylko około 10% odpowiedzi zawierało naukowe uwagi, można wnioskować, że na publikacje mają wpływ inne czynniki niż sama treść artykułu i przeprowadzone w nim badania.

W pracy sprawdzono, czy na publikację mają wpływ: statystyki autorów, bibliografia artykułu a także statystyki autorów cytowanych publikacji. Zbadane zostały też zależności pomiędzy kolejnymi artykułami autora pod względem możliwości aproksymacji współczynników wpływu czasopism jego przyszłych publikacji.

2 Analiza zagadnień

2.1 Publikacja naukowa

Powołując się na definicję zawartą w Wikipedii [7] za publikację naukową uznaje się pracę naukową opublikowaną w czasopiśmie naukowym bądź książce, zawierającą oryginalne badania lub odnoszących się do już istniejących. Publikacja naukowa powinna również spełniać pewne formalne wymogania. Na podstawie komunikatu Ministra Nauki i Szkolnictwa Wyższego [8] należą do nich m.in.:

- tytuł publikacji,
- listę autorów wraz z afiliacją,
- cytowana literatura (bibliografia).

W mojej pracy odnoszę się tylko do publikacji naukowych zawartych w czasopismach naukowych. Miało to na celu umożliwić poprawną klasyfikację poszczególnych artykułów, z uwagi na punktację czasopism naukowych.

2.2 Czasopismo naukowe

Zgodnie z definicją zawartą w Wikipedii [9] w czasopiśmie naukowym ukazują się artykuły, które przed publikacją są recenzowane, w celu zapewnienia o ich jakości. Jeśli artykuł przedstawia badania, eksperymenty lub obliczenia muszą one być tak opisane, aby niezależny naukowiec mógł powtórzyć opisywane badania oraz potwierdzić otrzymane przez autora rezultaty.

NATURE

A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE

"To the solid ground
Of Nature trusts the mind which builds for aye."—WORDSWORTH

THURSDAY, NOVEMBER 4, 1869

NATURE: APHORISMS BY GOETHE

NATURE! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her.

Without asking, or warning, she snatches us up into her circling dance, and whirls us on until we are tired, and drop from her arms.

She is ever shaping new forms: what is, has never yet been; what has been, comes not again. Everything is new, and yet nought but the old.

We live in her midst and know her not. She is incessantly speaking to us, but betrays not her secret. We constantly act upon her, and yet have no power over her.

The one thing she seems to aim at is Individuality; yet she cares nothing for individuals. She is always building up and destroying; but her workshop is inaccessible.

Her life is in her children; but where is the mother? She is the only artist; working-up the most uniform material into utter opposites; arriving, without a trace of effort, at perfection, at the most exact precision, though always veiled under a certain softness.

Each of her works has an essence of its own; each of her phenomena a special characterisation: and yet their diversity is in unity.

She performs a play; we know not whether she sees it herself, and yet she acts for us, the lookers-on.

Incessant life, development, and movement are in her, but she advances not. She changes for ever and ever, and rests not a moment. Quietude is inconceivable to her, and she has laid her curse upon rest. She is firm. Her steps are measured, her exceptions rare, her laws unchangeable.

She has always thought and always thinks; though not as a man, but as Nature. She broods over an all-comprehending idea, which no searching can find out.

Mankind dwell in her and she in them. With all men she plays a game for love, and rejoices the more they win. With many, her moves are so hidden, that the game is over before they know it.

That which is most unnatural is still Nature; the stupidest philistinism has a touch of her genius. Whoso cannot see her everywhere, sees her nowhere rightly.

She loves herself, and her innumerable eyes and affections are fixed upon herself. She has divided herself that she may be her own delight. She causes an endless succession of new capacities for enjoyment to spring up, that her insatiable sympathy may be assuaged.

She rejoices in illusion. Whoso destroys it in himself and others, him she punishes with the sternest tyranny. Whoso follows her in faith, him she takes as a child to her bosom.

Her children are numberless. To none is she altogether miserly; but she has her favourites, on whom she squanders much, and for whom she makes great sacrifices. Over greatness she spreads her shield.

She tosses her creatures out of nothingness, and tells them not whence they came, nor whither they go. It is their business to run, she knows the road. Her mechanism has few springs—but they never wear out, are always active and manifold.

The spectacle of Nature is always new, for she is always renewing the spectators. Life is her most exquisite invention; and death is her expert contrivance to get plenty of life.

She wraps man in darkness, and makes him for ever long for light. She creates him dependent upon the earth, dull and heavy; and yet is always shaking him until he attempts to soar above it.

Rysunek 1 Okładka pierwszego wydania czasopisma Nature - 04.11.1869
Źródło: Wikipedia, https://en.wikipedia.org/wiki/Scientific_journal, 05.02.2017

Według raportu [10] w 2014 roku było na świecie około 34 tysięcy czasopism naukowych, w tym 28 tysięcy anglojęzycznych. Ważnym podzbiorem z tych czasopism jest zbiór wybrany w *Thomson Reuter's Journal Citation Report*. W 2016 roku [11] liczył on 11365

czasopism z 81 krajów. Czasopisma na tej liście posiadają wyliczony współczynnik miary danego czasopisma – Impact Factor (z ang. wskaźnik wpływu) – IF.

2.3 Wskaźnik wpływu

Wskaźnik wpływu (Impact Factor) został powszechnie wprowadzony w corocznych raportach *Journal Citation Reports* w 1975 roku. Jego zadaniem jest sklasyfikowanie czasopism mierząc jego oddziaływanie na świat naukowy. Współczynnik ten wylicza się bazując na liczbie artykułów w czasopiśmie w stosunku do liczby cytowań artykułów z tego czasopisma z ostatnich dwóch lat.

$$IF = B/C$$

- IF – Wskaźnik wpływu na dany rok.
- B – Liczba cytowań artykułów z tego czasopisma w ciągu dwóch poprzednich lat.
- C – Liczba artykułów opublikowanych w tym czasopiśmie w ciągu dwóch poprzednich lat.

Podstawową rolą wskaźnika wpływu – IF jest zgromadzenie czasopism naukowych, które najsilniej oddziałują na świat nauki. Znalezienie się na liście tych czasopism jest bardzo prestiżowe. Lista sklasyfikowanych czasopism z IF jest uznawana na całym świecie, jako pewien wyznacznik najlepszych czasopism naukowych.

Wskaźnik wpływu – Impact Factor ma również inne zastosowania. Jest powszechnie uznany przez autorów artykułów, którzy się starają o przyjęcie swoich prac w sklasyfikowanych czasopismach. Z drugiej strony IF stosuje się również w celu odniesienia się do sukcesu naukowego danej osoby. Często jest on miarą osiągnieć naukowca, co nie do końca musi być miarodajne.

Wskaźnik ten również stał się ważny dla instytucji naukowych, których sukcesy lub aktualny status w świecie nauki mierzy się za pomocą liczby artykułów w czasopismach znajdujących się na listach *Journal Citation Reports*. Status instytutu zależy od liczby publikacji jego pracowników w punktowanych czasopismach.

W wielu państwach powszechnie stosowane są systemy ocen instytucji naukowych w oparciu o IF. W Polsce instytucje naukowe są oceniane zgodnie z rozporządzeniem wydanym przez Ministerstwo Nauki i Szkolnictwa Wyższego [12], w którym na ocenę składają się między innymi łączne punkty za publikacje w wybranych czasopismach. Obecnie ministerstwo publikuje wykazy czasopism wraz z punktacją przyznawaną za każdy artykuł opublikowany w danym czasopiśmie [13]. Wykaz ten składa się z 3 części:

- A – zawierająca wyłącznie czasopisma sklasyfikowane w *Journal Citation Reports.*
- B – zawierająca czasopisma nieposiadające współczynnika wpływu – IF, głównie polskie czasopisma.
- C – zawierające czasopisma znajdujące się w bazie *European Reference Index for the Humanities* (ERIH).

Najwięcej punktów jest przyznawane za publikację w czasopismach z wykazu części A, posiadających współczynnik wpływu – IF.

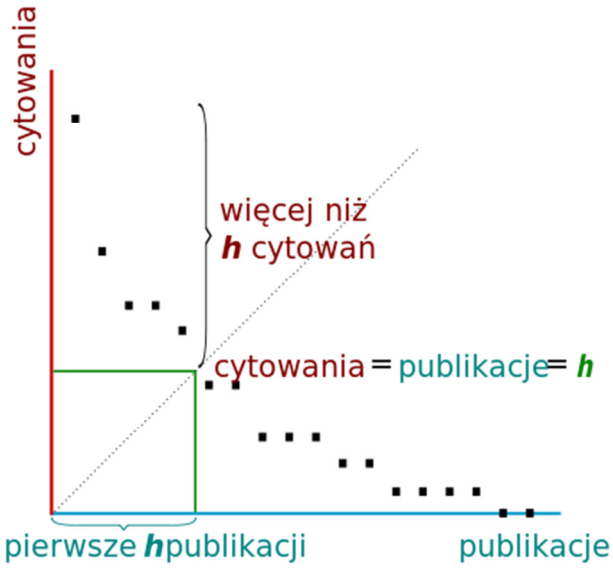
2.4 Miara dorobku naukowca

Podobnie jak dla czasopism, również dla autorów wprowadzono pewne wskaźniki. Zastosowanie ich ma na celu wyznaczenie dorobku naukowego danej osoby. Jedną z najprostszych metod jest liczenie bezpośrednio liczby publikacji, ale może to być niewystarczające. Pojawia się między innymi problem jakościowy – czy każda publikacja jest tak samo istotna, czy każda publikacja wnosi identyczną wartość do dorobku naukowca?

Inną metodą miary dorobku naukowego może być sumowanie współczynników wpływu czasopism, w których ukazywane były artykuły danego autora. Również i tej metodzie można zastrzec, czy dobrym artykułem naukowym czyni go tylko to, że został opublikowany w renomowanym czasopiśmie.

Kolejną miarą dorobku naukowego może być liczba cytowań artykułów danego autora. Uwzględniane są wszystkie publikacje naukowca i sprawdzana jest liczba odniesień we wszystkich innych publikacjach do prac autora. Ta metoda pozwala określić pewien wpływ autora na świat nauki. Jednak w dalszym ciągu można tę metodę zakwestionować. Czy można porównywać autora, który ma jeden artykuł z dużą liczbą cytowań z autorem, który ma wiele artykułów z większą liczbą cytowań, ale sumarycznie mniejszą niż autor z jednym artykułem?

Ostatnią przytoczoną metodą jest klasyfikowanie za pomocą h-indeksu [14] – wskaźnika Hirscha. Tak jak w poprzedniej metodzie uwzględnia się cały dorobek autora. Jednak wskaźnik Hirscha wyróżnia to, że mówi o tym ile prac autora zostało, co najmniej tyle razy cytowane przez innych autorów. Wyliczany on jest na podstawie danych o cytowaniach prac naukowca. Zgodnie z definicją naukowiec ma wskaźnik wynoszący *h*, jeżeli co najmniej *h* publikacji spośród jego wszystkich publikacji było cytowana *h* razy każda oraz każda z pozostałych publikacji ma mniej niż *h* cytowań.



Rysunek 2 Sposób liczenia h-indeksu
Źródło: https://pl.wikipedia.org/wiki/Plik:H-index-pl.svg, dostęp 28.01.2017

Na rysunku 2 pokazano sposób obliczania indeksu Hirscha. Istnieje wiele możliwości mierzenia dorobku naukowców. Stosuje się je w zależności od sytuacji i żądanego efektu.

3 Pobór danych

3.1 Przygotowanie do realizacji

Realizacja pracy magisterskiej rozpoczęła się od sformułowania szczegółowych zadań na podstawie ogólnie przyjętych założeń. Aby zrealizować cele pracy, czyli zweryfikować postawione hipotezy należało uzyskać dostęp do różnych danych. Na tej podstawie w porozumieniu z opiekunem pracy zostały ustalone następujące minimalne wymagania dotyczących danych:

- dane o publikacji naukowej:
  - tytuł, DOI,
  - lista autorów,
  - lista cytowań,
  - tytuł, ISSN czasopisma naukowego;
- dane o autorach:
  - imię, nazwisko,
  - afiliacja;
- dane o czasopismach:
  - współczynnik punktowy.

Dodatkowym, ale obowiązkowym wymaganiem była możliwość jednoznacznej identyfikacji autorów. Należało przygotować mechanizm, który pozwoli na rozróżnienie autorów. Wielu autorów może tak samo się nazywać. W artykułach naukowych często podawane są tylko inicjały imienia, co może prowadzić do błędnej identyfikacji, kiedy kilku autorów ma to samo nazwisko i imię zaczynające się od tej samej litery. Można dodać kolejny poziom rozpoznania autora ze względu na jego afiliację, ale to również może prowadzić do błędów.

Kolejnym aspektem było wydzielenie specjalnego atrybutu pozwalającego w pewien sposób agregować powyższe dane. Takimi atrybutami mogłyby być np. kraj afiliacji autorów lub dziedzina, w której artykuł jest publikowany. Powszechnie wiadomo jest, że artykuły z różnych dziedzin znacząco różnią się zarówno liczbą cytowań, jak i ilością autorów publikujących pracę lub sposobem recenzowania.

Mając tak postawione wymagania dotyczących danych można było przystąpić do implementacji bazy danych w celu przechowywania znalezionych danych.

3.2 Wykorzystana technologia

Dużą częścią wykonanej pracy było pobranie danych oraz ich obsługa, aby móc je analizować i mieć do nich stały dostęp. Aby to zrealizować konieczne było napisanie własnych programów do wykonania tych zadań. Zdecydowałem się implementować moduły pobierające w języku JAVA, ponieważ posiadałem już doświadczenie w tego typu zadaniach. Do przechowywania danych, oczywistym rozwiązaniem było stworzenie bazy danych.

3.2.1 Sposób pobierania danych

Programowanie w języku JAVA daje możliwość korzystania z wielu dostępnych darmowych bibliotek. Jedną z tych bibliotek, których postanowiłem wykorzystać do pobierania danych była *JSoup* [26]. Jest to biblioteka umożliwiająca wysyłanie zapytań HTTP i otrzymywania odpowiedzi. Wystarczy w tym celu podać URL strony internetowej oraz opcjonalnie dodać dodatkowe pola do zapytania. W odpowiedzi otrzymywany jest dokument HTML, zawierający treść odpowiadającą na wysłane zapytanie. Z takiego dokumentu można wyłuskać wszelkie potrzebne informacje za pomocą znaczników HTML. W swojej pracy korzystałem z RESTful API, które w odpowiedzi na zapytania zwracało wyniki w formie XML. Z tego powodu zdecydowałem się używać biblioteki *Jackson* [27]. Biblioteka ta umożliwia proste parsowanie dokumentów HTML w celu uzyskania żądanych pól XML w zadanym formacie. Okazała się ona bardzo pomocna przy zbieraniu danych.

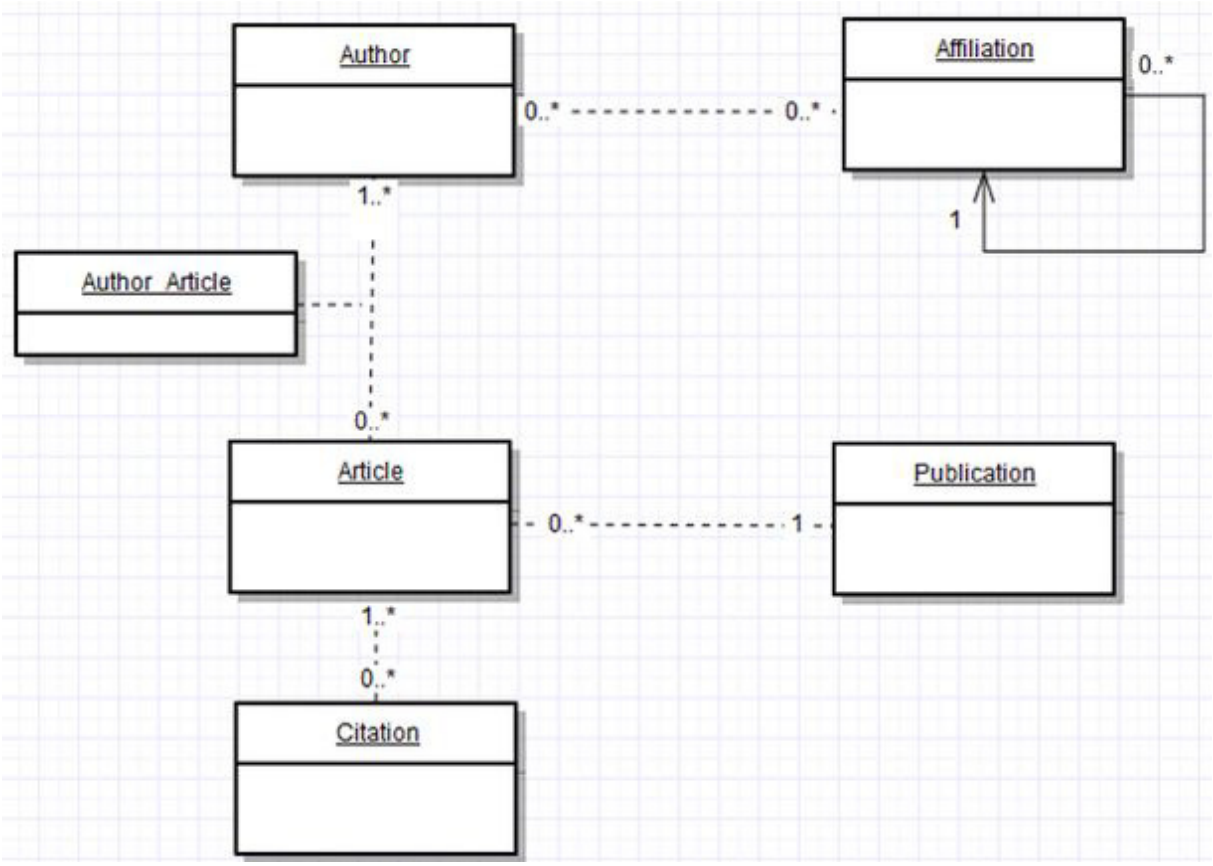
3.2.2 Sposób przechowywania oraz obsługi danych

W celu przechowywania pobranych danych stworzyłem lokalną bazę danych w technologii MySQL. Została ona opisana w rozdziale 3.2.3. Natomiast do zapisywania w bazie danych oraz jej obsługi skorzystałem z frameworka *Hibernate* [28]. Pozwala on na translację danych pomiędzy relacyjną bazą danych, a programami obiektowymi. Framework wykorzystuje w tym celu opis struktur danych w języku XML. Dzięki zastosowaniu *Hibernate* obsługa bazy danych odbywała się praktycznie automatycznie. Wszelkie zapisy, odczyty oraz modyfikacje obiektów odbywały się tylko za pomocą wywołania odpowiedniej metody, bez potrzeby pisania poleceń SQL. Wstępna konfiguracja tego frameworka była dosyć skomplikowana, ale była to inwestycja, która się z pewnością zwróciła. Aczkolwiek z zapytań SQL również korzystałem w celu wykonania złożonych zapytań do bazy.

W Dodatku Dodatek A – Fragmenty kodu źródłowego zostały przedstawione zastosowania powyżej opisywanych technologii w praktyce.

3.2.3 Baza danych

Do przeprowadzenia analiz potrzebny był stały i szybki dostęp do danych. W tym celu została stworzona relacyjna baza danych MySQL działająca na lokalnym serwerze. W jej architekturze zostały uwzględnione wymagania w tym również możliwość jej rozbudowy w celu zapewnienia integralności danych.



Rysunek 3 Schemat bazy danych

Na rysunku 3 przedstawiono architekturę bazy danych wraz z powiązaniami między tabelami. Głównymi tabelami są:

- Author – przechowuje dane o autorach.
- Affiliation – przechowuje dane o afiliacjach, tabela hierarchiczna, zawierająca odwołania do samej siebie w przypadku posiadania przez instytut instytutu nadrzędnego (np. wydział – uczelnia).
- Publication – przechowuje dane o czasopismach naukowych.
- Article – przechowuje dane o artykułach.
- Citation – przechowuje dane o cytowaniach.

Występują również tabele asocjacyjne, pozwalające na związki wiele do wielu:

- Author-Article – powiązania autorów do artykułów.
- Author-Affiliation – powiązania autorów do instytutów.

Posiadając gotową bazę danych można było przystąpić do szukania serwisu, który mógłby zapewnić dane do zapełnienia bazy danych. Dopiero mając dostęp do takich danych można było zająć się dalszą częścią pracy – analizą.

3.3 Poszukiwanie danych

Informacje o publikacjach naukowych są często łatwo dostępne na różnych serwisach internetowych, jednakże problematycznym może się okazać to, jakiego typu dane są potrzebne. W przypadku mojej pracy dane te były różnorodne, zarówno informacje o samej publikacji jak i o autorach, jak i o czasopismach. Należało znaleźć serwisy, które takie dane mogą dostarczyć w taki sposób, żeby można było je ze sobą powiązać, mimo pochodzenia z różnych środowisk.

Poszukiwania rozpoczęły się od znalezienia serwisów, które zajmują się pracami naukowymi. Zostały wytypowane następujące serwisy/źródła danych:

- Google Scholar,
- Web of Science,
- Polska Bibliografia Naukowa,
- ResearchGate,
- lista ministerialna czasopism punktowanych,
- Scopus.

3.4 Google Scholar

Poszukiwania rozpoczęły się od serwisu Google Scholar[15], jako że Google słynie z ogromu danych, jaki posiada w swoich bazach. Kolejnym atutem jest posiadanie szeregu API do obsługi swoich serwisów. Serwis Google Scholar jest darmową wyszukiwarką z otwartym dostępem dla każdego. Pozwala znaleźć informacje o artykułach naukowych. Według znalezionych informacji [16] serwis ten posiada największą bazę danych artykułów w skład, w której wchodzi również baza Web of Science. Google Scholar wydawał się być najodpowiedniejszym serwisem do realizacji mojej pracy.

Pierwszym problemem, który się pojawił był brak API Google do obsługi tego serwisu, ale była to trudność, którą można obejść implementując własny program – moduł zbierający dane. Kolejnym z problemów było wyszukiwanie artykułów. Nie było możliwości wyszukiwania po dziedzinie lub kraju pochodzenia autorów, jedyną opcją umożliwiającą pobranie ważnych danych było wyszukiwanie po autorach publikacji bądź po tytułach czasopism lub przez wyszukiwanie konkretnych artykułów przez ich tytuł lub DOI.

Portal umożliwia pobranie opisu bibliograficznego dla danego artykułu w jednym z proponowanych formatów. Wśród dostępnych formatów jest BibTex, który został wybrany ze względu na swoje podobieństwo do formatu XML. Format ten był często wykorzystywany w pracy przy parsowaniu stron.

```
@article{niewiadomska2014dynamic,
title={Dynamic power management in energy-aware computer networks and data intensive computing systems},
author={Niewiadomska-Szynkiewicz, Ewa and Sikora, Andrzej and Arabas, Piotr and Kamola, Mariusz and Mincer, Marcin and Ko{\l}odziej, Joanna},
 journal={Future Generation Computer Systems},
 volume={37},
 pages={284--296},
 year={2014},
 publisher={Elsevier}
}
```

Rysunek 4 Opis bibliograficzny w formacie BibTex
Źródło: https://scholar.googleusercontent.com/scholar.bib?q=info:ZuIrCxYe85MJ:scholar.google.com/&output=citation &scisig=AAGBfm0AAAAAWJdC3gDVmUy6Ail29B2T6O3O_Sm5kotE&scisf=4&ct=citation&cd=-1&hl=pl , 05.02.2017

Na rysunku 4 przedstawiono przykładowy opis, dzięki takiemu rozwiązaniu wiadomo było, który artykuł posiada podstawowe atrybuty jak rok publikacji, tytuł czasopisma oraz listę autorów.

Po krótkiej analizie okazało się, że identyfikowanie autorów bezpośrednio przez bazę Google Scholar może być również problematyczne. Ze względu na fakt, że Google prowadzi bazę tylko dla autorów, którzy się zarejestrowali w ich serwisie i są do nich przypisane prace naukowe. Pozostali autorzy posiadają jedynie dane widoczne na liście autorów, która składa się z nazwiska i imienia autora (czasem tylko inicjał imienia) – identyfikacja autorów w ten sposób staje się bardzo utrudniona. Bardzo złożonym problemem jest identyfikacja autora na podstawie tylko jego nazwiska. Jednoznaczne rozróżnienie staje się wręcz niemożliwe. Z powyższych powodów zdecydowałem się na oddzielne pobranie bazy autorów oraz artykułów, a następnie wyszukanie ich cytowań w bazie Google Scholar. Należało jeszcze przeanalizować inne serwisy.

3.5 Web of Science

Baza Web of Science [18] jest prowadzona przez wydawnictwo *Thomson Reuters*; odpowiada ono również za prowadzenie i publikację corocznych raportów zawierających aktualne czasopisma naukowe z ich współczynnikiem wpływu na dany rok – *Journal Citation Reports*. Z tego powodu to źródło danych również wygląda na odpowiednie. Dostęp do Web of Science jest ograniczony, wymaga on posiadania specjalnego konta założonego przez zarejestrowany w portalu instytut lub dostęp przez sieć instytutu.

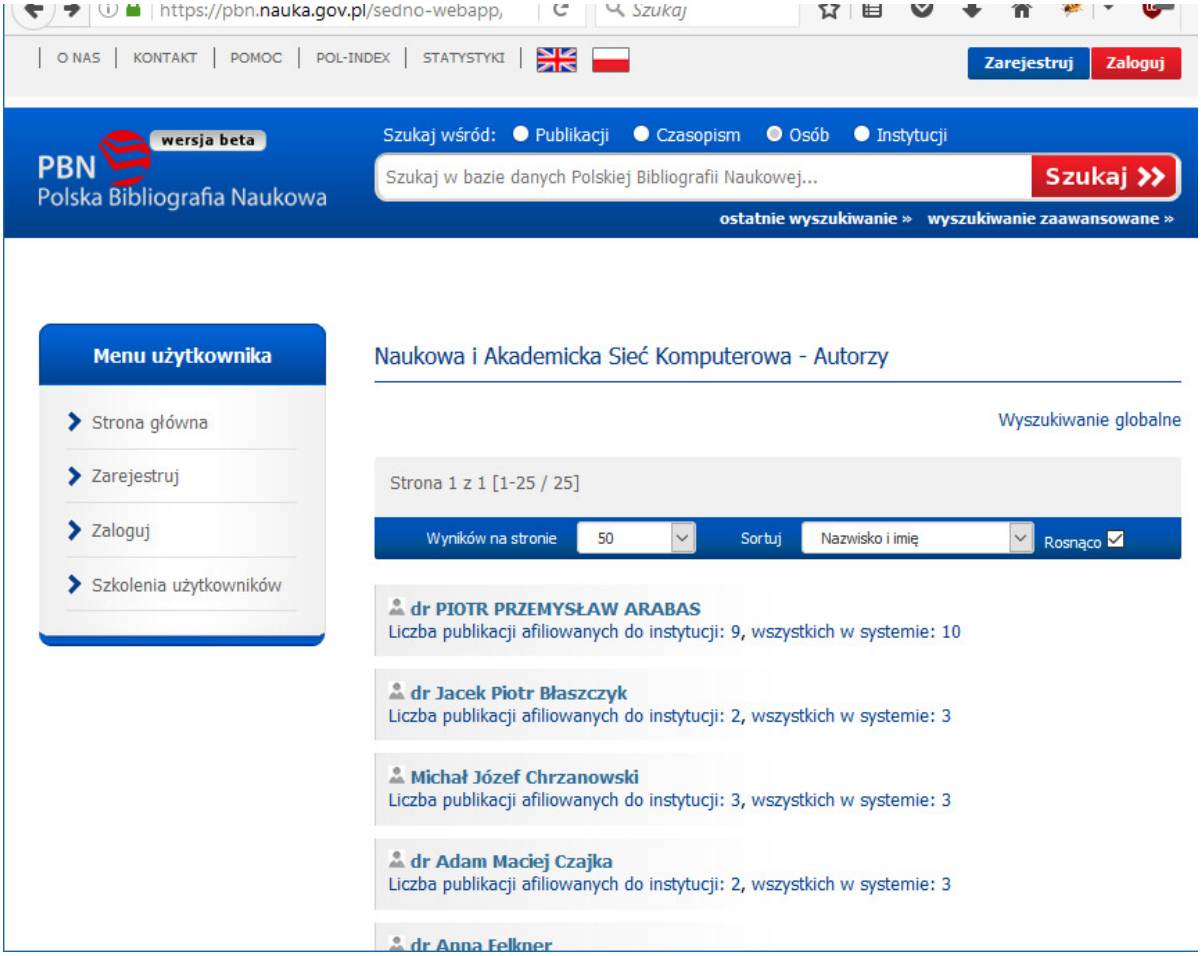
W moim przypadku korzystałem z dostępu przez wydziałową sieć. Aby móc być połączonym z dowolnego miejsca, nie będąc fizycznie połączonym przez sieć wydziałową

skorzystałem z połączenia tunelowego poprzez SSH [19], korzystając z własnego konta na serwerze MION. Przez konfigurację połączenia SSH mogłem w dowolnej aplikacji, programie ustawić połączenie przez serwer PROXY i korzystać z niej tak jakby komputer był fizycznie połączony z siecią wydziałową. Używałem tego serwera PROXY w przeglądarce internetowej, aby analizować serwisy oraz w modułach zbierających, napisanych przeze mnie w celu pobrania danych. Z tego rozwiązania korzystałem w ten sam sposób używając innych serwisów wymagających dostępu poprzez sieć instytutu.

Web of Science również nie posiadał API do korzystania ze swoich zasobów. Mając dostęp zarówno do bazy Web of Science jak i Google Scholar należało porównać oba te serwisy. Po wyszukaniu kilku pracowników Politechniki Warszawskiej okazało się, że w bazie Google było około 50% więcej artykułów sprawdzanych autorów.

3.6 Polska Bibliografia Naukowa

Kolejnym serwisem, przeze mnie badanym w celu znalezienia danych o artykułach i autorach, które poddane zostałyby analizie była Polska Bibliografia Naukowa – PBN [17]. Jest to portal prowadzony przez Ministerstwo Nauki i Szkolnictwa Wyższego, który agreguje informacje o czasopismach naukowych oraz polskich autorach publikacji naukowych jak i również samych publikacjach. Jest to darmowy serwis z otwartym dostępem. Korzystając z tego serwisu mógłbym zyskać wymagane informacje dotyczące artykułów oraz autorów z jednego kraju – Polski.



Rysunek 5 Przykładowy zrzut PBN
Źródło:www.pbn.nauka.gov.pl, dostęp 30.01.2017

Na rysunku 5 widać, że serwis oferuje przejrzystą wyszukiwarkę, oraz że do autorów

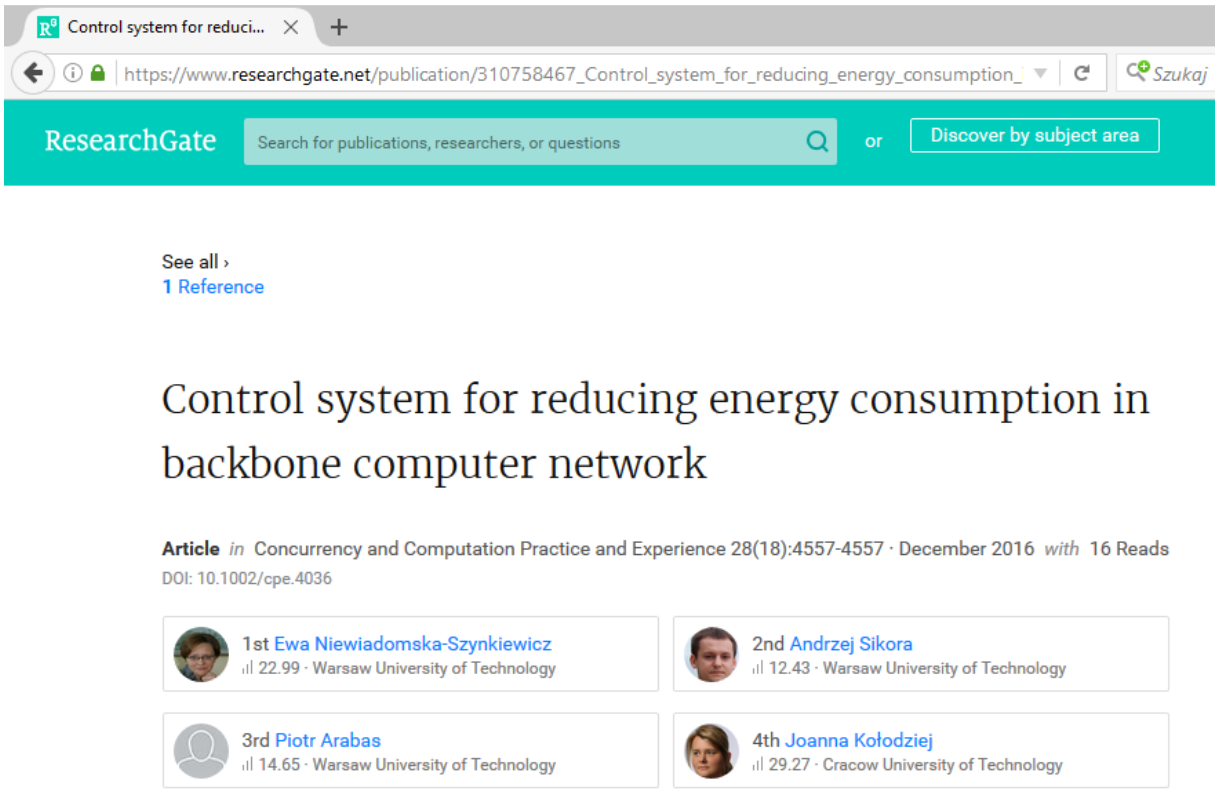
są przypisane publikacje. Jednak informacje zawarte o nich są bardzo często niepełne lub nawet szczątkowe. W zależności od artykułu mogą to być informacje takie jak DOI artykułu lub pełna bibliografia artykułu, a czasami nie będzie żadnej informacji poza tytułem. Natomiast dane o autorach są minimalistyczne i zawierają jedynie imię i nazwisko autora oraz niepotwierdzone informacje o jego artykułach. Sytuacja wygląda podobnie jak w przypadku serwisu Google Scholar, gdzie poza imieniem i nazwiskiem autora nie występują inne dane.

Z wyżej wymienionych powodów zdecydowałem się nie używać tego portalu do pobierania informacji ani o artykułach, ani o autorach.

3.7 ResearchGate

Portal ResearchGate[20] został założony w 2008 roku, i jak sam się ogłasza jest stworzony przez naukowców dla naukowców. Ma na celu współdzielenie się pracami naukowymi oraz nawiązywaniu współpracy, jak również tworzeniu statystyk naukowców i instytutów. Według danych podanych na stronie portalu korzysta z niego ponad 11 milionów badaczy z całego świata. Dostęp do serwisu wymaga jedynie rejestracji za pomocą e-maila instytutu naukowego. Aby mieć dostęp do serwisu stworzyłem konto na uczelniany e-mail w domenie pw.edu.pl.

Portal udostępnia dane na temat instytutów, naukowców oraz podstawowe informacje o publikacjach tj. DOI, lista autorów, czasopismo, data publikacji.



Rysunek 6 Przykładowe dane z portalu ResearchGate
Źródło: www.researchgate.net , 05.02.2017

Portal umożliwia przeglądanie instytutów według krajów oraz odnalezienie pracowników danego instytutu, oraz finalnie wgląd w publikacje autorów. Na rysunku 6 został przedstawiony przykładowy wgląd w artykuł jednego z pracowników Politechniki Warszawskiej. Zostały wyświetlone dane, wśród których jest m.in.: tytuł artykułu; czasopismo, w którym został opublikowany; data publikacji oraz DOI artykułu. Widoczna jest także lista autorów z ich liczba porządkową dla danej publikacji. Każdy autor jest jednoznacznie określony za pomocą odnośnika do strony swojego profilu. Oprócz tego widoczne są również punkty przypisane do autorów. Jest to *RG Score*[21], współczynnik stworzony przez ResearchGate, który bazuje na ocenie wkładu naukowego na tym portalu przez innych naukowców. Instytuty również mają ten współczynnik, jako pewną wypadkową punktową swoich pracowników. Jest to współczynnik, którego ustalanie nie do końca jest wiadome, więc jego wykorzystanie w pracy byłoby bezcelowe.

Portal ResearchGate oferuje bardzo wiele przydatnych danych o afiliacjach wraz z ich strukturą – uczelnia, wydziały, pracownicy z rozróżnieniem na kraje. Są to dane, które niewątpliwie mogłyby zostać wykorzystane w późniejszych badaniach. W celu pobrania danych o artykułach został zaimplementowany specjalny moduł, który wysyłał żądania HTTP, dostawał odpowiedź i ją parsował wyszukując pożądanych informacji.

3.8 Lista ministerialna

W moich badaniach odwołuję się do czasopism z IF oraz koncentruję na polskich naukowcach jako jednej z badanych podgrup. Z tego powodu najodpowiedniejszą bazą czasopism jest lista ministerialna, która zawiera wszystkie czasopisma, które są klasyfikowane przez *Thomson Reuters* jak i również krajowe i europejskie czasopisma.

CZĘŚĆ A WYKAZU CZASOPISM NAUKOWYCH CZASOPISMA NAUKOWE POSIADAJĄCE WSPÓŁCZYNNIK WPŁYWU IMPACT FACTOR (IF), ZNAJDUJĄCE SIĘ W BAZIE JOURNAL CITATION REPORTS (JCR) WRAZ Z LICZBĄ PUNKTÓW PRZYZNAWANYCH ZA PUBLIKACJĘ W TYCH CZASOPISMACH			
Lp.	TYTUŁ CZASOPISMA	NR ISSN	LICZBA PUNKTÓW ZA PUBLIKACJĘ W CZASOPIŚMIE NAUKOWYM
1	*2*	*3*	*4*
1	4OR-A Quarterly Journal of Operations Research	1619-4500	20
2	AAPG BULLETIN	0149-1423	35
3	AAPS Journal	1550-7416	40
4	AAPS PHARMSCITECH	1530-9932	25

Rysunek 7 Wycinek listy ministerialnej, cześć A
Źródło:http://www.nauka.gov.pl/g2/oryginal/2017_01/964b9d4fd07c847ec0150745fae26feb.pdf, dostęp 30.01.2017

Lista jest wydawana w formacie PDF w formie tabeli. Jej fragment jest widoczny na rysunku 7. Przedstawione są niezbędne dane: tytuł czasopisma, ISSN oraz punktacja. Na liście ministerialnej nie ma współczynnika wpływu, który należało znaleźć do pobrania w innym źródle.

W tym wypadku do pobrania danych do bazy danych wystarczyło zaimplementować prosty moduł odczytu z tekstu.

3.9 Pobieranie danych – próba I

W porównaniu serwisów Google Scholar oraz Web of Science, Google Scholar wypadł o wiele lepiej. Miał znacznie więcej wyników oraz cytowań prac. W szczególności, jeżeli chodzi o polskich autorów. Baza Web of Science jest skierowana przede wszystkim na Stany Zjednoczone. Z tych powodów wybrany został serwis Google Scholar i został zaimplementowany kolejny moduł zbierający dane. Zarówno moduł do Google Scholar jak i do portalu ResearchGate implementacyjnie były do siebie bardzo zbliżone. Korzystały z biblioteki JSoup w celu pobierania i parsowania stron HTML. Największą wadą tego rozwiązania było ręczne szukanie potrzebnych danych na stronie i sprawdzanie, w jakich elementach się znajdują, żeby można je było sparsować. Fragment kodu źródłowego został przedstawiony w Dodatku Dodatek A – Fragmenty kodu źródłowego na listingu 6.

Mając już przygotowane narzędzia do pobierania informacji i zapisywania ich w bazie danych można było ten proces rozpocząć. Rozpoczęto od sparsowania listy ministerialnej,

aby zapisywane w bazie artykuły mogły się odwoływać do czasopism. Następnym krokiem było pobranie danych z serwisu ResearchGate, gdzie napotkane zostały pierwsze poważne problemy. Serwis założył blokadę na adres IP ze względu na zbyt dużą liczbę zapytań, a było ich ok. 5000. Moduł był zaimplementowany w sposób taki, żeby nie wzbudzał zbyt dużych podejrzeń (m.in. usypianie wątku wysyłającego zapytania HTTP na losową długość czasu między 0,5 a 1,5 sekundy między kolejnymi zapytaniami).

W celu obejścia tej blokady skorzystałem z Internetu mobilnego przez smartfon wyposażony w system Android. Mechanizm pozwalający obejść blokowanie adresu IP polegał na tym, iż mając dostęp do Internetu przez smartfon podłączony do komputera przez kabel USB i po nałożeniu blokady przez serwis ResearchGate, moduł uruchamiał specjalne skrypty, który komunikowały się z smartfonem. Skrypty te miały za zadanie wprowadzić telefon w tryb samolotowy, a następnie po paru sekundach powrócić w tryb normalny i sprawdzenie aktualnego adresu IP. Jeśli adres się zmienił to moduł zbierający wznawiał swoją pracę. W przeciwnym wypadku operacja przejścia w tryb samolotowy była powtarzana. Zabieg ten pozwalał na dostanie przez smartfon kolejnego adresu IP z puli adresów.

Zostały zebrane informacje o instytutach, autorach oraz ich publikacjach z Polski, Szwajcarii oraz Austrii. Łącznie było to 32732 artykułów oraz 170598 autorów i 10628 instytucji naukowych z wymienionych krajów.

Niestety po paru dniach oprócz adresu IP również moje konto zostało zawieszone przez administratorów portalu. Dalsze korzystanie było niemożliwe, jednakże posiadałem już całkiem dużo informacji.

Mając dane o artykułach mogłem przystąpić do pobierania informacji o cytowaniach. Google Scholar okazał się być mniej liberalnym od ResearchGate, jeśli chodzi o wysyłanie zapytań i blokował adresy IP już po 100 zapytaniach. Używając tego samego rozwiązania, co przy ResearchGate z Internetem mobilnym, doprowadziło to do skończenia się nowych adresów w puli adresów przyznawanych. Po około 12 godzinach zbierania danych adresy się zapętlały i należało przerwać pracę. Jednak Google nakłada blokady czasowe i po mniej więcej 48 godzinach pracę można było wznowić na nowo.

W ten sposób udało się pobrać około 100 tysięcy informacji o cytowaniach. Jednak powstał problem, który był w tych warunkach trudny do rozwiązania. Nałożenie permanentnej blokady przez portal ResearchGate uniemożliwiało znalezienie informacji o artykułach pobranych z Google Scholar. Należało znaleźć nowe źródło danych, które nie będzie blokowało adresów IP.

3.10 Scopus

Baza Scopus[22] jest konkurencyjna wobec Web of Science. W odróżnieniu od bazy WoS, która należy do wydawnictwa *Thomson Reuters* baza Scopus należy do wydawnictwa *Elsevier*. Dostęp do niej również jest ograniczony do sieci instytutu. Dodatkowo jak *Elsevier* zapewnia baza Scopus jest o wiele bardziej otwarta na Europę niż Web of Science, co może nieść ze sobą korzyści związane z większą liczbą danych o polskich autorach.

Jednak Scopus posiadał jeszcze jedną znaczącą przewagę nad bazami Google Scholar oraz Web of Science, mianowicie posiada RESTful API do korzystania ze swojej bazy[23]. API w pełni umożliwia pobieranie informacji o autorach, afiliacjach oraz artykułach i ich cytowaniach. Wadą była niekompletna dokumentacja, która nie ujawniała wszystkich możliwości ani sposobów wykorzystania tego API. Niewielką niedogodnością API było korzystanie ze specjalnych kluczy wygenerowanych na portalu *Elsevier*, które były ograniczone na liczbę zapytań w zależności od rodzaju zapytania (autor, artykuł, afiliacja) od 1 tysiąca do 10 tysięcy zapytań na jeden klucz, po czym należało wygenerować nowy klucz.

W ten sposób zostało znalezione źródło danych, które spełnia wszystkie postawione wymagania. Oprócz tego umożliwia jednoznaczną identyfikację autora. Każdy autor posiada unikalny identyfikator, który jest mu przypisany przez serwis Scopus.

Należało jeszcze porównać dotychczasowe dane z informacjami, które oferuje *Elsevier*. Jak opisuje Harzing [5] bazy Google Scholar, Scopus oraz Web of Science zawierają porównywalne dane, jednak używanie ich naprzemiennie nie jest wskazane, ponieważ każda baza ma różniące się informacje przede wszystkim o cytowaniach. Z tego powodu zdecydowałem się na zebranie danych od początku, co pozwoliło zapewnić najlepsze informacje oraz rzetelne wyniki.

Elsevier oprócz swojej bazy, proponuje również korzystanie z innych wskaźników ocen czasopism oprócz Impact Factor. Są to trzy współczynniki[24]:

- SJR – SCImago Journal Rank,
- SNIP – Source Normalized Impact per Paper,
- IPP – Impact per Publication (zastąpiony przez CiteScore[25]).

Współczynniki te wyliczane są na podstawie bazy Scopus, w związku z tym mają przewagę nad IF. Wynika ona z tego, że baza Scopus gromadzi ponad 20 tysięcy czasopism, podczas gdy IF jest wyliczany z basy Web of Science dla ponad 10 tysięcy czasopism. Współczynniki te podobnie jak IF są również wyliczane na podstawie cytowań, jednak z pewną różnicą ważności cytowań. Wskaźnik SJR jest wyliczany bazując na algorytmie podobnym do Google Page Rank. Stosuje on algorytm, w którym oprócz cytowania uwzględnia się, w jakim czasopiśmie jest artykuł cytowany. Bardziej wartościowe cytowania

są z prestiżowych czasopism. Ograniczono również uwzględnianie autocytowań przez czasopismo do 1/3 wszystkich cytowań. Wskaźnik SNIP również wartościuje cytowania, jednak w tym wypadku liczą się kontekstowe cytowania dla danej dziedziny. Wartościowane są cytowania występujące w danej dziedzinie nauki. Współczynnik IPP jest podobny w swoim algorytmie do IF z pewnymi różnicami. Bada on okres 3 ostatnich lat w przeciwieństwie do 2 ostatnich w IF oraz mierzy on jedynie cytowania artykułów, gdzie przy wyliczaniu IF są uwzględniane m.in. także recenzje i materiały konferencyjne.

Korzystanie z bazy *Elsevier* pozwoli na analizę nie tylko pod kątem IF, ale również innych współczynników. Wokół IF narasta krytyka, której przyczyną jest metoda wyliczania wskaźnika. Wylicza się go na podstawie całkowitej liczby cytowań danego czasopisma. Może to doprowadzać do sytuacji, w których kilka artykułów ma bardzo wiele cytowań, a pozostałe mają dużo mniej, aczkolwiek ogólny IF będzie nadal wysoki. Na podstawie artykułu [6] zdecydowałem się również prowadzić analizy dla współczynnika SJR, który obecnie jest najczęściej porównywany z IF. Sprawdzone zostało czy oba te współczynniki można używać z podobnym efektem końcowym.

Aby móc zacząć korzystać z danych *Elsevier* należało ponownie napisać nowe moduły zbierające dane wykorzystujące REST API oraz przystąpić do ich zbierania. Fragmenty modułów zostały przedstawione w Dodatku **Błąd! Nie można odnaleźć źródła odwołania.** na listingach 3, 4, 5.

Na listingu 1 pokazano fragment wyniku użycia tego API w celu uzyskania informacji o artykułach ze Stanów Zjednoczonych opublikowanych po 31.12.2011. Odpowiedź zwracana jest w formacie XML, wyniki zapytania znajdują się w znaczniku *search-results*. Na początku znajdują się metadane takie jak liczba znalezionych artykułów, indeks pierwszego wyniku, liczba zwróconych wyników na zadane zapytanie oraz zapytanie. Początkowy indeks oraz liczba wyników jest podana w zapytaniu przez parametry *start* oraz *count*, w tym przypadku równe odpowiednio 0 i 200. Poszczególne artykuły spełniające żądanie są podane w znacznikach *entry*, które należy iterować w celu wyłuskania wszystkich wyników. Z poszczególnego rezultatu można uzyskać wiele informacji, które są umieszczone w odpowiednich znacznikach np.:

- *prism:doi* – DOI artykułu,
- *dc*:*title* – Tytuł artykułu,
- *prism*:*issn* – ISSN czasopisma publikacji,
- *authors* – Lista autorów.

Listing 1 Fragment rezultatu użycia Elsevier API do uzyskania danych o artykułach

```
<search-results>
<opensearch:totalResults>7104</opensearch:totalResults>
<opensearch:startIndex>0</opensearch:startIndex>
<opensearch:itemsPerPage>200</opensearch:itemsPerPage>
<opensearch:Query role="request" searchTerms="affil(united states) and
pub-date AFT 20111231  and content-type(JL)" startPage="0"/>
<link ref="self"
href="http://api.elsevier.com/content/search/scidir?start=0&count=200&qu
ery=affil%28united+states%29+and+pub-date+AFT+20111231++and+content-
type%28JL%29&APIKey=XXXXXb&subj=computerscience"
type="application/xml"/>
<entry>
<link ref="self"
href="http://api.elsevier.com/content/article/pii/S1084804515002623"/>
<dc:identifier>DOI:10.1016/j.jnca.2015.11.006</dc:identifier>
<eid>1-s2.0-S1084804515002623</eid>
<prism:url>
</prism:url>
<dc:title>Smart grid communications: Modeling and validation</dc:title>
<dc:creator>Ming, Yu</dc:creator>
<prism:publicationName>Journal of Network and Computer
Applications</prism:publicationName>
<prism:issn>10848045</prism:issn>
<prism:volume/>
<prism:issueIdentifier/>
<prism:coverDate>2016-01-01</prism:coverDate>
<prism:coverDisplayDate>January 2016</prism:coverDisplayDate>
<prism:startingPage>247</prism:startingPage>
<prism:endingPage>249</prism:endingPage>
<prism:doi>10.1016/j.jnca.2015.11.006</prism:doi>
<authors>
<author>
<given-name>Ming</given-name>
<surname>Yu</surname>
</author>
```

3.11 Pobieranie danych – próba II

Korzystając z API *Elsevier* możliwym było pobranie danych o wybranych artykułach parametryzując dziedzinę, kraj pochodzenia autorów, datę publikacji. W tym wypadku dziedzina została określona, jako *Computer Science* i zostały pobrane artykuły po kolei z każdego kraju z datą publikacji od 2013 roku włącznie. W sumie zostały pobrane informacje o 99850 artykułach. Najwięcej było ich z Chin – 13896. Dla artykułów z Polski pobrane zostały wszystkie bez ograniczeń czasowych. Uzyskano ich 5478, z czego najstarszy z 1961 roku.

Dla artykułów zostały pobrane informacje o autorach. Łącznie z autorami z cytowanych artykułów udało się zebrać dane o 164670 naukowcach wraz z statystykami (H-indeks, liczba cytowań). Zostali oni przypisani do 154240 różnych afiliacji.

Na koniec zaktualizowano i rozszerzono bazę czasopism o brakujące parametry IF, a także dodano wskaźniki SJR, SNIP oraz IPP.

3.12 Podsumowanie pobierania

W efekcie końcowym przedsięwziętych działań otrzymana została wypełniona baza danych z następującymi informacjami pobranymi z bazy Scopus:

- Instytuty – nazwa, państwo, liczba pracowników, instytut nadrzędny (jeśli występuje).
- Autorzy – imię, nazwisko, h-indeks, liczba cytowań, liczba cytowanych przez autora artykułów, liczba współautorów, liczba artykułów, przynależność do instytutów.
- Artykuły – tytuł, DOI, kraj pochodzenia, data publikacji, dziedzina, czasopismo, autorzy, cytowania.
- Czasopisma – nazwa, ISSN, punkty przyznawane przez MNiSW, współczynniki IF, SJR, IPP, SNIP.

Struktura bazy danych pozostała niezmieniona, tak jak przedstawiono na rysunku Rysunek 3.

4 Wyniki

Po długotrwałym procesie zbierania potrzebnych informacji należało się dokładnie zapoznać z otrzymanymi danymi i ich ilością.

Tabela 1 Statystyki bazy danych

Czasopisma naukowe	łącznie	25271
	spoza listy ministerialnej	7485
	posiadające SJR, SNIP oraz IPP	12780
	posiadające IF	1675
Artykuły	łącznie	99843
	z różnych czasopism	1062
	z Polski	5478
	wydane od 01.01.2015	25468
	wydane od 01.01.2014	47497
Autorzy	łącznie	201202
	posiadający h-indeks, liczbę cytowań, artykułów, współautorów	164670
	posiadający afiliację w Polsce	4458
Afiliacje	łącznie	154240
	z Polski	2622
Cytowania	łącznie	117972
	dla poszczególnych artykułów	41671
	dla poszczególnych artykułów z Polski	3428

W tabeli 1 przedstawiono statystyki z bazy danych, ukazujące liczby pobranych danych. Liczba artykułów jest wystarczająca do przeprowadzenia rzetelnej analizy. Łączna liczba cytowań odnosi się do artykułów zindeksowanych w bazie Scopus, które posiadają DOI. Faktyczna liczba cytowań jest większa. Martwiącym był jedynie fakt, że czasopism posiadających Impact Factor jest tylko 1675, ale biorąc również pod uwagę fakt, że wszystkie artykuły zostały opublikowane w 1062 czasopismach należało dokładniej sprawdzić dla ilu artykułów czasopisma mają określony IF oraz SJR.

Sprawdzono ile faktycznie jest czasopism bez współczynników SJR oraz IF, w których były publikowane zebrane artykuły oraz wariacje tych współczynników.

Tabela 2 Dane czasopism, w których były publikowane artykuły

Czasopisma przypisane do artykułów	1062
w tym bez SJR	165
w tym bez IF	221
w tym bez IF i SJR	157
w tym bez IF lub SJR	229

W tabeli 2 można zobaczyć, iż IF pobranych było dla 1675 czasopism, co w porównaniu z pobranymi współczynnikami SJR, których było prawie 8 razy więcej nie robi różnicy dla posiadanych przeze mnie danych o czasopismach. Proporcje czasopism bez SJR lub IF są do siebie bardzo zbliżone. Z tabeli 2 wynika również, że czasopisma nieposiadające SJR nie posiadają również w większości współczynnika wpływu. Z tego wynika, że ok. 20% danych może być nie przydatna. Żeby to potwierdzić zostały przeanalizowane również artykuły.

Tabela 3 Statystyki artykułów i czasopism

Artykuły	99843
Artykuły bez SJR czasopisma	11400
Artykuły bez IF czasopisma	12051
Artykuły bez IF i SJR czasopisma	10548
Artykuły cytujące	41671
Artykuły cytujące bez SJR czasopisma	2204
Artykuły cytujące bez IF czasopisma	2302
Artykuły cytujące bez IF i SJR czasopisma	1975

W tabeli 3 pojawia się określenie „artykuły cytujące" są to artykuły, dla których zostały zebrane informacje o ich bibliografii. Na podstawie danych zawartych w tabeli 3, można stwierdzić, że ok. 10% artykułów nie posiada czasopisma z współczynnikiem SJR lub IF, jednakże tylko ok. 5% artykułów, które będą niezbędne w analizie nie posiada tych współczynników. W optymistycznym wariancie, cytowane artykuły bez współczynnika są cytowane również przez artykuły bez współczynnika i wtedy faktycznie tylko ok. 5% danych zostaje niewykorzystanych. W pesymistycznym wariancie nie zachodzi relacja opisana w poprzednim przypadku i w takiej sytuacji niewykorzystane będzie ok. 10%. Jednak w obu przypadkach nie powinna być to sytuacja problematyczna, z powodu dużej liczby danych do analizy.

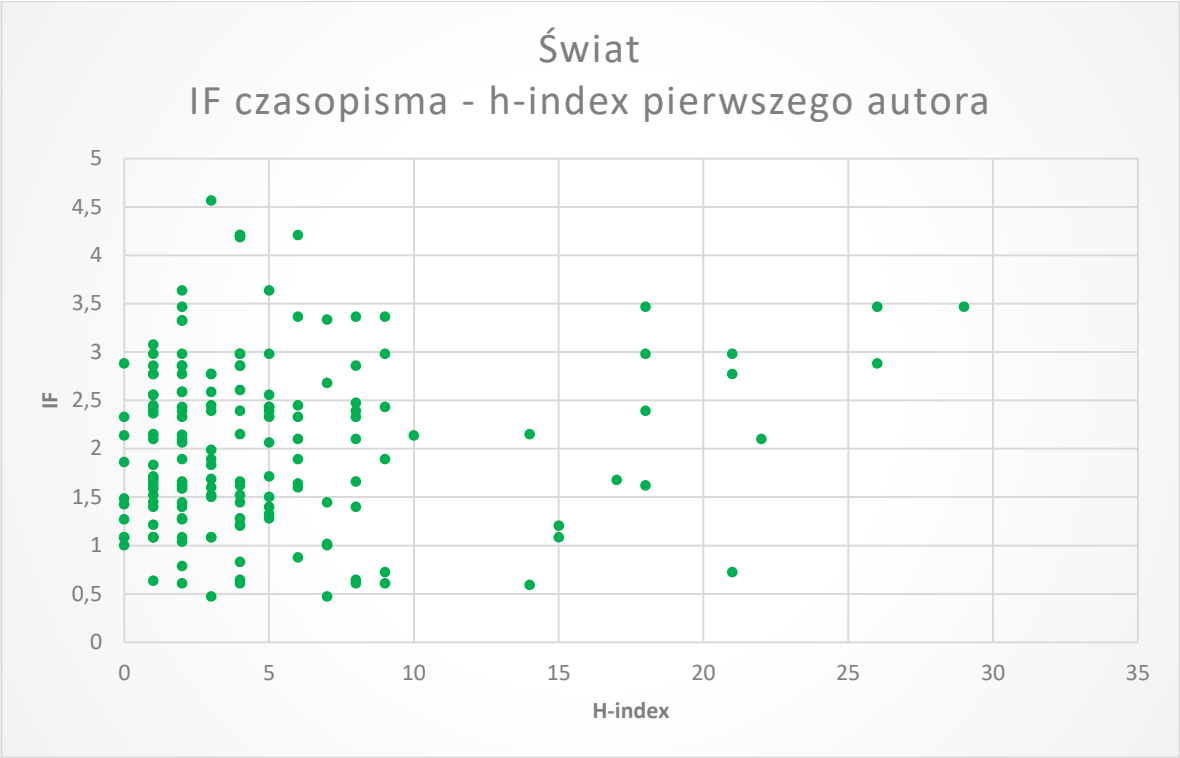
Po zapoznaniu z danymi właściwa analiza mogła być rozpoczęta, a hipotezy poddane weryfikacji.

4.1 Wpływ statystyk autora na publikację

Pierwszym aspektem do zbadania było sprawdzenie zależności h-indeksów autorów względem IF czasopisma publikacji. Z powodu, że zarówno h-indeks autorów jak i współczynniki czasopism były pobierane dla bieżących danych w bazie Scopus, która jest często aktualizowana (raz na kwartał) badaniu zostały poddane grupy artykułów opublikowanych od 2015 roku włącznie. Można przyjąć, że pobrane statystki autorów oraz czasopism są prawidłowe dla tych lat. Artykuły zostały również pogrupowane na kolejne dwie grupy ze względu na kraj pochodzenia autorów Polska oraz cały świat. Być może na mniejszej próbie danych mogą być widoczne inne zależności niż na całym zbiorze. Badane były tylko artykuły, z co najmniej 3 autorami, aby zapewnić jednolitość i pewne podobieństwo informacji wejściowych. Dla kraju pochodzenia artykułów z Polski uzyskano dane o 224 publikacjach i wszystkie zostały zaprezentowane na wykresach. Natomiast dla całego świata otrzymano 8719 artykułów, ponieważ diagramy przedstawiające tyle danych byłyby nieczytelne to zostały losowo wybrane dane o 166 artykułach z całego świata.

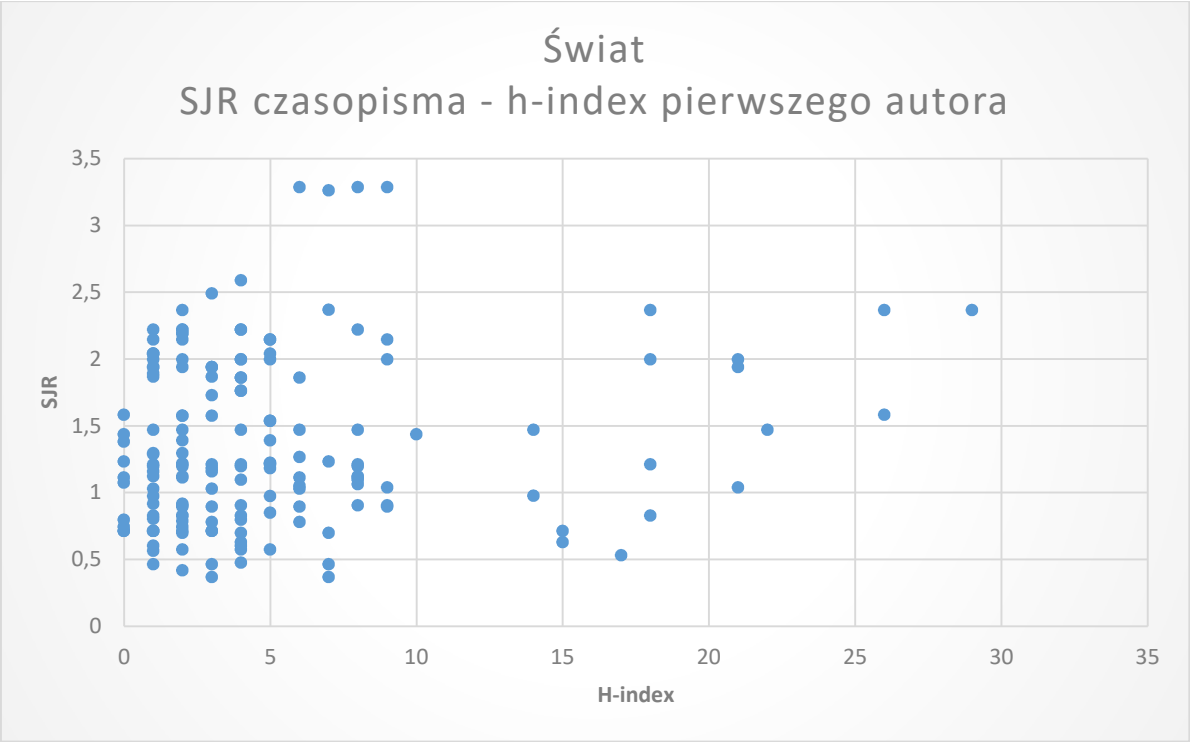
Zostały wydzielone następujące warianty dotyczące zależności h-indeksu autorów:

- H-indeks uśredniony wszystkich autorów publikacji,
- H-indeks pierwszego autora,
- H-indeks ostatniego autora,
- H-indeks maksymalny wśród autorów,
- H-indeks minimalny wśród autorów.



Rysunek 8 Wykres zależności h-indeksu pierwszego autora od IF czasopisma publikacji przez autorów z całego świata

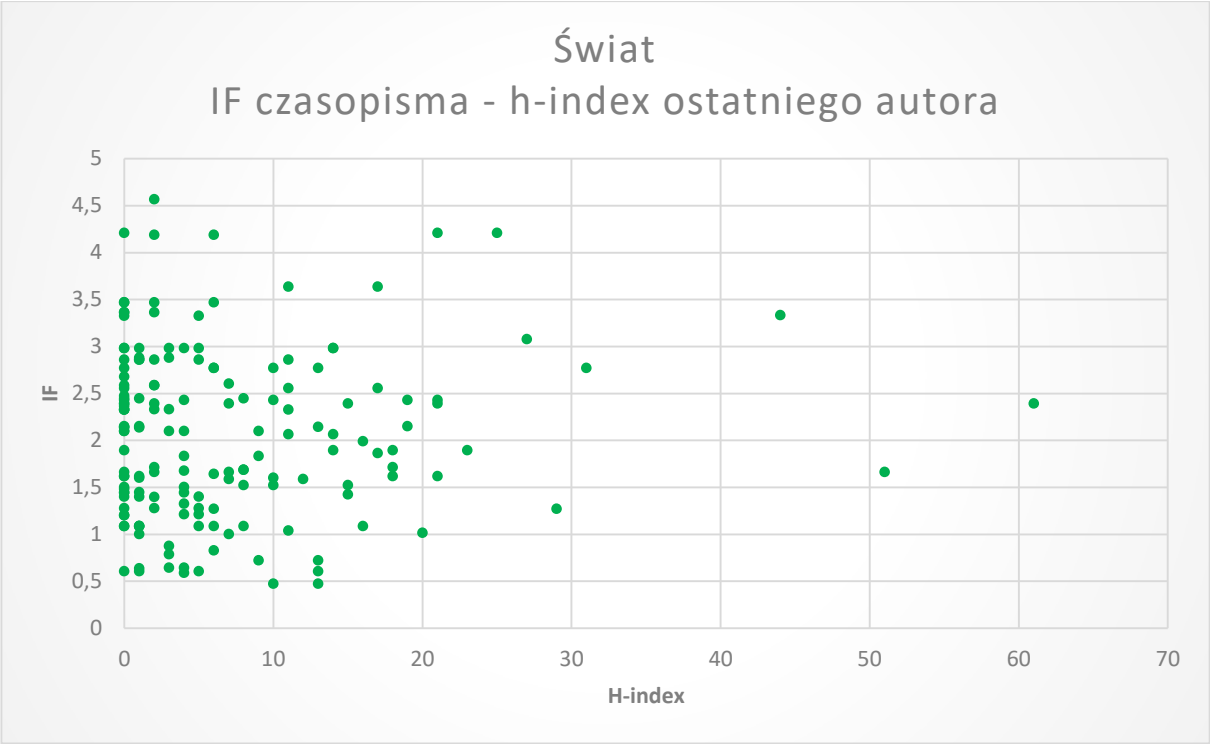
Na rysunku 8 pokazano zależność indeksu Hirscha pierwszego autora artykułu z współczynnikiem wpływu czasopisma, które opublikowało artykuł. Współczynnik korelacji Pearsona wynosi 0,0395569. Jest widoczna nieznaczna zależność dla autorów z h-indeksem powyżej 20, którzy nie publikują w czasopismach z IF poniżej 2. Aczkolwiek autorzy z niższym h-indeksem, których jest znacznie więcej publikują zarówno w czasopismach z wysokim jak i niskim Impact Factor. Z wymienionych powodów h-indeks pierwszego autora nie ma wpływu na to, w jakim czasopiśmie ukazał się artykuł.



Rysunek 9 Wykres zależności h-indeksu pierwszego autora od SJR czasopisma publikacji przez autorów z całego świata

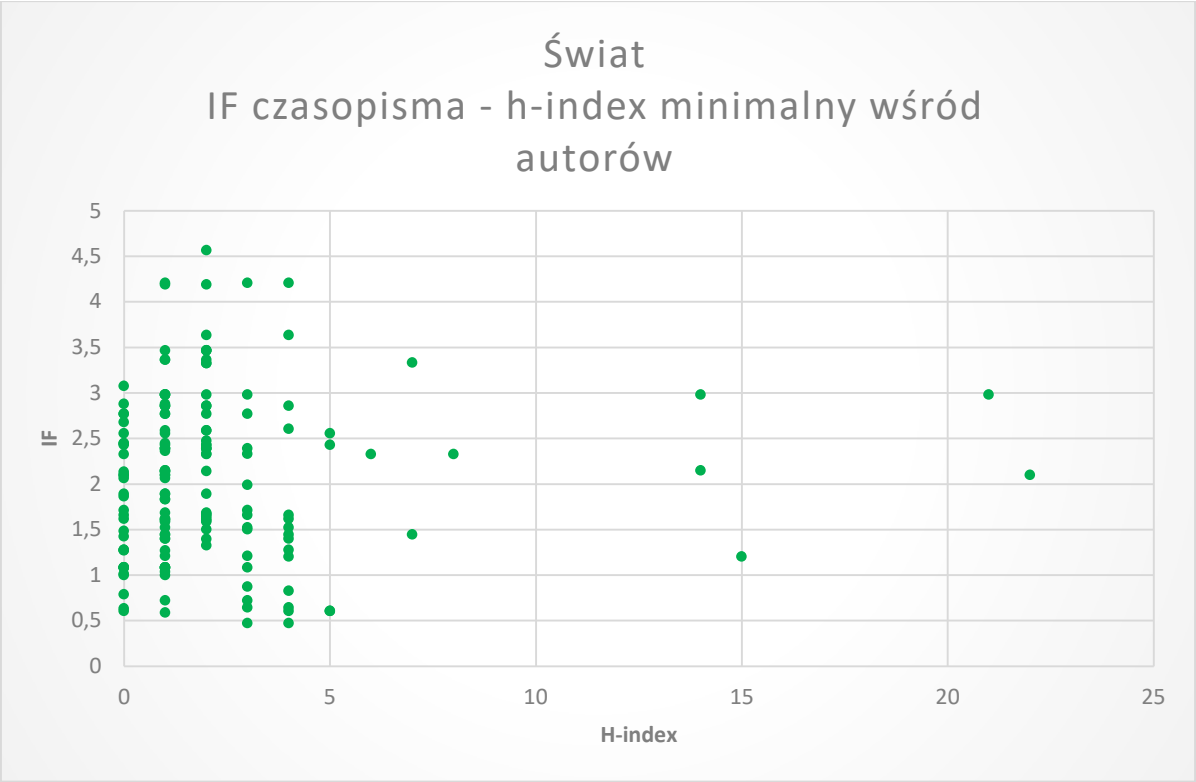
Na rysunku 9 pokazano tą samą zależność co na rysunku 8 z tą różnicą, że użyto współczynnika SJR zamiast IF. Współczynnik korelacji Pearsona wynosi 0,0557655, dlatego można również stwierdzić tak jak w poprzednim przypadku bardzo słabą zależność oraz brak wpływu h-indeksu pierwszego autora na publikacje w określonym czasopiśmie.

Porównując oba wykresy można dojść do wniosku, że wskaźniki te nie są identyczne, ale są podobne. Współczynnik korelacji Pearsona dla IF oraz SJR dla otrzymanych danych wynosi 0,6822405, także możemy tu mówić o dość silnej zależności liniowej między współczynnikami IF oraz SJR. Jednak nie jest ona tyle silna, aby móc używać naprzemiennie obu współczynników. Aczkolwiek dla dużej liczby przypadków hierarchia czasopism określona za pomocą IF będzie bardzo podobna jak w przypadku współczynnika SJR.



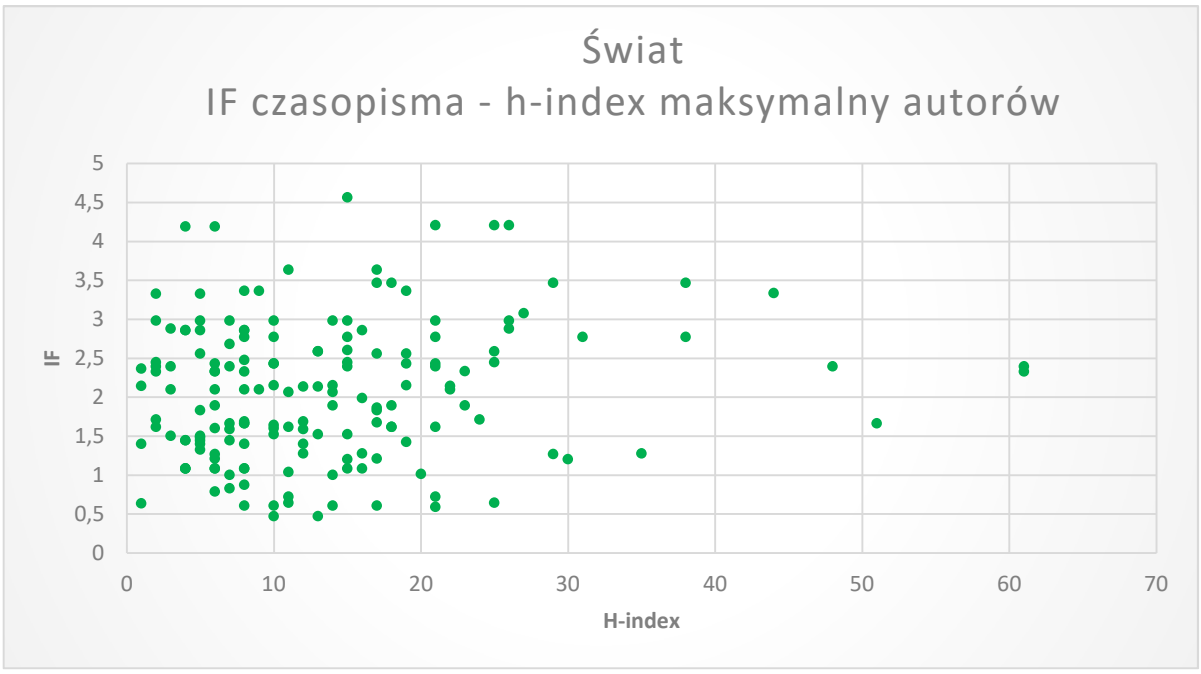
Rysunek 10 Wykres zależności h-indeksu ostatniego autora od IF czasopisma publikacji przez autorów z całego świata

Na wykresie przedstawionym na rysunku 10 przedstawiono zależność indeksu Hirscha ostatniego autora tzw. koordynatora od IF czasopisma publikacji. Współczynnik korelacji Pearsona wynosi 0,0180506. Na tej podstawie można stwierdzić, że artykuły ukazują się zarówno w czasopismach z wysokim współczynnikiem Impact Factor jak i z niskim IF bez względu na h-indeks ostatniego autora. Podobnie jak w poprzednim przypadku nie ma związku pomiędzy IF czasopisma a h-indeksem koordynatora publikacji. Przyczyną tego może być to, że większy wpływ odgrywa autor publikacji z maksymalnym, bądź minimalnym h-indeksem spośród autorów.



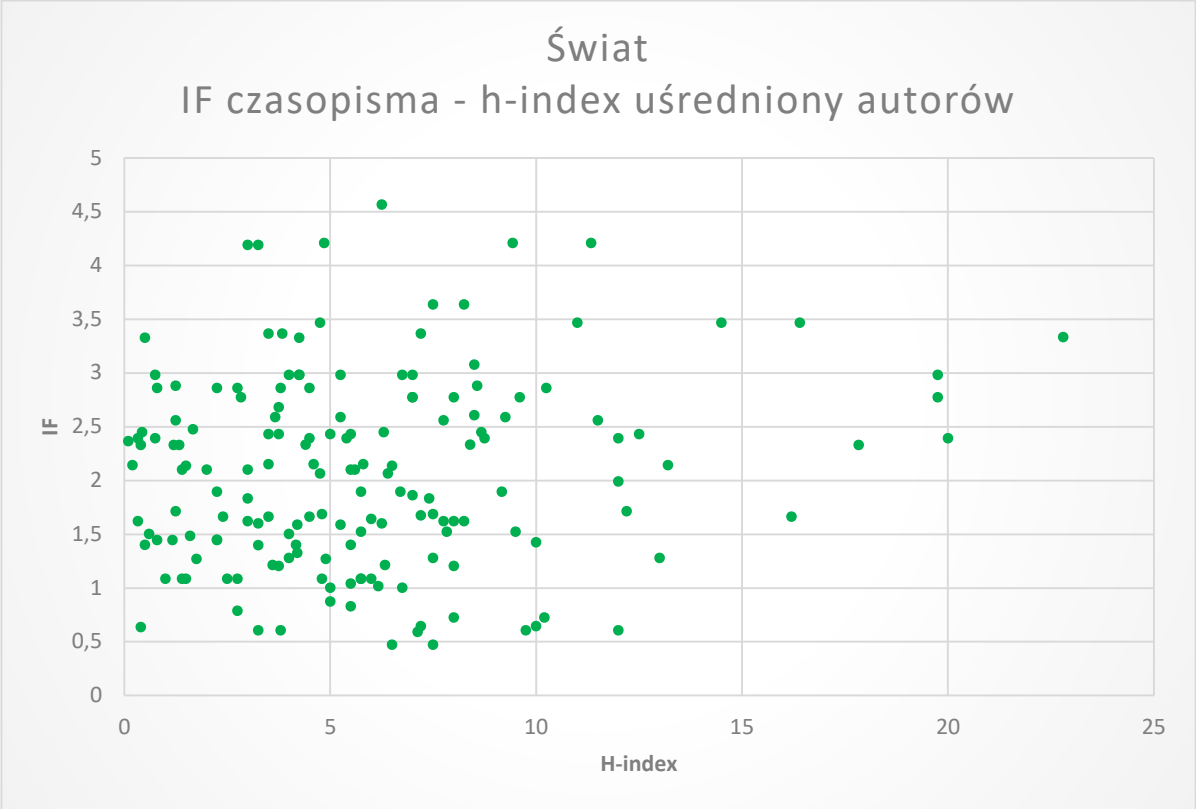
Rysunek 11 Wykres zależności minimalnego h-indeksu autorów od IF czasopisma publikacji przez autorów z całego świata

Na rysunku 11 ukazano zależność między minimalnym indeksem Hirscha wśród autorów publikacji, a IF czasopisma publikacji również na podstawie współczynnika korelacji wynoszącego 0,031439 można stwierdzić brak zależności pomiędzy tymi parametrami. Autor z najniższym h-indeksem nie ma wpływu na ukazanie się artykułu. Być może o sukcesie publikacji w większym stopniu decyduje autor z maksymalnym h-indeksem.



Rysunek 12 Wykres zależności maksymalnego h-indeksu autorów od IF czasopisma publikacji przez autorów z całego świata

Na rysunku 12 została pokazana relacja między maksymalnym indeksem Hirscha wśród autorów publikacji, a IF czasopisma publikacji. Współczynnik korelacji wynosi 0,0698137 z tego powodu zauważalne jest, że publikacje ukazują się w czasopismach o różnym IF bez znaczenia maksymalnego h-indeksu autorów. Podobnie jak w poprzednich przypadkach h-indeks autora nie ma wpływu na publikację artykułu. Wynika z tego, że h-indeks pojedynczego autora nie odgrywa żadnej roli w procesie publikacji. Następnie należało sprawdzić, czy h-indeks wszystkich autorów ma znaczenie przy publikacji artykułu.



Rysunek 13 Wykres zależności uśrednionego h-indeksu autorów od IF czasopisma publikacji przez autorów z całego świata

Relację uśrednionego h-indeksu wszystkich autorów publikacji do współczynnika wpływu pokazano na rysunku 13. Również i tym razem współczynnik korelacji wynoszący 0,0643021 wskazuje na brak zależności. Podobnie jak h-indeks poszczególnych wybranych autorów, tak samo średni h-indeks autorów publikacji nie ma wpływu na upublikowanie artykułu. Największy współczynnik korelacji odnotowano dla autora z najwyższym wskaźnikiem Hirscha 0,0698137, aczkolwiek jest on zbyt niski by móc mówić o wpływie na publikację.

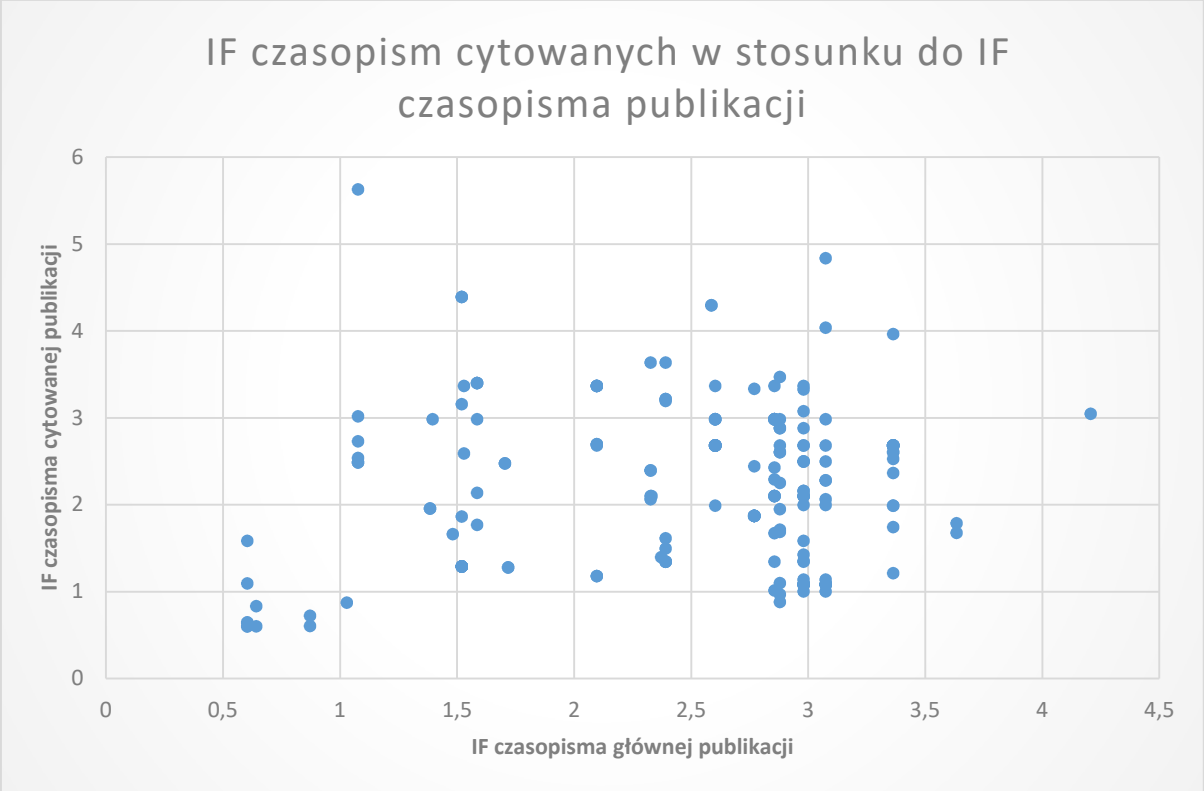
Pozostałe wykresy, w tym dla artykułów z Polski oraz wykresy dla współczynnika SJR zostały umieszczone w Dodatku B.1 Statystyki autorów artykułu, ze względu na ich dużą objętość i podobieństwo do siebie.

Podsumowując, nie znaleziono zależności pomiędzy indeksem Hirscha autorów publikacji, a współczynnikami oceny czasopisma publikacji. H-indeks nie ma wpływu na publikację. Aczkolwiek na podstawie otrzymanych wykresów można potwierdzić, że używanie współczynników SJR oraz IF niesie podobny skutek i można ich obu używać w celu odniesienia się do renomy czasopisma. Należy jednak pamiętać, że te współczynniki mogą inaczej klasyfikować dane czasopisma.

4.2 Wpływ cytowań na publikację

Kolejnym wariantem jest sprawdzenie zależności pomiędzy artykułem a jego artykułami cytowanymi. Sprawdzone zostanie czy współczynnik czasopisma cytowanego ma wpływ na to, w jakim czasopiśmie ukazuje się publikacja. Brane pod uwagę zostały artykuły opublikowane od 2015 roku. Z tej grupy zostało losowo wybranych 200 artykułów, w celu zapewnienia przejrzystości wykresów. Analizie zostały poddane następujące warianty:

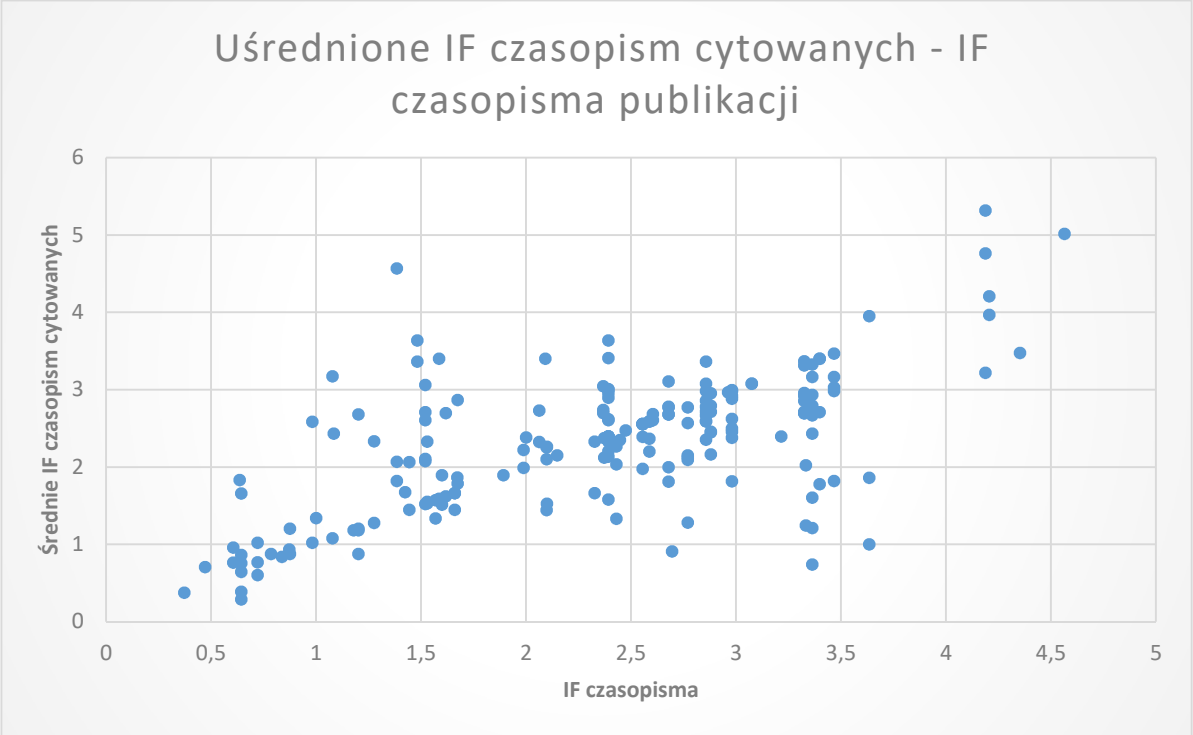
- współczynniki czasopism dla publikacji cytowanych,
- średnia współczynników czasopism dla publikacji cytowanych,
- minimalny współczynnik czasopisma z publikacji cytowanych,
- maksymalny współczynnik czasopisma z publikacji cytowanych.



Rysunek 14 Wykres zależności pomiędzy IF czasopisma publikacji do IF czasopism cytowanych publikacji

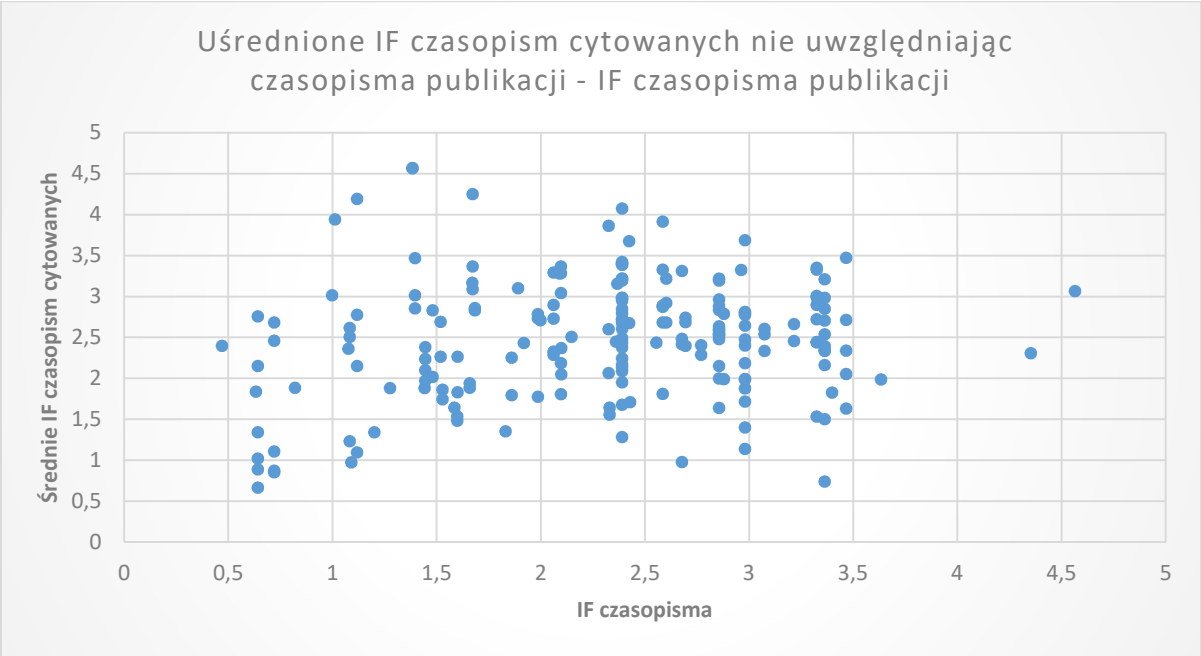
Na rysunku 14 przedstawiono zależność IF czasopisma publikacji wobec IF czasopism cytowanych publikacji. Z tego zbioru zostały wyłączone cytowania artykułów z tego samego

czasopisma, co czasopisma publikacji, ponieważ naturalnym jest, że najczęściej publikacja odwołuje się do artykułów z czasopisma, w którym próbuje być opublikowana. Takie wartości nic nie wniosłyby do analizy, co zostało przedstawione na rysunku 15. Na wykresie widocznych jest mniej niż 200 punktów. Jest to spowodowane tym, że często dochodzi do sytuacji, kiedy cytowane jest kilka artykułów z jednego czasopisma wtedy te punkty na wykresie się pokrywają. Współczynnik korelacji Pearsona dla powyższych argumentów wynosi 0,0639646. Również i tym razem nie występują tutaj żadne zależności, artykuł zawiera bibliografię z czasopism zarówno z niższym jak i wyższym współczynnikiem IF.



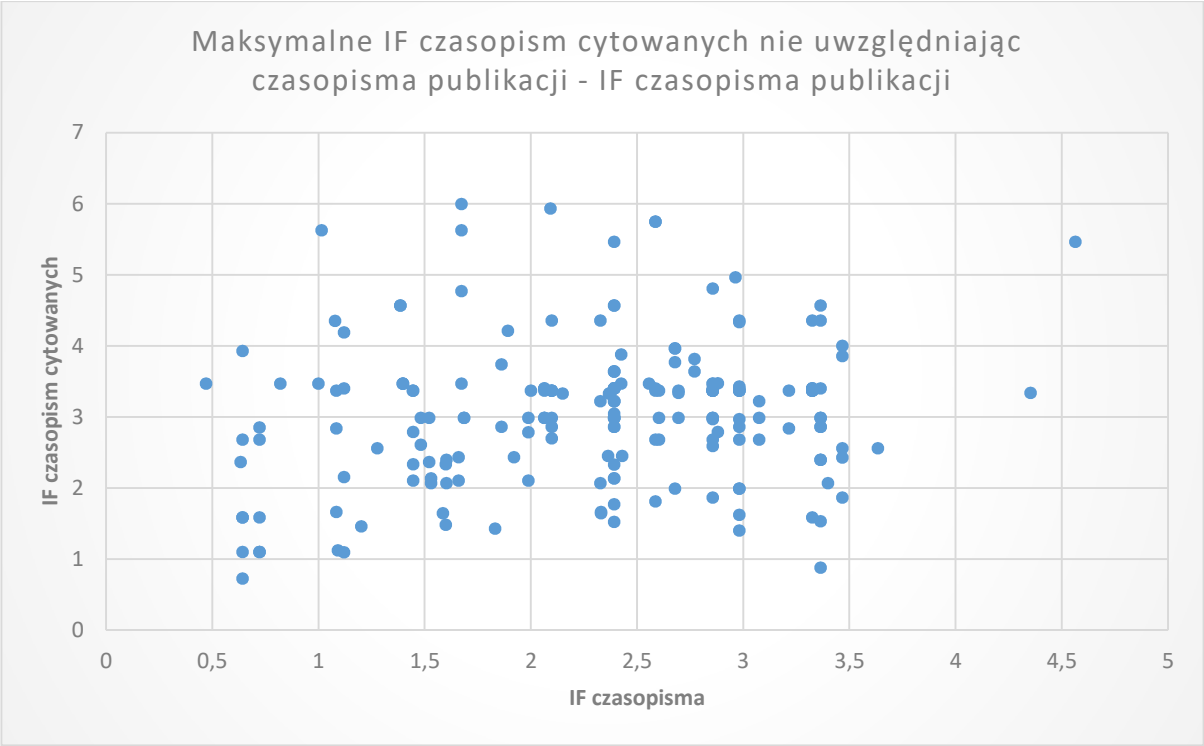
Rysunek 15 Wykres zależności uśrednionego IF czasopism cytowanej publikacji do czasopisma publikacji

Na rysunku 15 przedstawiono zależność pomiędzy średnim IF czasopism cytowanych przez artykuł a jego IF czasopisma publikacji. Współczynnik korelacji Pearsona wynosi 0,39949, co wskazuje na pewną liniową zależność, która również jest widoczna na tym wykresie. Jest ona spowodowana uwzględnieniem cytowań z tych samych czasopism, które jak zostało opisane powyżej nie wnoszą nic do analizy. Dlatego został przygotowany następny wykres nieuwzględniający cytowań z tego samego czasopisma.

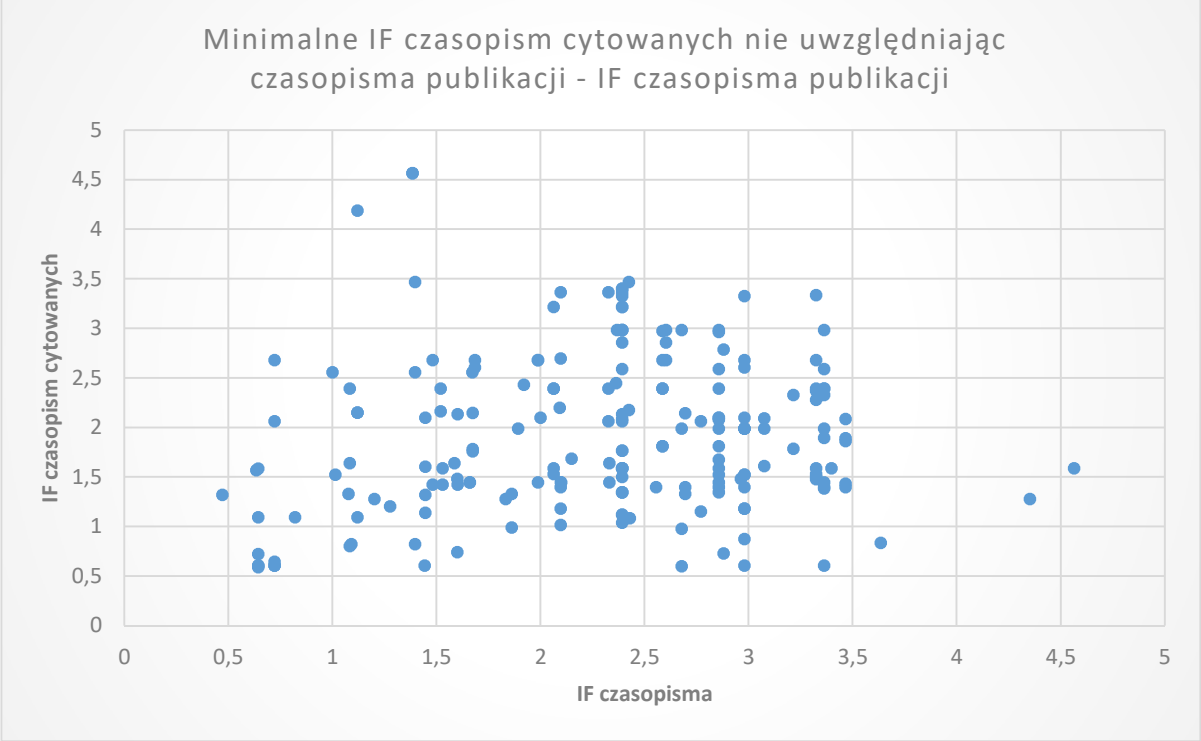


Rysunek 16 Wykres zależności uśrednionego IF czasopism cytowanej publikacji wyłączając cytowania z czasopisma publikacji do czasopisma publikacji

Na rysunku 16 przedstawiono zależność pomiędzy średnim IF czasopism publikacji cytowanych nieuwzględniających publikacji z tego samego czasopisma, co czasopismo publikacji, które były uwzględnione na rysunku 15. Widoczna jest zasadnicza różnica pomiędzy tymi wykresami. Współczynnik korelacji wynosi 0,2356515, co świadczy o dużo mniejszej zależności. Jednak można stwierdzić, że występuje pewna zależność. Jednak nie jest ona na tyle silna, żeby móc mówić o faktycznym wpływie doboru cytowań na publikację artykułu. Jest to również widoczne na powyższym wykresie, na którym nie ma zdecydowanej tendencji liniowej zależności.



Rysunek 17 Wykres zależności maksymalnego IF czasopism cytowanej publikacji wyłączając cytowania z czasopisma publikacji do czasopisma publikacji



Rysunek 18 Wykres zależności minimalnego IF czasopism cytowanej publikacji wyłączając cytowania z czasopisma publikacji do czasopisma publikacji

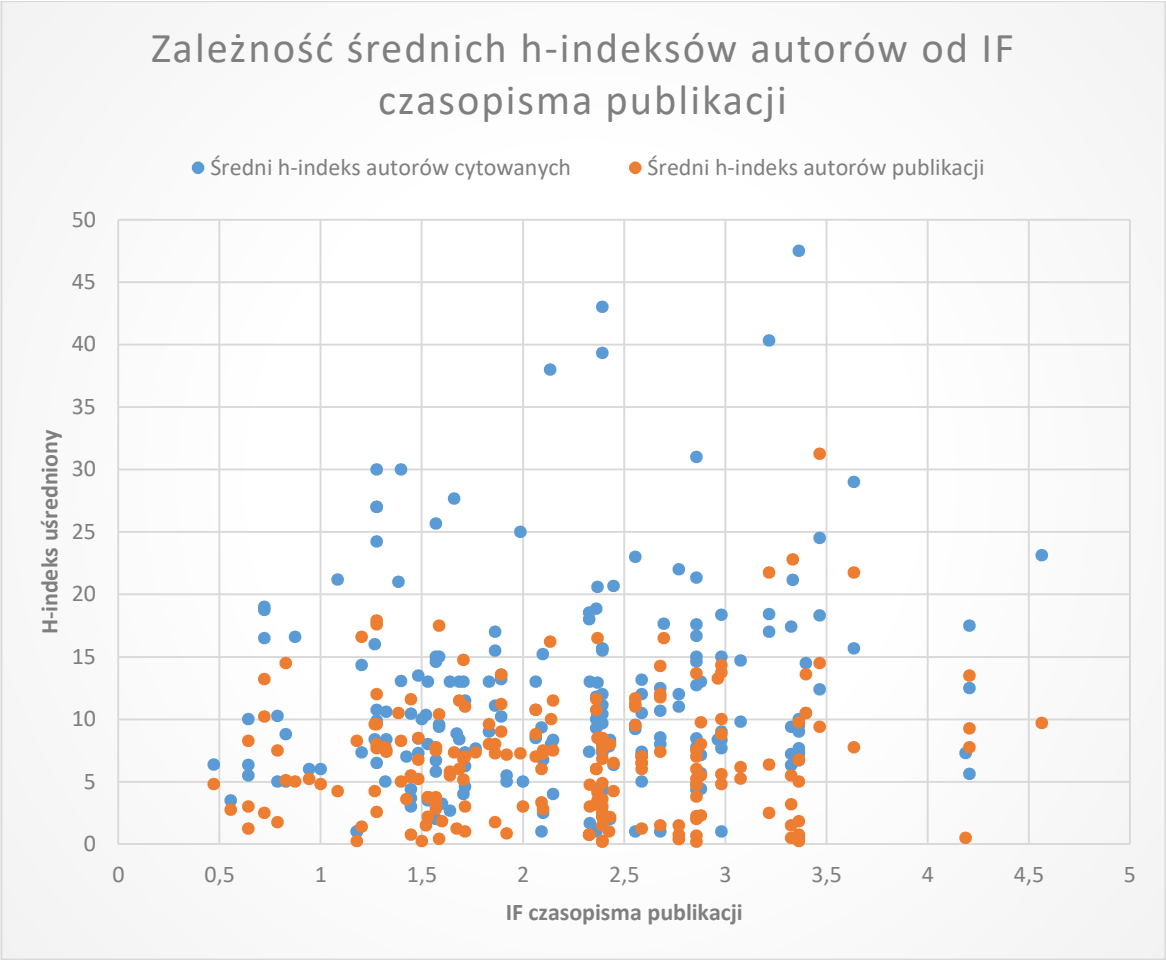
Na rysunkach 17 i 18 przedstawiono zależności pomiędzy maksymalnym oraz minimalnym wskaźnikiem wpływu czasopism cytowanych a IF czasopisma publikacji. Współczynniki korelacji wynoszą kolejno 0,1999063 i 0,1629341, co również świadczy

o słabej zależności. Na podstawie otrzymanych wyników można stwierdzić, że dobór cytowań nie ma bezpośredniego wpływu na publikację, aczkolwiek zależność cytowań do IF czasopisma publikacji jest 4-krotnie większa od h-indeksu autorów publikacji. Pozostałe wykresy zostały przedstawione w Dodatku B.2 Statystyki cytowań artykułuB.2 Statystyki cytowań artykułu, ze względu na ich znaczną objętość oraz podobieństwo do wykresów załączonych powyżej.

4.3 Wpływ autorów cytowanych artykułów

Zbadane zostaną również statystyki autorów z cytowanych artykułów. Sprawdzone zostanie czy H-indeks autorów z bibliografii ma wpływ na publikacje artykułu w danych czasopismach. Brane pod uwagę zostały artykuły opublikowane od 2015 roku. Z tej grupy zostało losowo wybranych 200 artykułów, w celu zapewnienia przejrzystości wykresów. Przygotowano następujące scenariusze dla każdej publikacji:

- H-index autorów uśredniony dla cytowanej pozycji,
- H-index minimalny autora dla cytowanej pozycji,
- H-index maksymalny autora dla cytowanej pozycji,
- H-index autorów uśredniony dla wszystkich cytowanych pozycji,
- H-index minimlany autora dla wszystkich cytowanych pozycji,
- H-index maksymalny autora dla wszystkich cytowanych pozycji.

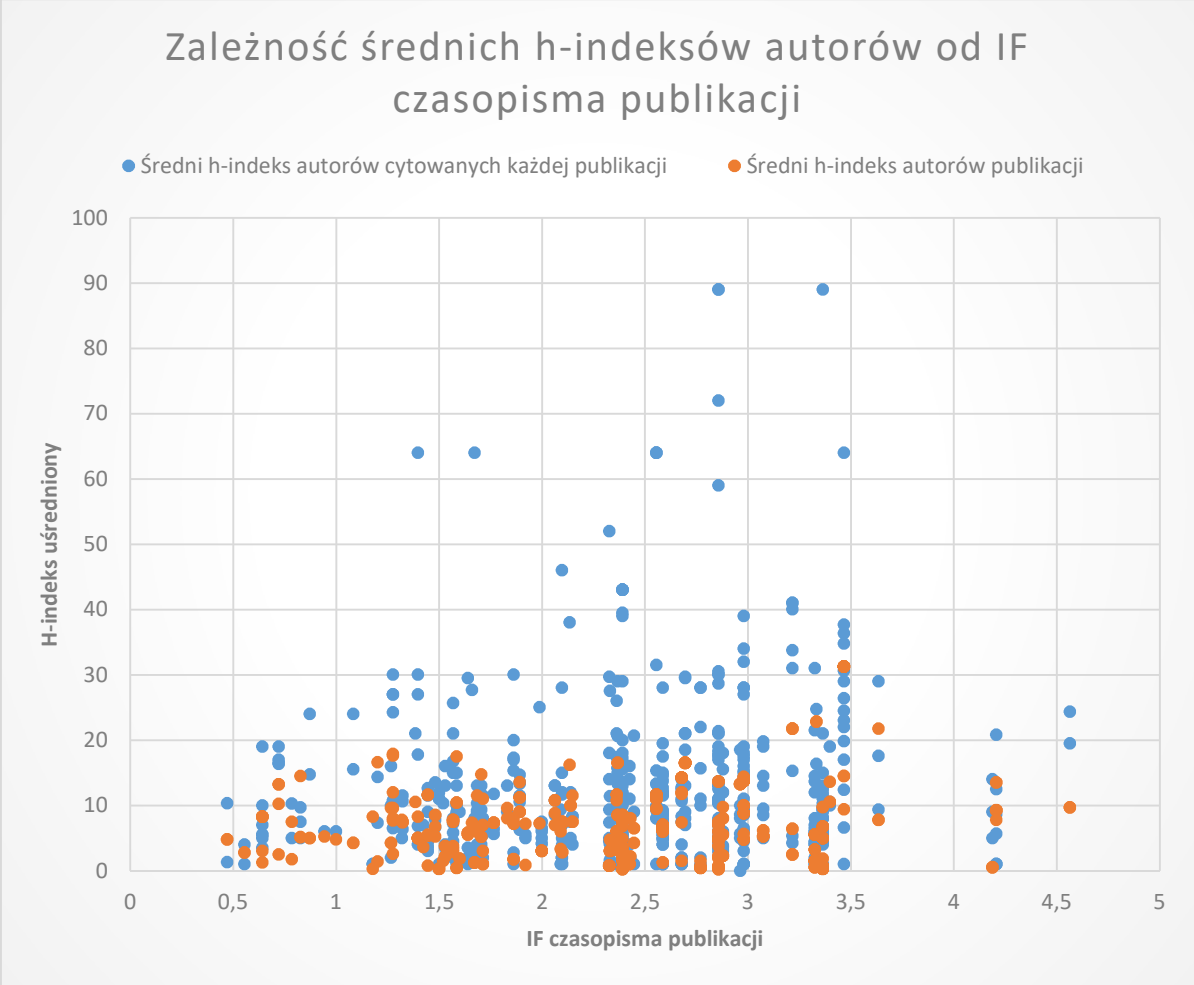


Rysunek 19 Wykres zależności średnich h-indeksów cytowanych wszystkich autorów od IF czasopisma publikacji

Na rysunku 19 zostały przedstawione zależności pomiędzy IF czasopisma publikacji, a średnią h-indeksów autorów publikacji oraz średnią h-indeksów autorów cytowanych publikacji zsumowanych razem. Współczynnik korelacji dla średniej h-indeksów autorów publikacji względem IF czasopisma wynosi 0,1140797 oraz dla średniej h-indeksów autorów cytowanych publikacji wynosi 0,0865633. Jest to bardzo słaba zależność pomiędzy h-indeksami autorów cytowanych a IF czasopisma. Natomiast jest tutaj pewna zależność pomiędzy autorami publikacji a autorami cytowanymi, która pokazuje, że autorzy cytowani mają wyższy średni indeks Hirscha od autorów publikacji. Współczynnik korelacji dla tej zależności wynosi 0,19747, zatem występuje tutaj pewna zależność, która zostanie bardziej przeanalizowana przy następnym wykresie.

Na rysunku 20 zaprezentowano związki między IF czasopisma publikacji, a średnią h-indeksów jej autorów oraz średnią h-indeksów autorów cytowanych publikacji, ale tym razem każdej z osobna. W tym przypadku współczynniki korelacji informują o bardzo słabej zależności pomiędzy średnią h-indeksów, a IF czasopism, które wynoszą dla autorów cytowanych każdej publikacji 0,0354771 i 0,1140797 dla autorów publikacji. Natomiast współczynnik korelacji między średnim h-indeksem autorów cytowanych dla każdej

publikacji z osobna a średnim h-indeksem autorów publikacji wynosi 0,1867499, co może wskazywać na słabą zależność pomiędzy nimi.



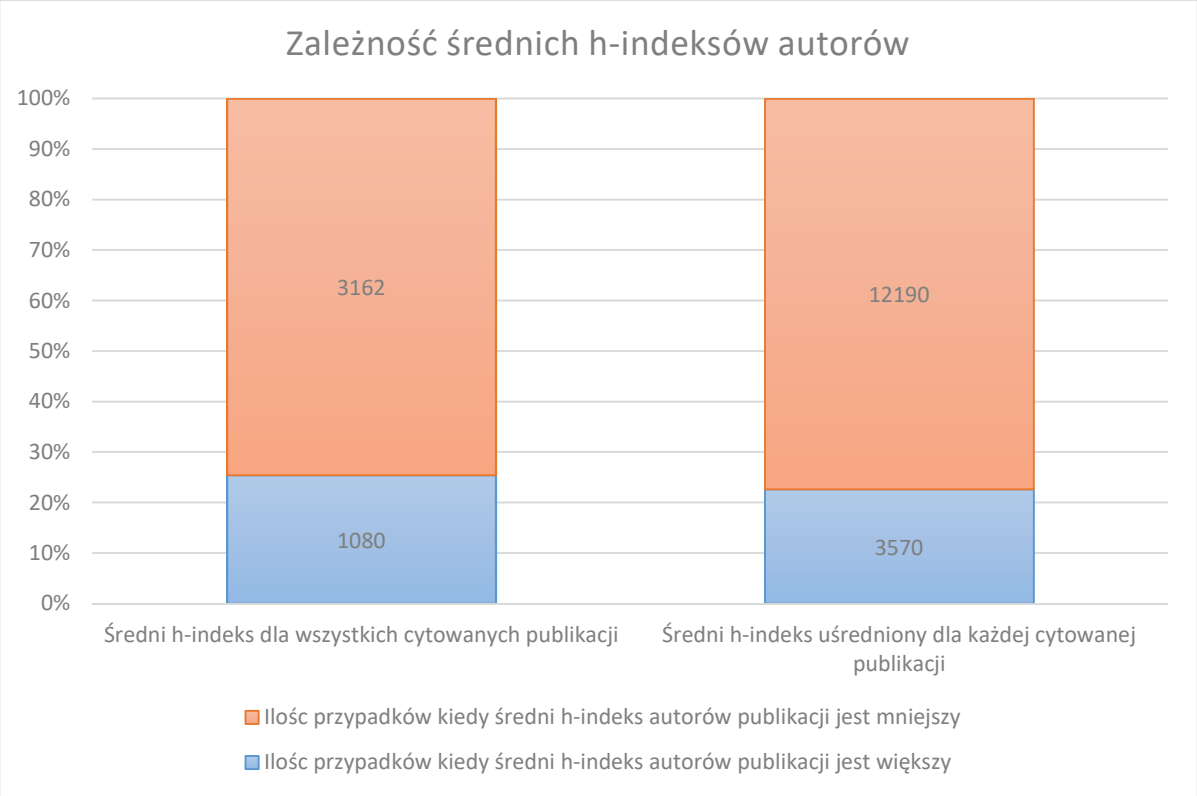
Rysunek 20 Wykres zależności średnich h-indeksów cytowanych autorów dla każdej publikacji od IF czasopisma publikacji

Na wykresie przedstawionym na rysunku 20 zauważalna jest pewna tendencja wzrostowa dla średnich h-indeksów wraz z wzrostem IF czasopisma publikacji. Wzrost jest dużo silniejszy dla średnich h-indeksów autorów cytowanych niż dla autorów publikacji. Aczkolwiek pomimo wzrostu średnich h-indeksów nadal przeważają wartości mniejsze, co również jest odzwierciedlone przez współczynnik korelacji. Wzrost wartości średniego h-indeksu może być spowodowany tym, że w czasopismach z wyższym wskaźnikiem wpływu cytowani są również autorzy z wyższym indeksem Hirscha, którzy również publikują w renomowanych czasopismach. Naturalnym jest to, że autorzy z wysokim h-indeksem publikują swoje artykuły w bardziej renomowanych czasopismach. Następnie autorzy, z niekoniecznie tak wysokim h-indeksem, którzy również w tych czasopismach publikują odwołują się do tych prac.

W celu potwierdzenia związków dotyczących cytowania autorów z wyższym indeksem Hirscha zostały przebadane matematycznie średnie h-indeksów autorów publikacji oraz

autorów cytowanych publikacji. Dało to następujące wyniki dla całego zakresu artykułów z 2015 roku spełniających podane kryteria.

Po wykonaniu obliczeń okazało się, że w ok. 75% artykułach dochodzi do sytuacji, w której autorzy cytowanych publikacji posiadają wyższe indeksy Hirscha od autorów publikacji. Może to być spowodowane tym, że autorzy odwołują się do przeszłych prac naukowców, którzy zdobyli uznanie w świecie nauki. Wyliczenia zostały przedstawione na rysunku 21.

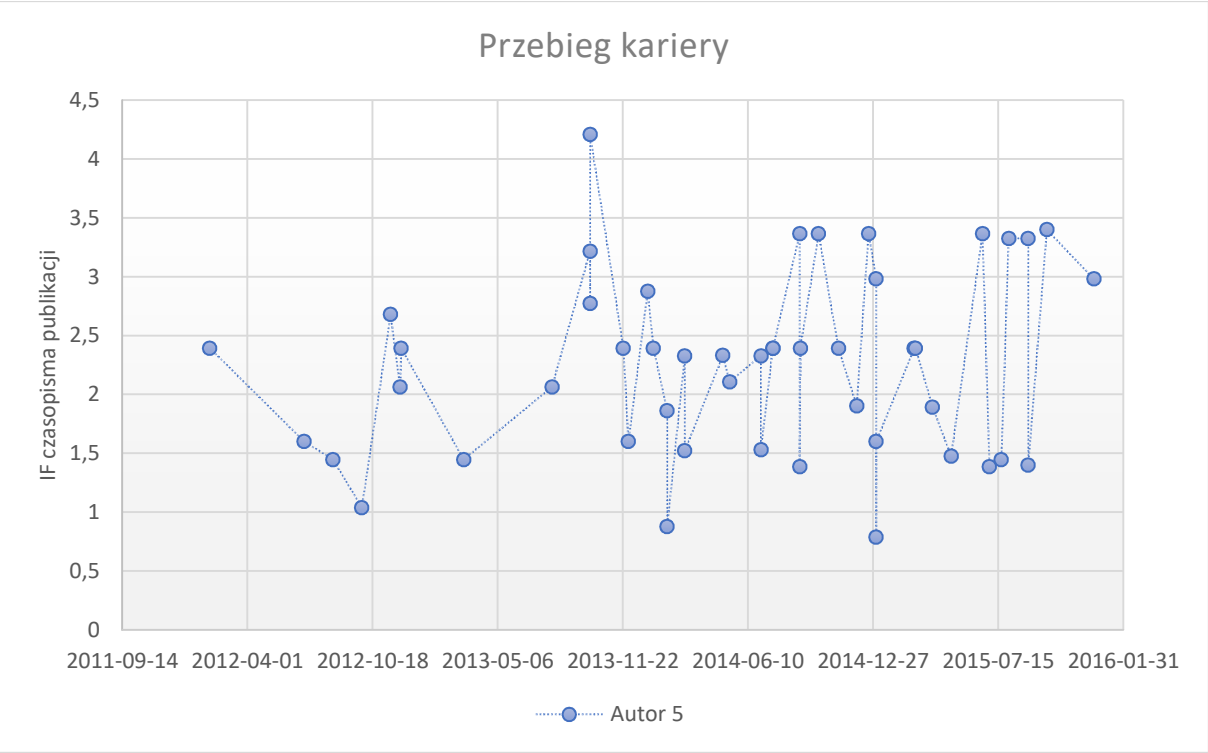


Rysunek 21 Zależność średnich h-indeksów autorów od średniej h-indeksów cytowanych autorów

Pozostałe wykresy pokazujące zależności między maksymalnym oraz minimalnym h-indeksem zostały ukazane w Dodatku B.3 Statystyki autorów cytowanych publikacji.

4.4 Wpływ historii autora

W celu sprawdzenia jak przebiega kariera naukowca zostały sporządzone wykresy ukazujące kolejne artykuły danego autora rozłożone w czasie ich publikacji.



Rysunek 22 Przebieg kariery wybranego autora

Na rysunku 16 przedstawiono zależność pomiędzy datą publikacji a IF czasopisma publikacji autora. Dodatkowe wykresy przedstawiające przebiegi karier różnych autorów zostały zamieszczone w Dodatku B.4 Przebieg kariery. Na podstawie wykresów można stwierdzić, że nie ma reguły dla kolejnych publikacji danego autora. Można zaobserwować, że jeśli autor publikuje w czasopiśmie o określonej renomie to jego następne publikacje oscylują wokół pewnego poziomu czasopism. Jeżeli autor ma już kilkanaście publikacji w uznanych czasopismach to rzadko zdarzają się publikacje w czasopismach o niższym IF, jednak publikacje mogą się ukazać w bardziej renomowanych czasopismach. Otrzymane wyniki mogą świadczyć o tym, że każdy z autorów ma wpływ na jego publikację.

4.5 Model matematyczny

Jak zostało ukazane w pracy [1] historia cytowań publikacji ma wpływ na jej przyszłe cytowania. Można przy pomocy modelu matematycznego na podstawie 5-letniej historii cytowań wyznaczyć przebieg ilości cytowań w następnych 20 latach. Dlatego wartym jest zbadanie czy przeszłe publikacje autora również mają wpływ na jego przyszłe publikacje. W poprzednim podrozdziale zostały pokazane kolejne artykuły autora, gdzie nie było widać zależności pomiędzy artykułami. Dlatego w wypadku tworzenia modelu matematycznego należy mieć na uwadze to, że publikacje naukowe mają najczęściej wielu autorów. Po ustaleniach z opiekunem pracy zostało postanowione, że w modelu matematycznym pod uwagę będzie brany pierwszy i ostatni autor publikacji. Wynika to z przyjętego założenia, że jeżeli czynniki autorów mogą mieć jakiś wpływ to powinien to być pierwszy autor tzw.

„główny autor” oraz ostatni autor tzw. „koordynator”. Za parametry autorów przyjęto wskaźnik IF czasopism ich poprzednich publikacji. Przyjęto następujący model matematyczny:

$$f = \sum_{t=1}^{T} (a_{1t} f_{1t} + a_{0t} f_{0t}) + c \tag{1}$$

Gdzie:

$f -$ współczynnik IF czasopisma przyszłej publikacji,
$T -$ liczba poprzednich publikacji autorów,
f_{1t} – IF czasopisma t ostatniej publikacji pierwszego autora,
$a_{1t} -$ waga parametru f_{1t},
f_{0t} – IF czasopisma t ostatniej publikacji ostatniego autora,
$a_{0t} -$ waga parametru f_{0t},
$c -$ współczynnik stały.

Wagi oraz współczynnik stały zostaną wyznaczone regresją liniową poprzez metodę najmniejszych kwadratów na różnych zbiorach uczących. Wyodrębnione zostały 3 zbiory składające się z autorów:

- z co najmniej 10 publikacjami,
- z co najmniej 5 publikacjami,
- z co najmniej 2 publikacjami.

Postanowiono również sprawdzić zachowanie modelu matematycznego w 3 odmiennych podejściach, które wynikają z założenia, że największe znaczenie przy przyszłych publikacjach mają ostatnie publikacje. Wynikły z tego następujące próby:

- biorąc pod uwagę po dwie ostatnie publikacje autorów (*P* = 2),
- biorąc pod uwagę po jednej ostatniej publikacji autorów (*P* = 1),
- nie biorąc pod uwagę żadnej publikacji autorów (*P* = 0).

Finalnie model regresyjny przyjął postać:

$$arg\ min_{a,c} = \sum_{i=1}^{I} \left(\sum_{t=1}^{T} (a_{1t} f_{1ti} + a_{0t} f_{0ti}) + c - f_i \right)^2 \tag{2}$$

Gdzie:

$i -$ i-ta publikacja.

Opis pozostałych parametrów jest zgodny z równaniem (1).

Tabela 4 Fragment zbioru uczącego model regresyjny

f_{11i}	f_{12i}	f_{01i}	f_{02i}	f_i
2,679	0,722	2,679	0,722	**3,364**
2,392	3,075	2,392	1,862	**3,216**
2,679	1,446	2,857	3,325	**3,325**
0,634	3,635	1,482	2,368	**2,857**
2,327	4,188	2,331	1,988	**3,325**
3,399	2,142	3,399	3,325	**2,857**
2,392	2,392	2,392	2,392	**2,981**
1,446	2,163	3,399	3,399	**2,392**

W tabeli 4 przedstawiono fragment zbioru uczącego. Zawiera on 2 poprzednie współczynniki IF czasopism publikacji dla pierwszego i ostatniego autora a także IF, w której dana para autorów opublikowała swój artykuł.

Powyższy model regresyjny został zaimplementowany w środowisku AMPL oraz Octave. Uczenie modelu odbywało się na 70% zebranych próbek, a następnie na 30% pozostałych danych następowała weryfikacja modelu za pomocą błędu średniokwadratowego oraz średniego błędu względnego. Wyniki zostały przedstawione w tabeli 8 i 9.

Rozmiar uzyskanych danych do wyznaczenia modelu został przedstawiony w poniższej tabeli 5.

Tabela 5 Rozmiar danych do modelu regresyjnego

	Razem	Dane uczące	Dane sprawdzające
Autorzy, z co najmniej 10 publikacjami	99	70	29
Autorzy, z co najmniej 5 publikacjami	339	237	102
Autorzy, z co najmniej 2 publikacjami	1224	856	368

4.5.1 AMPL

Listing 2 Implementacja modelu regresyjnego w środowisku AMPL

```
param f{i in 1..r};           #IF czasopism publikacji
param ft{1..r, 1..n};         #IF poprzednich publikacji autorów
                              #r-liczba próbek, n-liczba parametrów - 4
var a{1..n}:= 0;              #współczynniki a
var c := 0;                   #współczynnik stały c
var y{i in 1..r} = sum{j in 1..n}( a[j] * ft[i, j ] ) + c;

minimize x : sum{i in 1..r} (f[i] - y[i])^2 ;

subject to ogr1:        #ograniczenia dla jednej ostatniej publikacji
a[2] = 0;
subject to ogr2:
a[4] = 0;

subject to ogr3:        #ograniczenie, żeby nie uwzględniać publikacji
a[1] = 0;
subject to ogr4:
a[3] = 0;
```

Za pomocą implementacji zaprezentowanej na listingu 2, otrzymano wyniki przedstawione w tabelach 8 i 9 razem z wynikami z środowiska Octave. W celu uzyskania wyników pożądanych dla trzech różnych podejść modyfikowano ograniczenia w modelu AMPL. Bez ograniczeń dla dwóch ostatnich publikacji, z ograniczeniami *ogr1* oraz *ogr2* dla jednej ostatniej publikacji, z ograniczeniami *ogr1*, *ogr2*, *ogr3* i *ogr4* dla podejścia bez publikacji. Badania przeprowadzone kolejno dla autorów, z co najmniej 10, 5 oraz 2 publikacjami.

W poniższej tabeli 6 przedstawiono otrzymane parametry dla autorów, z co najmniej 2 publikacjami. Można zaobserwować, w jaki sposób zmieniają się parametry. Dla modelu bez publikacji otrzymywany jest tylko współczynnik stały, który jest średnią arytmetyczną danych uczących. W następnych modelach współczynnik ten maleje prawie 4-krotnie. W modelu z jedną ostatnią publikacją można zaobserwować, że znaczącymi stają się współczynniki ostatniej publikacji z lekką przewagą publikacji koordynatora. Zachowanie takie sugeruje, że publikacje ostatniego autora mają większy wpływ niż publikacje pierwszego autora. W modelu z dwoma ostatnimi publikacjami najważniejszym współczynnikiem nadal jest ostatnia publikacja ostatniego autora, aczkolwiek od drugiej publikacji tego autora ważniejsze są dwie ostatnie publikacje pierwszego autora. Świadczy to o tym, że najważniejsza jest ostatnia publikacja koordynatora artykułu, a pozostałe publikacje odgrywają mniejszą rolę.

Tabela 6 Otrzymane parametry modelu AMPL dla autorów, z co najmniej 2 publikacjami

	2 ostatnie publikacje	1 ostatnia publikacja	Bez publikacji
a11	0.167511	0.264129	0
a12	0.188504	0	0
a21	0.240302	0.323832	0
a22	0.135406	0	0
c	0.577918	0.882604	2.19752

4.5.2 Octave

W środowisku Octave wyniki uzyskano za pomocą operacji algebraicznej. Otrzymane wyniki przedstawiono w tabelach 8 i 9 razem z wynikami z środowiska AMPL. W celu zapewnienia wyników dla 3 różnych podejść względem ostatnich publikacji odpowiednie współczynniki zostały zerowane. Podobnie jak w przypadku środowiska AMPL badania zostały przeprowadzone dla autorów, z co najmniej 10, 5 oraz 2 publikacji. W tabeli 7 przedstawiono parametry uzyskane dla autorów, z co najmniej 2 publikacjami.

Tabela 7 Otrzymane parametry modelu Octave dla autorów, z co najmniej 2 publikacjami

	2 ostatnie publikacje	1 ostatnia publikacja	Bez publikacji
a11	0.167511	0.264129	0
a12	0.188504	0	0
a21	0.240302	0.323832	0
a22	0.135406	0	0
c	0.577918	0.882604	2.19752

Dla obu środowisk uzyskano identyczne wyniki, dlatego zostały one przedstawione w dwóch zbiorczych tabelach 8 – błąd średniokwadratowy oraz w tabeli 9 – średni błąd względny.

Tabela 8 Błąd średniokwadratowy dla uzyskanych modeli

	Autorzy, z co najmniej 10 publikacjami	Autorzy, z co najmniej 5 publikacjami	Autorzy, z co najmniej 2 publikacjami
2 ostatnie publikacje	0.43548	0.47940	**0.46001**
1 ostatnia publikacja	0.43019	0.49853	0.50839
Bez ostatnich publikacji	**0.41744**	0.74731	0.73732

Tabela 9 Średni błąd względny dla uzyskanych modeli

	Autorzy, z co najmniej 10 publikacjami	Autorzy, z co najmniej 5 publikacjami	Autorzy, z co najmniej 2 publikacjami
2 ostatnie publikacje	25,577%	29,282%	**29,204%**
1 ostatnia publikacja	25,66%	30,746%	31,567%
Bez ostatnich publikacji	**25,054%**	41,828%	41,309%

Na podstawie otrzymanych wyników najlepszym modelem wydaje się być ten, który nie bierze pod uwagę ostatnich publikacji autorów z błędem średniokwadratowym wynoszącym 0,41744 oraz średnim błędem względnym równym 25,054%. Jednak trzeba zaznaczyć, że model w tym przypadku korzystał z najmniej licznego zbioru uczącego. Również w modelach powstałych na podstawie grupy złożonej z publikacji autorów, z co najmniej 10 publikacjami w swoim dorobku występują najniższe błędy. Wynik ten spowodowany jest najmniejszą liczbą próbek oraz tym, że doszło do stabilizacji karier autorów, przez co publikują oni głównie w czasopismach o zbliżonym IF. Jest to wyraźnie widoczne przy następnych już większych zbiorach z mniej doświadczonymi autorami, w których średnie błędy względne są większe o ok. 10% dla modelu bez ostatnich publikacji w porównaniu do modelu uwzględniających ostatnie publikacje autorów. Z wymienionych powodów najlepszy model został osiągnięty dla najliczniejszego zbioru składającego się z autorów, z co najmniej 2 publikacjami, który uwzględnia dwie poprzednie publikacje. Błąd średniokwadratowy dla tego rozwiązania wynosi 0,46001 i średni błąd względny wynoszący 29,204% stanowią najmniejsze błędy nie uwzględniając pierwszego zbioru danych.

Na podstawie otrzymanych parametrów modelu, można dodatkowo stwierdzić, że największy wpływ na publikację ma ostatnia publikacja ostatniego autora, który jest prawie dwukrotnie większy od drugiej publikacji ostatniego autora oraz znacząco większy od publikacji pierwszego autora. Publikacje pierwszego autora mają zbliżone do siebie wagi z nieznaczną przewagą drugiej publikacji nad ostatnią.

Niestety wyniki te są obarczone dość wysokim błędem z tego względu aproksymacja stworzonymi modelami i wyznaczonymi współczynnikami może nie dać zadawalających wyników. Głównym powodem tego jest brak bezpośrednich zależności pomiędzy przeszłymi artykułami autora, co zostało przedstawione w rozdziale 4.4. Aczkolwiek model dowiódł, że przeszłe publikacje dwóch autorów mają wpływ na jego publikację, chociaż nie jest on tak duży jak by tego oczekiwano.

Z powodu osiągnięcia niezadowalających wyników zostało postanowione ulepszenie modelu matematycznego przez zmianę parametrów. Pierwszym podejściem było rozszerzenie zbioru uczącego o więcej niż dwie ostatnie publikacje autorów, aczkolwiek podejście to nie dało efektywniejszych estymacji.

Następnym krokiem było dodanie nowych parametrów do przyjętego modelu matematycznego. Mając do dyspozycji pobrane statystyki autorów można było podjąć próby dobrania odpowiednich parametrów w celu ulepszenia modelu.

$$f = \sum_{t=1}^{T} (a_{1t} f_{1t} + a_{0t} f_{0t}) + \sum_{p=1}^{P_1} w_{1p} p_{1p} + \sum_{p=1}^{P_0} w_{0p} p_{0p} + c \qquad (3)$$

Gdzie:

$f -$ współczynnik IF czasopisma przyszłej publikacji,
$T -$ liczba poprzednich publikacji autorów,
f_{1t} – IF czasopisma t ostatniej publikacji pierwszego autora,
$a_{1t} -$ waga parametru f_{1t},
f_{ot} – IF czasopisma t ostatniej publikacji ostatniego autora,
$a_{ot} -$ waga parametru f_{ot},
$p_{1p} -$ statystyka p autora pierwszego,
$w_{1p} -$ waga parametru p_{1p},
$p_{op} -$ statystyka p autora ostatniego,
$w_{op} -$ waga parametru p_{op},
$P_1 -$ liczba parametrów autora pierwszego,
$P_0 -$ liczba parametrów autora ostatniego,
$c -$ współczynnik stały.

Mając podany model w przedstawionej formie należało zdecydować jakie parametry autorów są istotne dla publikacji. Zostały wybrane następujące parametry autorów:

- h-indeks,
- liczba artykułów,
- liczba miesięcy od ostatniej publikacji,
- liczba wszystkich cytowań.

Należało sprawdzić, które z pojedynczych parametrów mogą mieć wpływ na publikację, a następnie sprawdzić ich wspólny dobór oraz różne kombinacje. Poniżej zaprezentowane zostały tylko istotne wyniki. Obliczenia bazowały na modelu stworzonym dla autorów, z co najmniej dwiema publikacjami.

4.5.3 H-indeks

Na początku jako kolejny parametr został przyjęty h-indeks zarówno pierwszego jak i ostatniego autora. W modelu (3) przyjęto następujący zestaw dodatkowych parametrów:

$$P_1 = 1,$$
$$P_0 = 1,$$
$$p_{11} - \text{ h-indeks autora pierwszego},$$
$$p_{o1} - \text{ h-indeks autora ostatniego},$$

W rozwiązaniu jak w poprzednich próbach w celu znalezienia najmniejszego błędu uwzględniono h-indeksy pierwszego i ostatniego autora, a także tylko pierwszego oraz tylko ostatniego autora. Wyniki są przedstawione w poniższej tabeli 10.

Tabela 10 Błędy estymacji dla modelu z h-indeksem autorów

	h-indeks obu autorów	h-indeks pierwszego autora	h-indeks ostatniego autora
Błąd średniokwadratowy	0,459853	0,460102	0,460093
Średni błąd względny	29,271%	29,209%	29,284%

Porównując otrzymane wyniki z wynikami otrzymanego dla modelu uwzględniającego tylko ostatnie publikacje (Tabela *9* Średni błąd względny dla uzyskanych modeli, kolumna *4*) można stwierdzić, że wprowadzenie dodatkowego parametru h-indeksu autora nie poprawia stworzonego już modelu regresji. Otrzymane modele różnią się bardzo nieznacznie pod kątem błędu średniokwadratowego oraz błędu względnego. H-indeks został w modelu regresji praktycznie pomijalny. Otrzymane wagi dla modelu przedstawiono w tabeli 14.

4.5.4 Liczba artykułów

Kolejnym rozpatrywanym parametrem była całkowita liczba artykułów autora. W (3) zdefiniowano dodatkowe parametry:

$$P_1 = 1,$$
$$P_0 = 1,$$
$$p_{11} - \text{ liczba artykułów autora pierwszego},$$
$$p_{o1} - \text{ liczba artykułów autora ostatniego}.$$

Również jak w poprzednim przypadku brano pod uwagę parametr całkowitej liczby artykułów dla pary autorów, jak i dla każdego z osobna. Wyniki przedstawiono w tabeli 11.

Tabela 11 Błędy estymacji dla modelu z całkowitą liczbą artykułów autorów

	liczba artykułów	liczba artykułów	liczba artykułów

	dla obu autorów	pierwszego autora	ostatniego autora
Błąd średniokwadratowy	0,459645	0,459302	0,460355
Średni błąd względny	29,212%	29,173%	29,241%

Otrzymane wyniki są bardzo zbliżone do pierwotnych wyników. Podobnie jak w przypadku h-indeksu autorów liczba artykułów również jest praktycznie nieznacząca w procesie publikacji.

4.5.5 Liczba miesięcy od ostatniej publikacji

Następnym analizowanym parametrem był czas między kolejnymi publikacjami autora. Został on wyrażony za pomocą liczb całkowitych oznaczających liczbę miesięcy pomiędzy publikacjami autora. W (3) tak zostały przedstawione dodatkowe parametry:

$$P_1 = 1,$$
$$P_0 = 1,$$
$$p_{11} - \text{ liczba miesięcy od poprzedniej publikacji autora pierwszego,}$$
$$p_{o1} - \text{ liczba miesięcy od poprzedniej publikacji autora ostatniego.}$$

Tym razem również badano wpływ czasu dla pary autorów oraz dla każdego autora osobno. Tabela 12 zawiera wyniki uzyskane za pomocą otrzymanego modelu regresji.

Tabela 12 Błędy estymacji dla modelu z liczbą miesięcy od ostatniej publikacji autorów

	Liczba miesięcy dla obu autorów	Liczba miesięcy dla pierwszego autora	Liczba miesięcy dla ostatniego autora
Błąd średniokwadratowy	0,462097	0,459258	0,461201
Średni błąd względny	29,405%	29,245%	29,415%

Podobnie jak w poprzednich przypadkach również i tym razem dodatkowy parametr jakim jest liczba miesięcy między publikacjami nie wpływa na polepszenie modelu.

4.5.6 Liczba cytowań autora

Ostatnim parametrem dodanym do modelu matematycznego była liczba cytowań autora. Liczba ta przedstawia ile razy artykuły autora były cytowane. W (3) tak zostały przedstawione dodatkowe parametry:

$$P_1 = 1,$$

$P_0 = 1,$
$p_{11} -$ liczba cytowań autora pierwszego,
$p_{o1} -$ liczba cytowań autora ostatniego.

Tak jak poprzednio tak i tym razem badano wpływ liczby cytowań pierwszego, ostatniego oraz pary autorów. Tabela 13 przedstawia zebrane wyniki uzyskane za pomocą otrzymanego modelu regresji.

Tabela 13 Błędy estymacji dla modelu z liczbą cytowań

	Liczba cytowań dla obu autorów	Liczba cytowań dla pierwszego autora	Liczba cytowań dla ostatniego autora
Błąd średniokwadratowy	0,460231	0,460163	0,460151
Średni błąd względny	29,281%	29,207%	29,279%

Również i tym razem dodatkowy parametr nie ma wpływu na publikację, ponieważ wyniki pozostają bardzo zbliżone do wyników otrzymanych podstawowym modelem regresji (2).

Tabela 14 Współczynniki modelu matematycznego dla dodatkowych parametrów

	h-indeks	liczba artykułów	Liczba miesięcy od publikacji	Liczba cytowań autora
a11	0.158811	0.157538	0.154907	0.158777
a12	0.205444	0.20342	0.201847	0.207358
a01	0.237674	0.238099	0.231728	0.237928
a02	0.132927	0.136452	0.14795	0.129039
w11	-0.000390381	-0.000126954	-0.00141937	0.00000146
w01	0.00225191	0.000060097	0.00631	0.00002211
c	0.577918	0.569988	0.607883	0.556442

W tabeli 14 przedstawiono otrzymane współczynniki dla poszczególnych modeli regresji. Tak jak w przypadku modelu (2) największe znaczenie odgrywają współczynniki dla poprzednich publikacji autorów. Natomiast praktycznie pomijalne są dodatkowe wprowadzone parametry, których wagi oscylują wokół zera. Badanie to udowadnia, że statystyki autorów nie mają praktycznie wpływu na publikację. W tym celu przeprowadzono dodatkową analizę wykluczając z modelu poprzednie publikacje autorów.

Tabela 15 Wagi współczynników dla modelu bez poprzednich publikacji autorów

	wagi dla pierwszego autora	wagi dla ostatniego autora
h-indeks	0.00967081	0.00405463

Liczba artykułów	−0.00150342	−0.000492085
Liczba miesięcy od publikacji	0.00319803	0.00614287
Liczba cytowań	0.000022314	0.00003829
c	2.24002	
Błąd średniokwadratowy	0,723725	
Średni błąd względny	41,121%	

W tabeli 15 przedstawiono otrzymane wagi dla parametrów, na której podstawie można potwierdzić, że statystki pierwszego i ostatniego autora nie mają bezpośredniego wpływu na publikację artykułu. Wagi dla parametrów są bliskie zeru i znaczącym jest jedynie stały parametr *c*. Błąd również znacząco wzrósł. Błąd średniokwadratowy zwiększył się o 0,26 natomiast średni błąd względny o ponad 10%.

Wykonano jeszcze próbę z wszystkimi dostępnymi parametrami. Otrzymane wyniki zostały przedstawione w poniższej tabeli 16.

Tabela 16 Model regresyjny dla wszystkich parametrów

	wagi dla pierwszego autora	wagi dla ostatniego autora
a1	0.154315	0.231356
a2	0.20481	0.13897
h−indeks	0.00195567	0.00267095
Liczba artykułów	−0.000402507	−0.000340641
Liczba miesięcy od publikacji	−0.00121021	0.00646716
Liczba cytowań	0.00002748	0.00002461
c	2.24002	
Błąd średniokwadratowy	0,459879	
Średni błąd względny	29,35%	

Na podstawie otrzymanych wyników można potwierdzić, że statystyki autorów nie są istotne przy publikacji artykułów, oraz że najważniejszym współczynnikiem podczas publikacji jest ostatni artykuł ostatniego autora. Waga tego współczynnika nie jest wystarczająco duża, aby móc mówić o znaczącej zależności. Uzyskane wysokie błędy również oznaczają brak silnej zależności.

5 Podsumowanie

W celu przeprowadzenia analizy dotyczącej znalezienia relacji między cytowanymi artykułami, samym artykułem, a także innych określonych w pracy relacji należało najpierw przygotować wiele informacji.

Nie byłoby to możliwe bez znalezienia odpowiedniego źródła danych. Mimo pierwszych prób z bazą Google Scholar rozwiązanie zostało to porzucone ze względu na napotkane problemy, a także dzięki znalezieniu lepszego źródła, jakim okazała się być baza Scopus od wydawnictwa Elsevier. Elsevier proponuje korzystanie ze swojego API w celu pobierania informacji z ich bazy, co bardzo ułatwia pracę oraz umożliwia sprawne przeprowadzenie procesu wyłuskiwania danych. Łącznie z bazy Scopus pobrano informacje o blisko 100 000 artykułach, 200 000 autorach oraz 150 000 afiliacjach i wiele więcej w celu powiązania tych danych.

W pracy nie stwierdzono występowania zależności pomiędzy indeksem Hirscha autorów a publikacjami w czasopismach o określonym IF, bądź SJR. Natomiast zostało pokazane, że można używać obu współczynników IF oraz SJR w celu odniesienia się do klasyfikacji czasopisma. Oczywiście nie można tych współczynników mieszać ze sobą, ponieważ ich wartości są rozbieżne, aczkolwiek w ogólnym rozrachunku hierarchia czasopism jest bardzo podobna.

Nie znaleziono również zależności pomiędzy artykułem a artykułami przez niego cytowanymi, biorąc pod uwagę IF czasopism ich publikacji. Analizując h-indeksy autorów publikacji oraz autorów cytowanych publikacji w stosunku do IF czasopisma publikacji również nie znaleziono bezpośredniej zależności, natomiast znaleziona została zależność pomiędzy samymi autorami. Według tej zależności blisko 75% autorów publikacji cytuje autorów z większym średnim h-indeksem od własnych h-indeksów. Aczkolwiek sytuacja ta raczej nie ma wpływu na publikację, a może wynikać z prostego faktu, że w artykułach zazwyczaj są odwołania do uznanych publikacji w celu określenia dziedziny artykułu.

Na podstawie wykresów kolejnych artykułów danego autora, również nie są widoczne żadne zależności pomiędzy kolejnymi publikacjami. Można stwierdzić, że autor z wieloma publikacjami utrzymuje swoje prace w czasopismach o zbliżonym IF, co może świadczyć o pewnej stabilizacji kariery.

Model regresji nie sprawdził się w sposób w jaki zakładano. Korzystanie z niego w celu estymacji następnej publikacji wybranych naukowców jest obarczone dużym błędem wynoszącym około 30%, aczkolwiek jest to możliwe. Na podstawie otrzymanych parametrów zostało pokazane, że największy wpływ na publikację ma ostatni artykuł koordynatora publikacji. W karierze naukowca nie zachodzą takie relacje jak w przypadku cytowań jego

publikacji, gdzie po danym okresie czasu można estymować z sukcesem, jakie będą dalsze losu danego artykułu. Natomiast udowodniono w modelu, że statystki autorów nie mają wpływu na publikację.

Kolejnym krokiem w celu znalezienia pewnych zależności oraz stworzenia lepszego modelu matematycznego mogłoby być znalezienie danych historycznych zarówno o autorach jak i czasopismach. Pozwoliło by to w pełni zbadać przebieg kariery opierając się na właściwych statystykach dla właściwego momentu w karierze danego naukowca. Innym podejściem mogłoby być również wykonanie analizy sieci społecznych. Analiza wśród autorów, mogłaby wykazać pewne zależności dotyczące wyboru bibliografii do artykułów.

Reasumując, nie udało się w mojej pracy wyznaczyć zależności pomiędzy statystykami autorów lub cytowań a publikacją w określonym czasopiśmie. Stworzony model jest obarczony dużym błędem względnym, aby mógł sukcesywnie estymować IF czasopism przyszłych prac naukowców. Wyniki te, chociaż są negatywne to powinny być odbierane pozytywnie. Brak znalezionych zależności oznacza, że zarówno statystyki publikacji zawartej w bibliografii jak i autorów cytowanych oraz statystyki autorów publikujących nie odgrywają roli w procesie publikacji artykułów naukowych.

6 Literatura

1. *Wang D, Song C, Barabási AL. , Quantifying long-term scientific impact, 2013* http://science.sciencemag.org/content/342/6154/127.full [dostęp 05.02.2017]
2. *Richard Van Noorden, Publishers withdraw more than 120 gibberish papers, 2014* http://www.nature.com/news/publishers-withdraw-more-than-120-gibberish-papers-1.14763 [dostęp 01.02.2017]
3. *Cyril Labbé, Ike Antkare one of the great stars in the scientific firmament, 2010* http://evaluation.hypotheses.org/files/2010/12/pdf_IkeAntkareISSI.pdf [dostęp 01.02.2017]
4. *John Bohannon, „Who's Afraid of Peer Review?", 2013* http://science.sciencemag.org/content/342/6154/60.full [dostęp 02.02.2017]
5. *Anne-Wil Harzing, Satu Alakangas, Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, 2016* http://link.springer.com/article/10.1007/s11192-015-1798-9 [dostęp 05.02.2017
6. *Ramin Sadeghi, Comparison between impact factor(IF), and Scimago Journal Rank Indicator (SJR) for scientometrics, 2014* https://www.linkedin.com/pulse/20141116123259-267597660-comparison-between-impact-factor-if-scimago-journal-rank-indicator-sjr-for-scientometrics [dostęp 01.02.2017]
7. https://en.wikipedia.org/wiki/Academic_publishing#Scholarly_paper [dostęp 27.01.2017]
8. http://www.nauka.gov.pl/g2/oryginal/2015_06/57d62136155875b12419981aa086b9f9.pdf [dostęp 27.01.2017]
9. https://en.wikipedia.org/wiki/Scientific_journal [dostęp 27.01.2017]
10. http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf [dostęp 27.01.2017]
11. http://scientific.thomsonreuters.com/imgblast/JCRFullCovlist-2016.pdf [dostęp 27.01.2017]
12. http://www.bip.nauka.gov.pl/g2/oryginal/2017_01/eef3ce53f11d75c8345087358d8f65e4.pdf [dostęp 28.01.2017]
13. http://www.nauka.gov.pl/lista-czasopism-punktowanych/ [dostęp 28.01.2017]
14. https://en.wikipedia.org/wiki/H-index [dostęp 28.01.2017]
15. https://scholar.google.pl/intl/pl/scholar/about.html [dostęp 30.01.2017]
16. https://en.wikipedia.org/wiki/Google_Scholar [dostęp 30.01.2017]
17. https://pbn.nauka.gov.pl/sedno-webapp/about [dostęp 30.01.2017]
18. http://webofknowledge.com [dostęp 30.01.2017]
19. http://www.dobreprogramy.pl/okokok/How-To-Windows-8-i-tunelowanie-ruchu-sieciowego-po-SSH,37760.html [dostęp 30.01.2017]

20. https://www.researchgate.net/about [dostęp 30.01.2017]
21. https://www.researchgate.net/RGScore/FAQ [dostęp 30.01.2017]
22. https://www.scopus.com/home.uri [dostęp 30.01.2017]
23. http://dev.elsevier.com/ [dostęp 31.01.2017]
24. https://blog.scopus.com/posts/2014-snip-sjr-and-ipp-journal-metrics-now-freely-available-online [dostęp 31.01.2017]
25. https://www.cwts.nl/blog?article=n-q2y254 [dostęp 31.01.2017]
26. https://jsoup.org/ [dostęp 04.02.2017]
27. https://github.com/FasterXML/jackson [dostęp 04.02.2017]
28. http://hibernate.org/orm/ [dostęp 04.02.2017]

Spis rysunków

Spis tabel

Dodatek A – Fragmenty kodu źródłowego

Listing 3 Przygotowanie połączenia za pomocą JSoup

```
private Connection MakeConnection(String href) {
  Connection connection = Jsoup
    .connect(href)
    .timeout(300000)
    .ignoreContentType(true)
    .userAgent(
       "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:25.0)
Gecko/20100101 Firefox/25.0")
    .followRedirects(true)
    .header("Accept", "application/json, application/atom+xml,
      application/xml");
  return connection;
}
```

Listing 4 Użycie Elsevier API do pobrania artykułów o zadanych parametrach

```
Connection connection =
  MakeConnection(http://api.elsevier.com/content/search/
                  index:SCIDIR);
connection.data("APIKey", api_key)
  .data("count", 200)
  .data("subj", ComputerScience)
  .data("start", 0)
  .data("query","affil(Poland) and pub-date AFT 20150101
      content-type(JL)")
  .method(Method.GET);
Document document = connection.execute().parse();
```

Listing 5 Pobranie DOI artykułu z pobranego dokumentu

```
ObjectMapper mapper = new ObjectMapper();
JsonNode node = mapper.readTree(document.text());
Int total = node.get("search-results").
  get("opensearch:totalResults").asInt()
for (int i = 0; i < total; i++) {
  JsonNode resultNode = node.get("search-results").
      get("entry").get(i);
  String doi = resultNode.get(getDoi).asText();
}
```

Listing 6 Ręczne parsowanie stron HTML

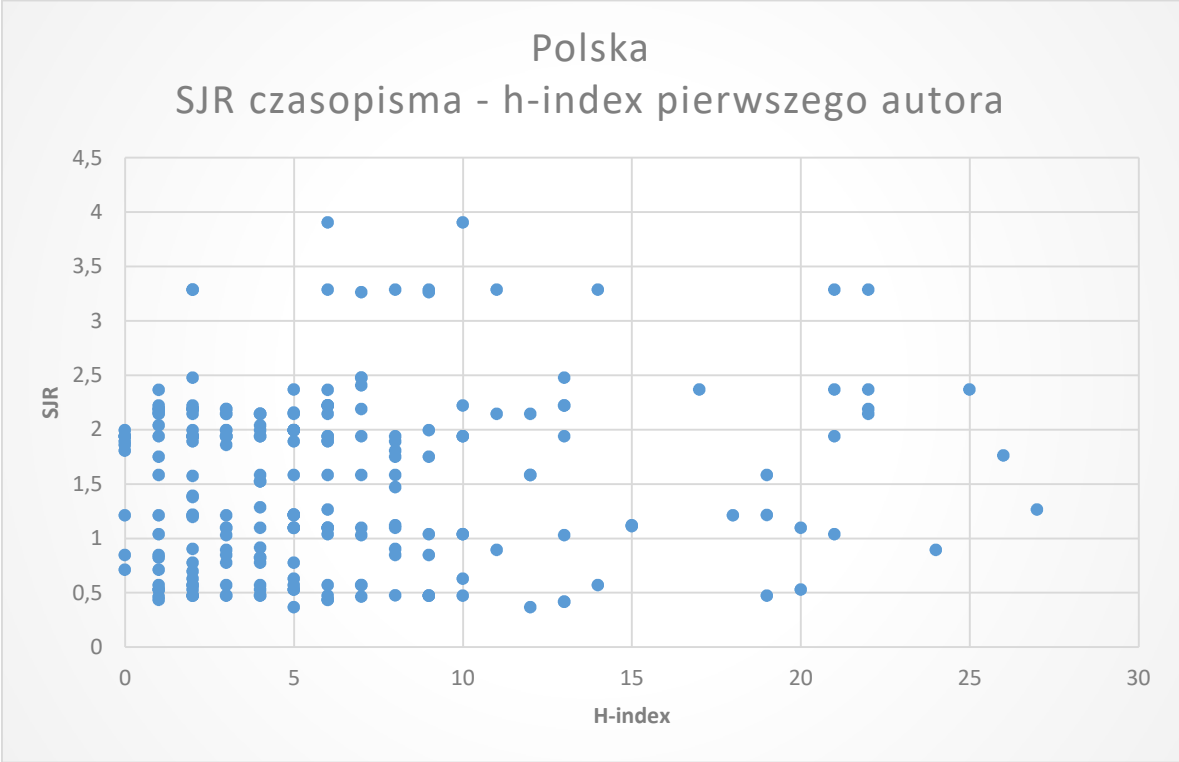
```
Document doc = Jsoup.connect(researchGate).timeout(20000).get();
Elements elements =
doc.getElementsByClass("list").first().getElementsByClass(
  "department-list-item-with-stats");
  for (Element element : elements) {
    Element elName = element.getElementsByClass("name").first();
    Element elMembers = element.getElementsByClass("stats").first()
        .getElementsByClass("stat").first();
    String name = elName.text();
    String link = elName.getElementsByTag("a").attr("href");
    elMembers = elMembers.getElementsByClass("caption").first();
    Integer members = 0;
    if (elMembers.text().equals("Members")) {
      members = Integer.parseInt(elMembers.getElementsByClass("number")
                .first().text());
    }
  }
```

Listing 7 Użycie Hibernate do dodawania oraz modyfikacji wierszy do bazy danych

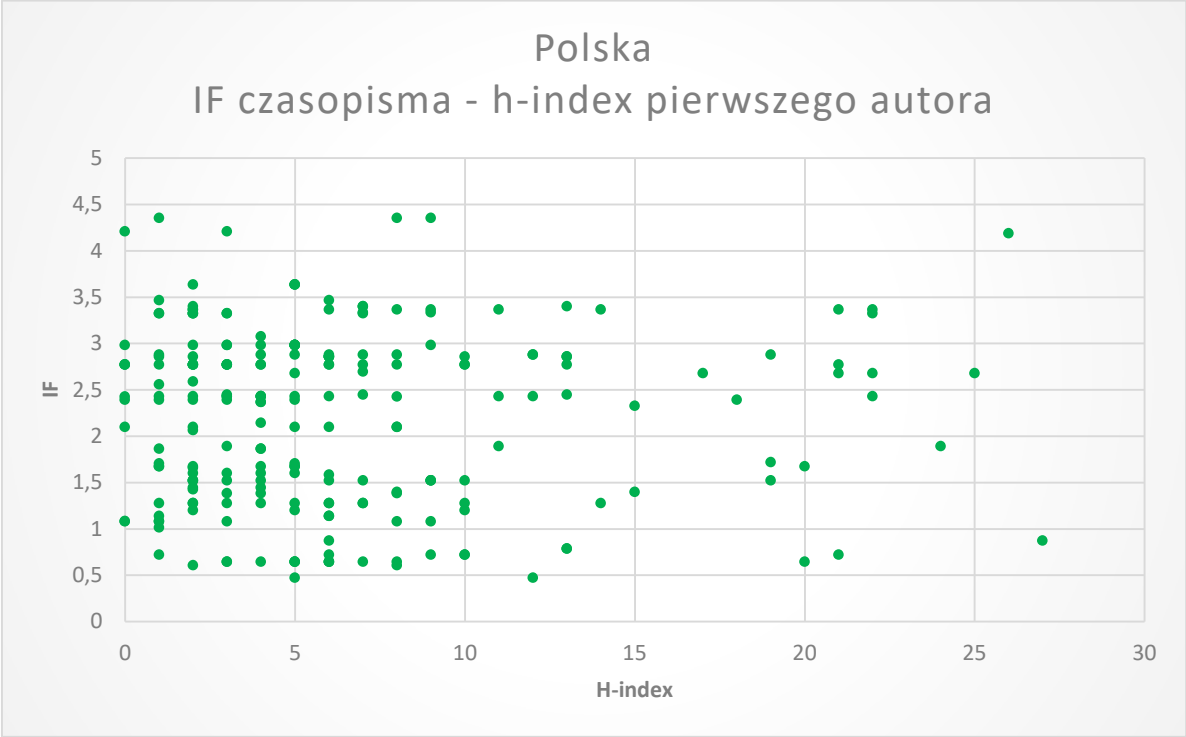
```
public static void updateDBobject(Object obj) {
  Session session =
    HibernateUtil.getSessionFactory().openSession();
  Transaction tx = session.beginTransaction();
  session.saveOrUpdate(obj);
  tx.commit();
  session.close();
}
```

Dodatek B - Wykresy

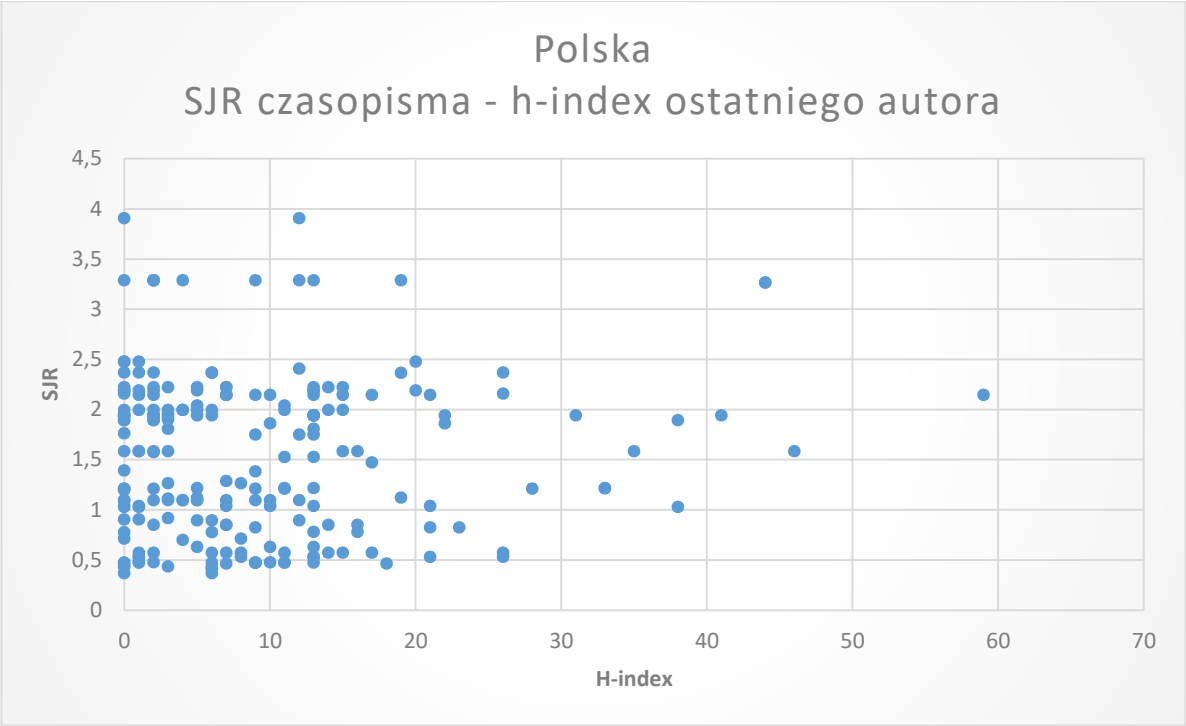
B.1 Statystyki autorów artykułu



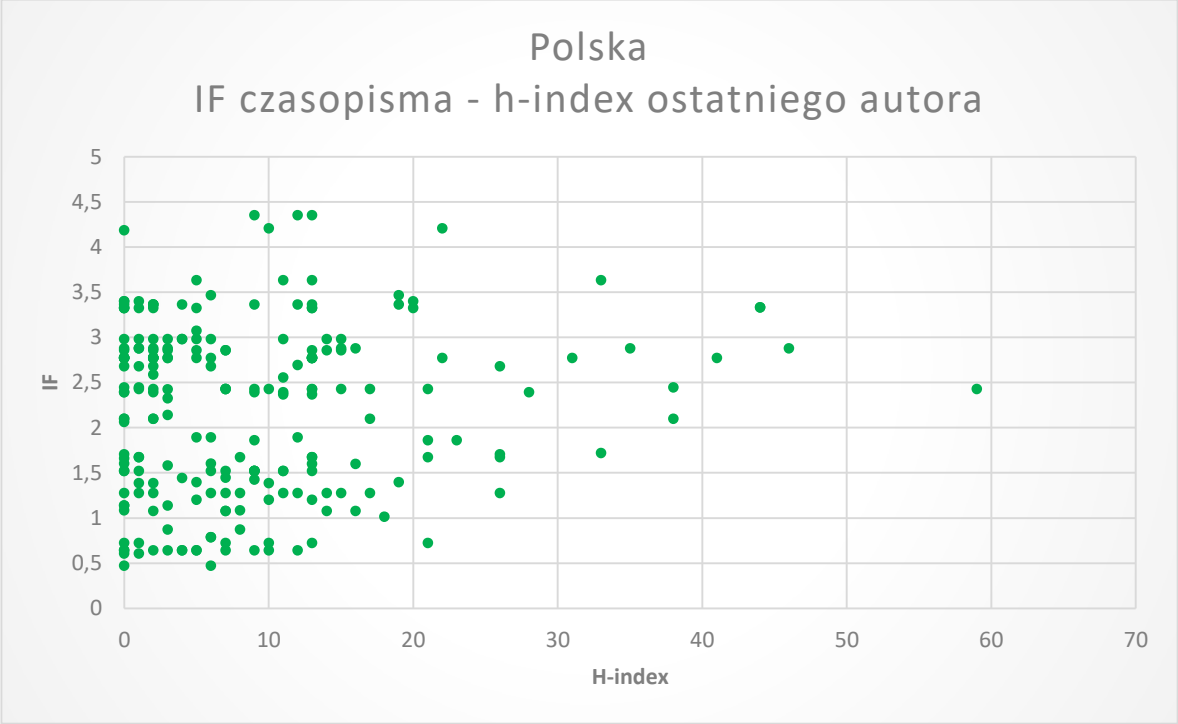
Rysunek 23 Wykres zależności h-indeksu pierwszego autora od SJR czasopisma publikacji przez polskich autorów



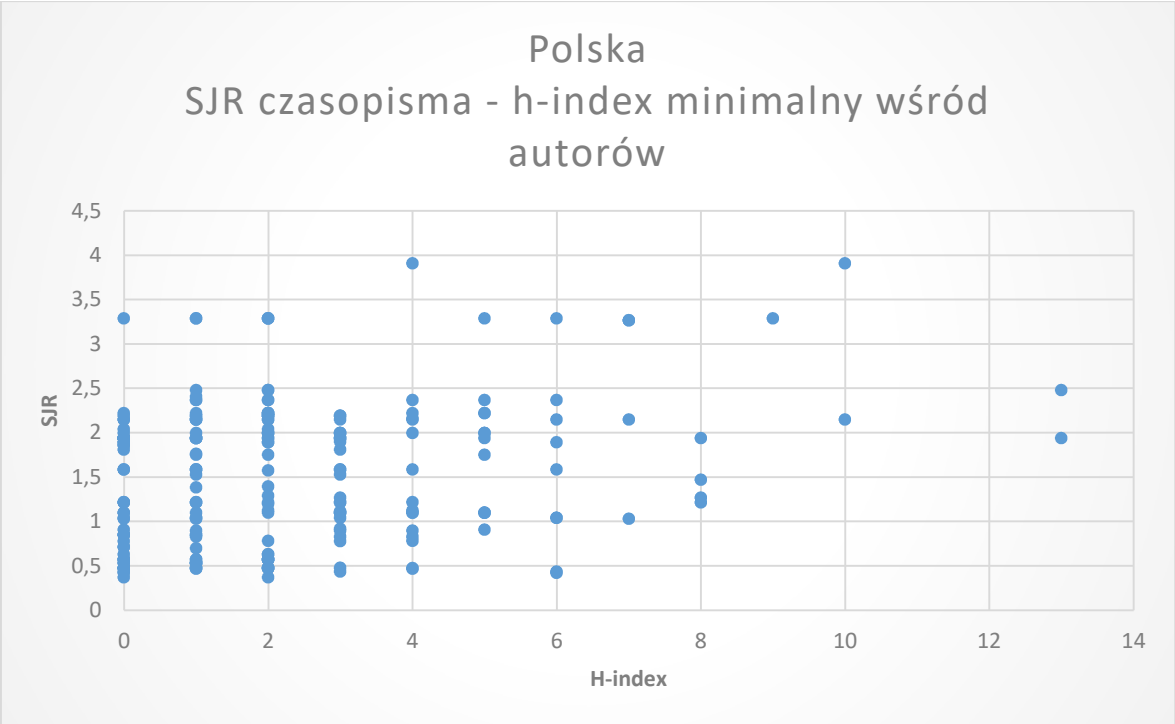
Rysunek 24 Wykres zależności h-indeksu pierwszego autora od IF czasopisma publikacji przez polskich autorów



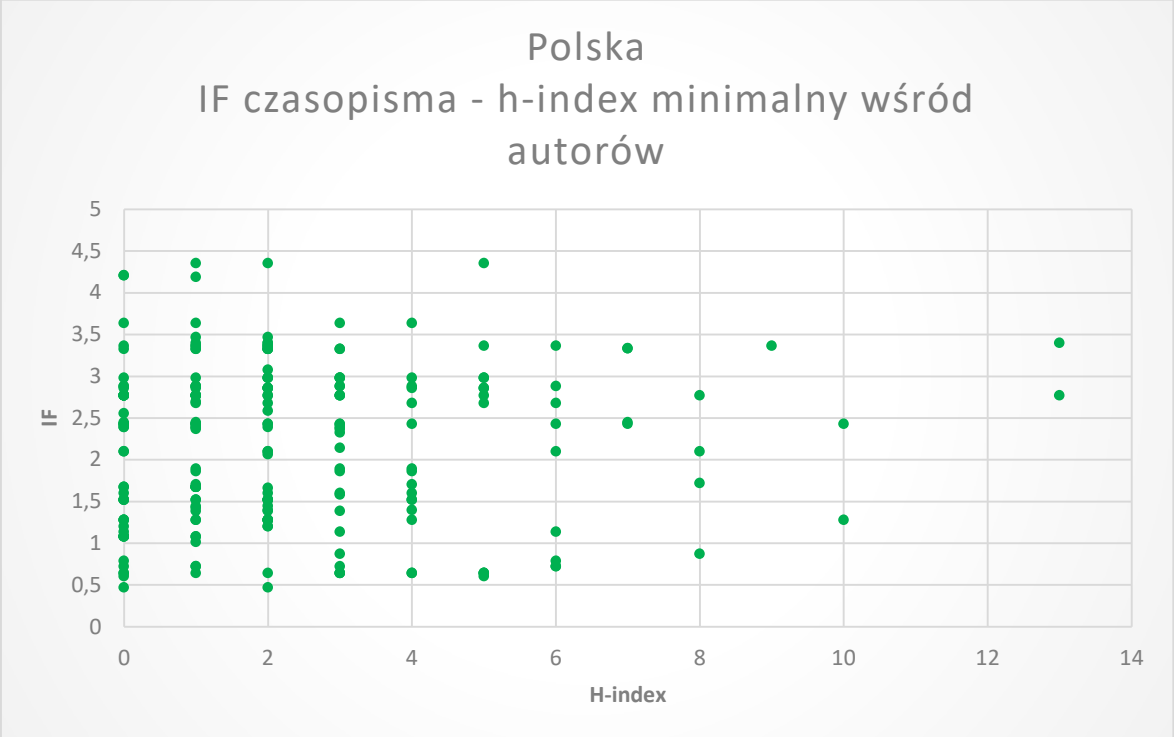
Rysunek 25 Wykres zależności h-indeksu ostatniego autora od SJR czasopisma publikacji przez polskich autorów



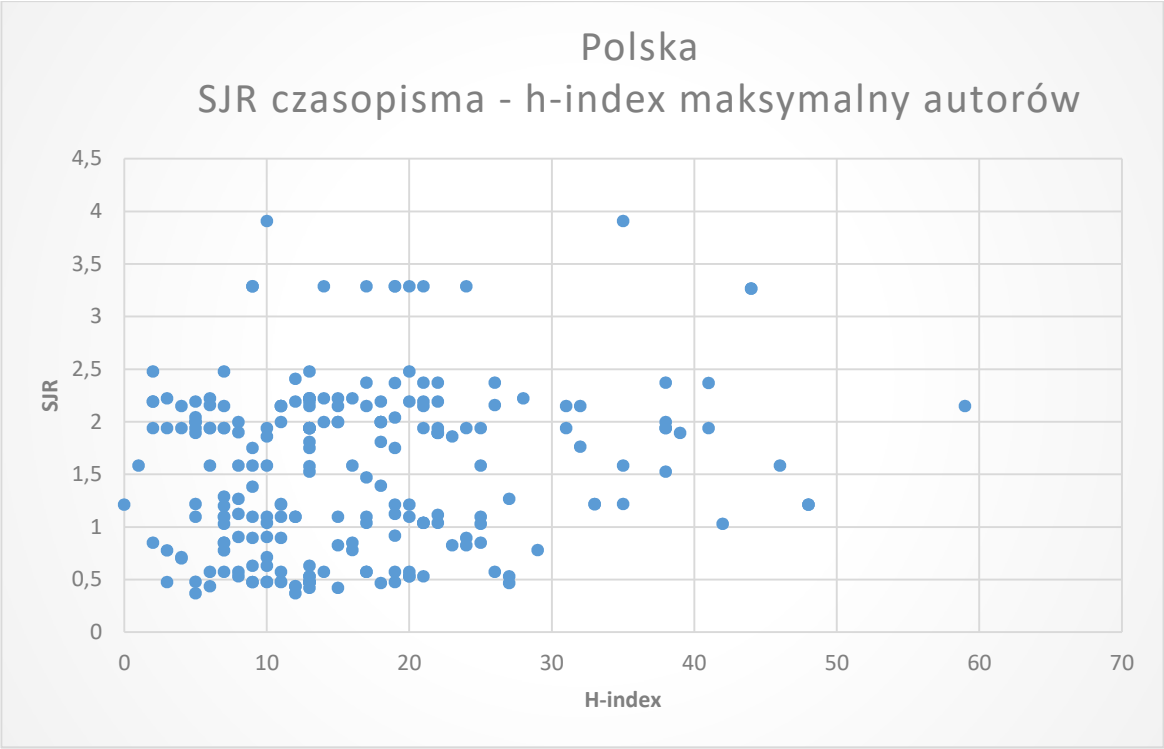
Rysunek 26 Wykres zależności h-indeksu ostatniego autora od IF czasopisma publikacji przez polskich autorów



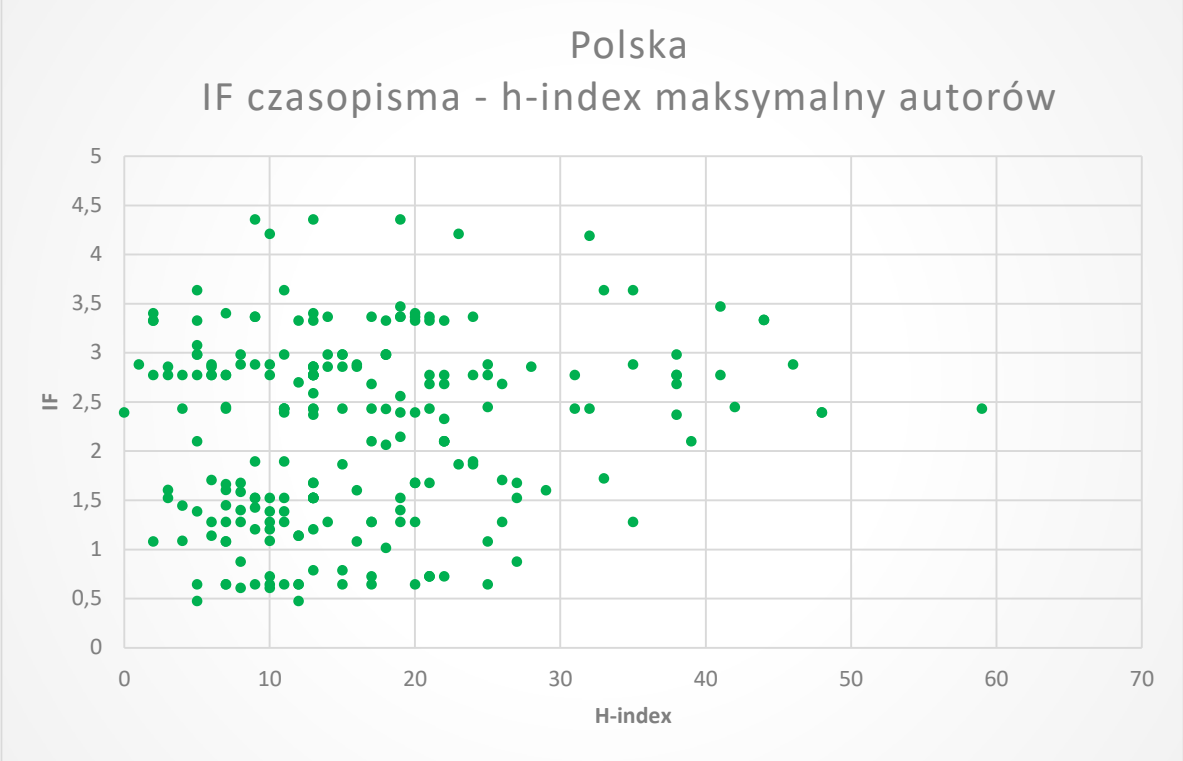
Rysunek 27 Wykres zależności minimalnego h-indeksu autorów od SJR czasopisma publikacji przez polskich autorów



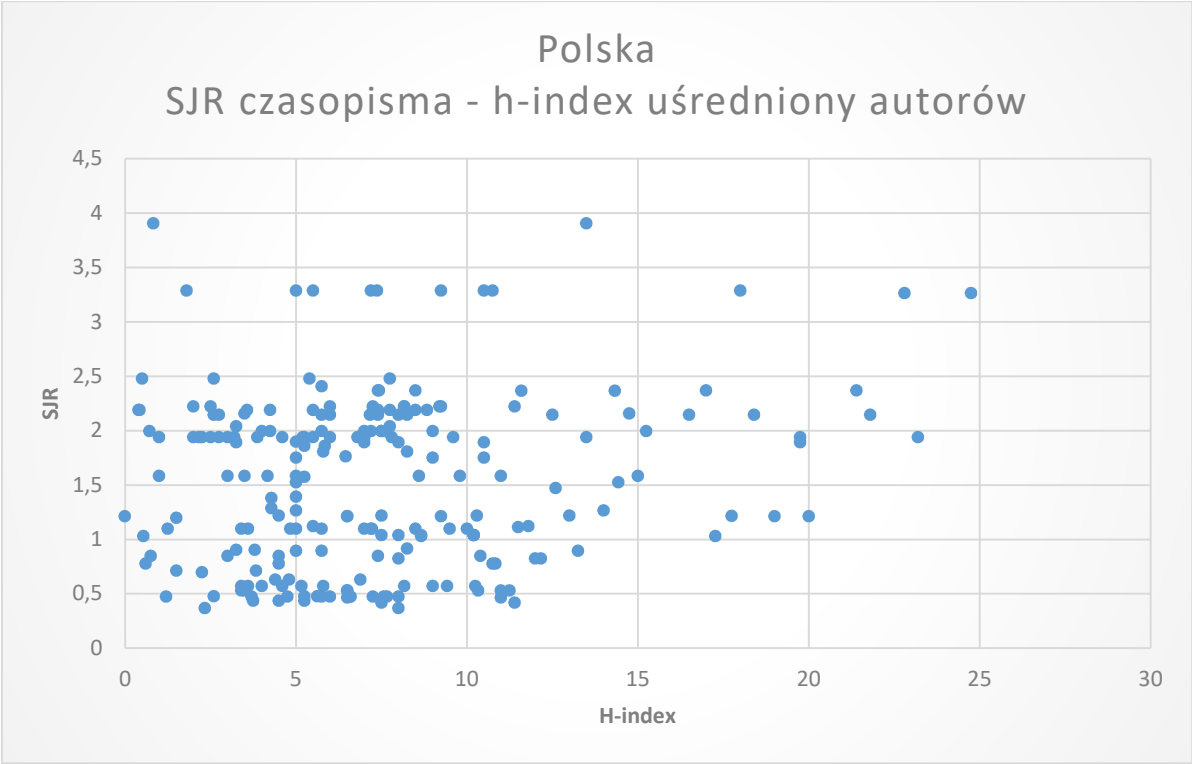
Rysunek 28 Wykres zależności minimalnego h-indeksu autorów od IF czasopisma publikacji przez polskich autorów



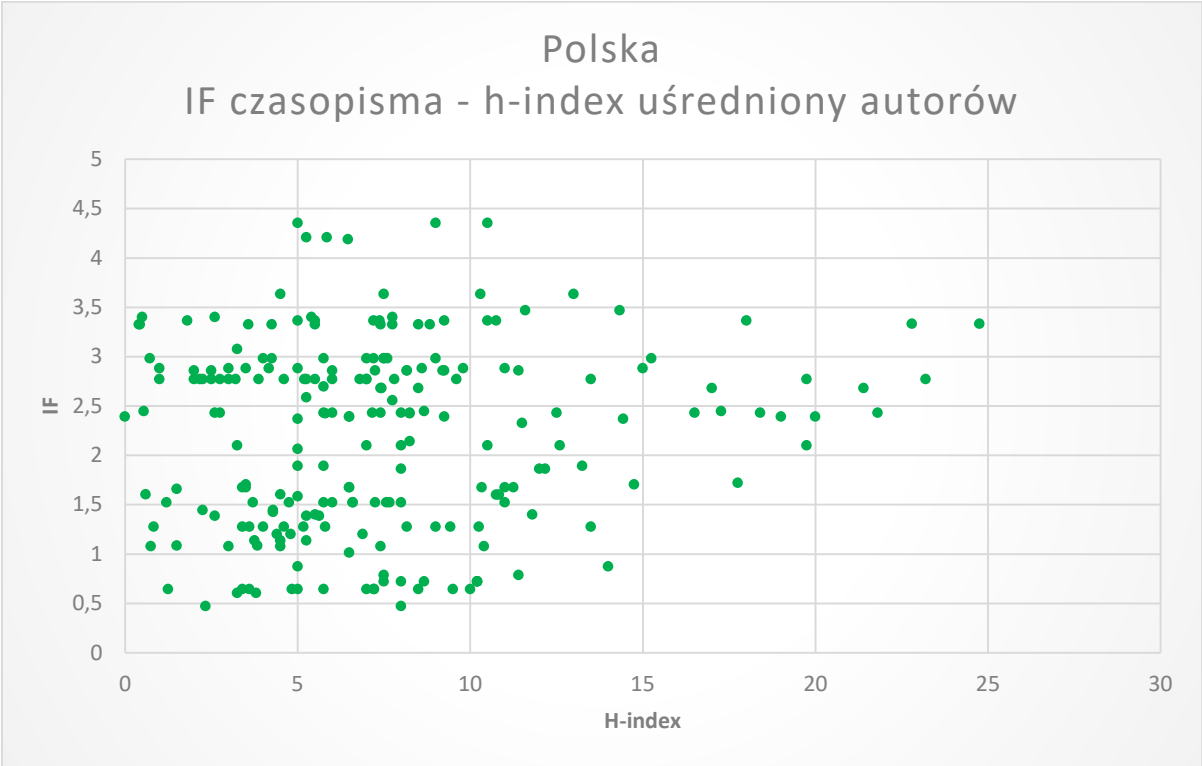
Rysunek 29 Wykres zależności maksymalnego h-indeksu autorów od SJR czasopisma publikacji przez polskich autorów



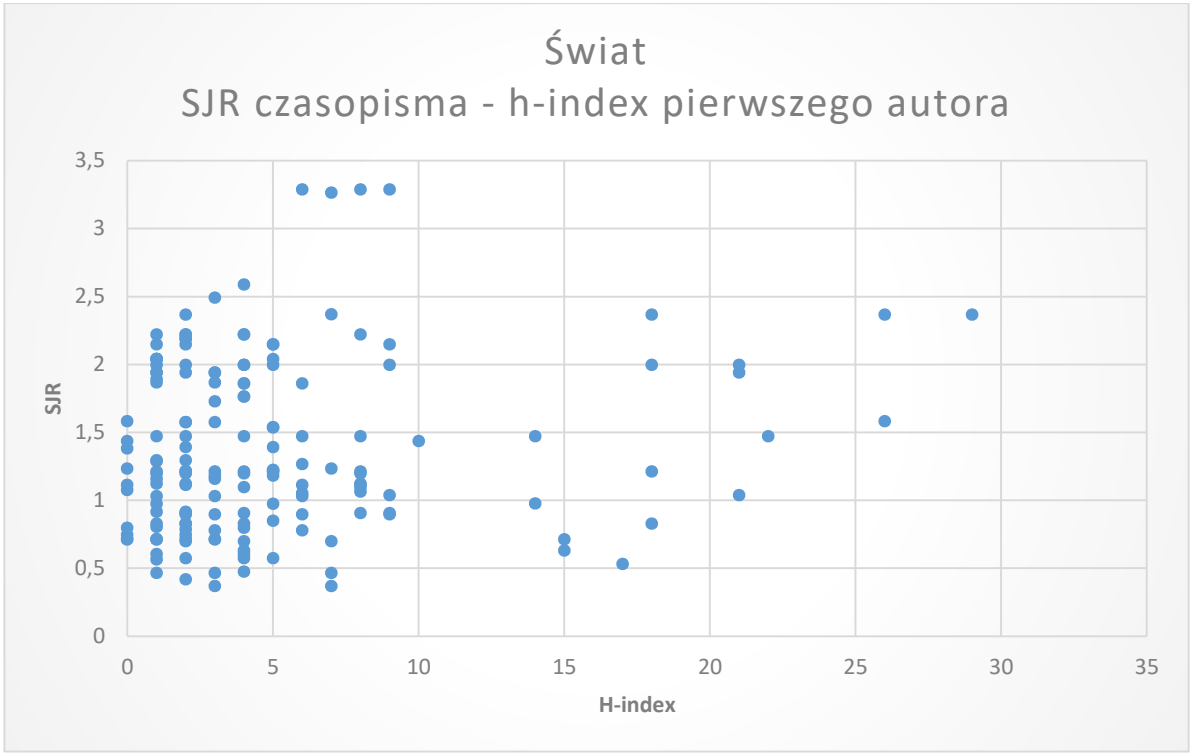
Rysunek 30 Wykres zależności maksymalnego h-indeksu autorów od IF czasopisma publikacji przez polskich autorów



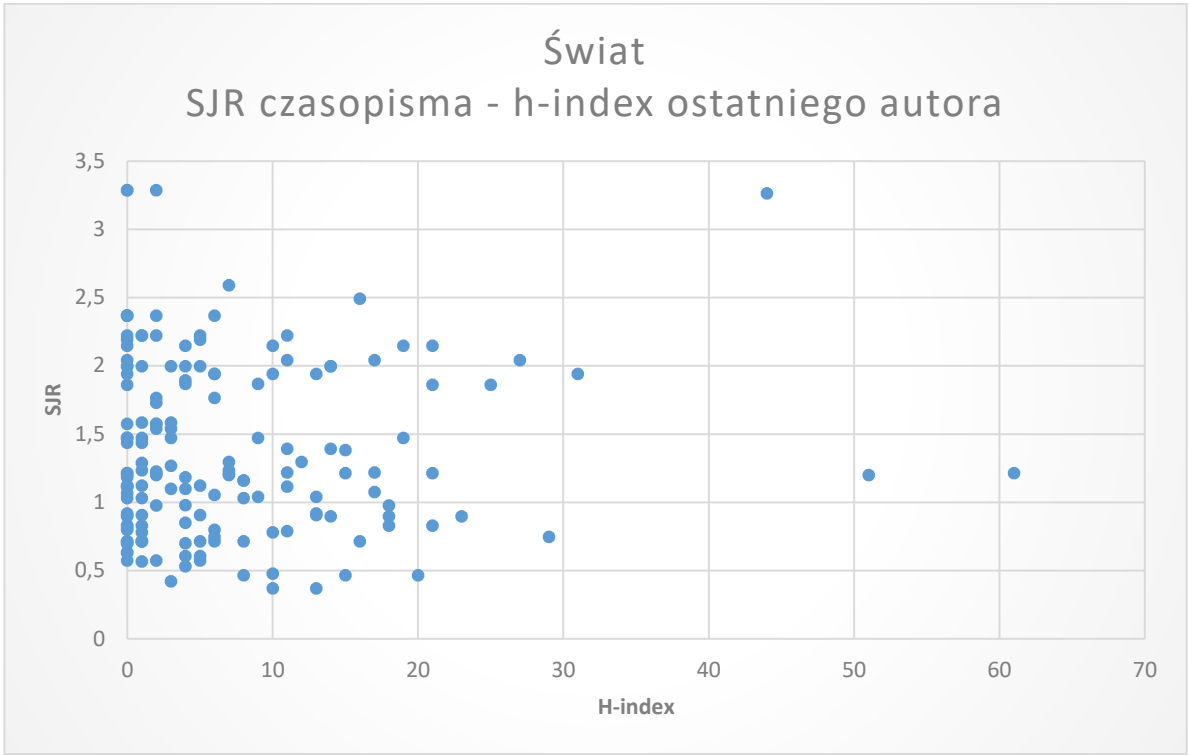
Rysunek 31 Wykres zależności uśrednionego h-indeksu autorów od SJR czasopisma publikacji przez polskich autorów



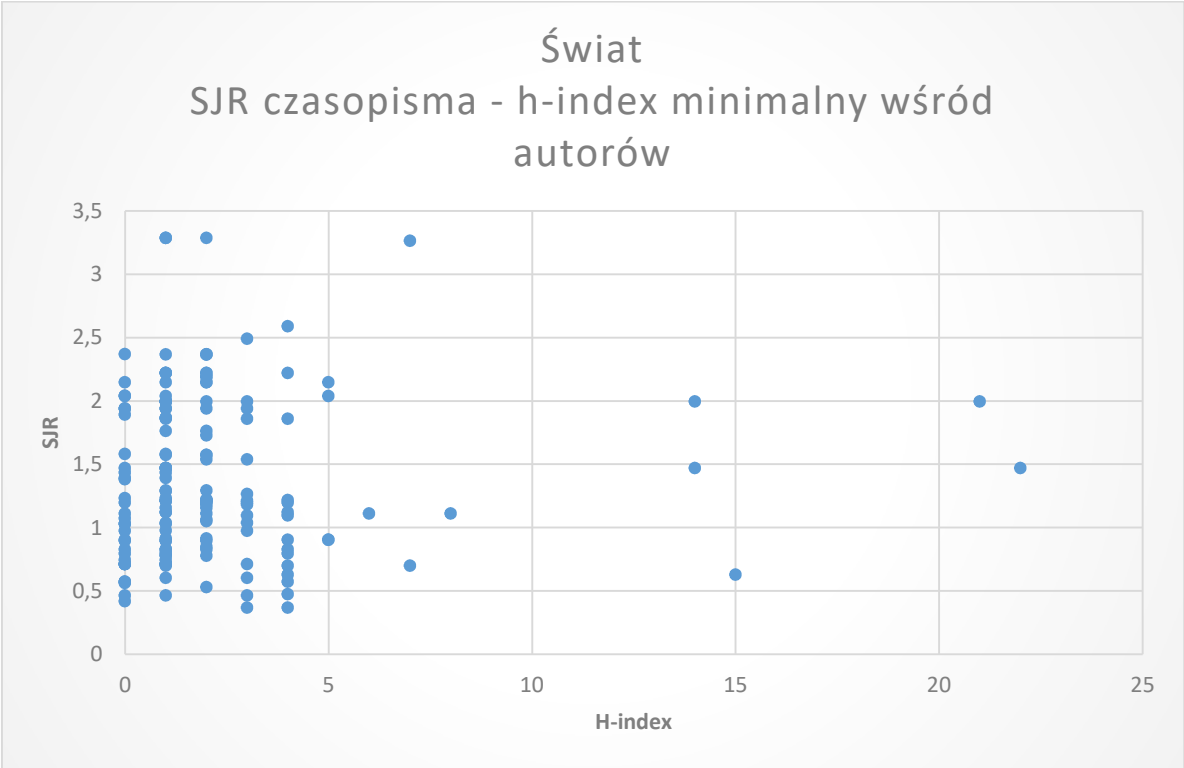
Rysunek 32 Wykres zależności uśrednionego h-indeksu autorów od IF czasopisma publikacji przez polskich autorów



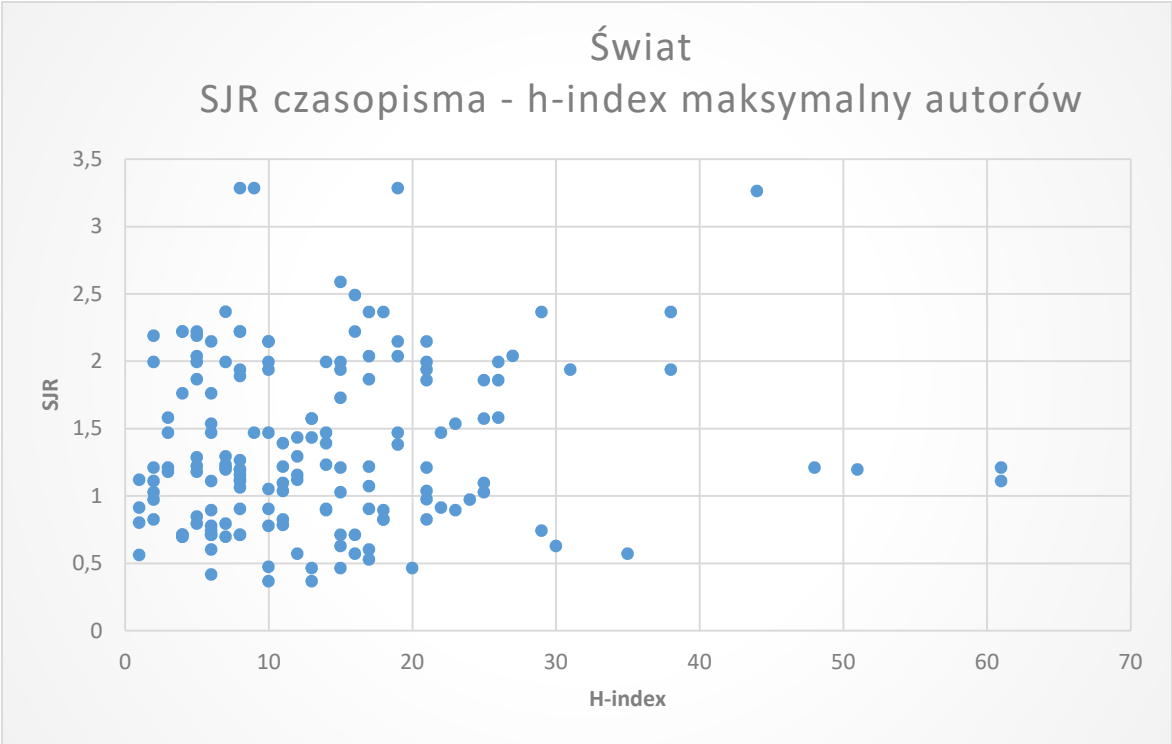
Rysunek 33 Wykres zależności h-indeksu pierwszego autora od SJR czasopisma publikacji



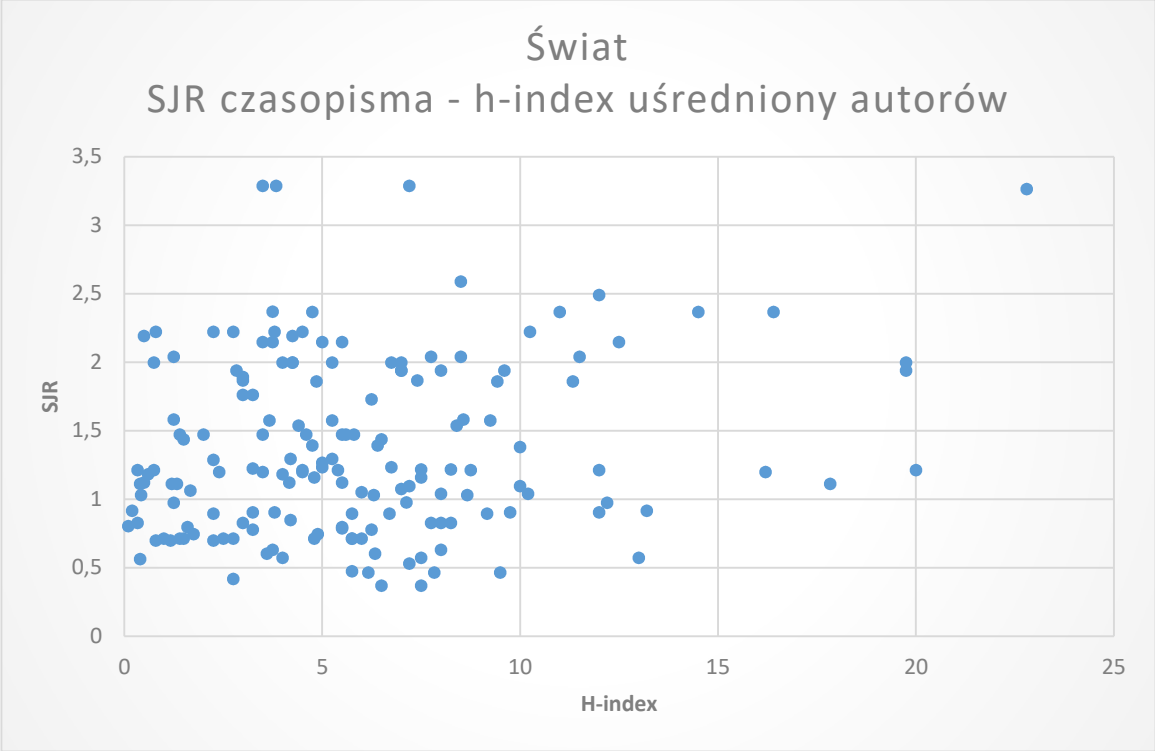
Rysunek 34 Wykres zależności h-indeksu ostatniego autora od SJR czasopisma publikacji



Rysunek 35 Wykres zależności minimalnego h-indeksu autorów od SJR czasopisma publikacji

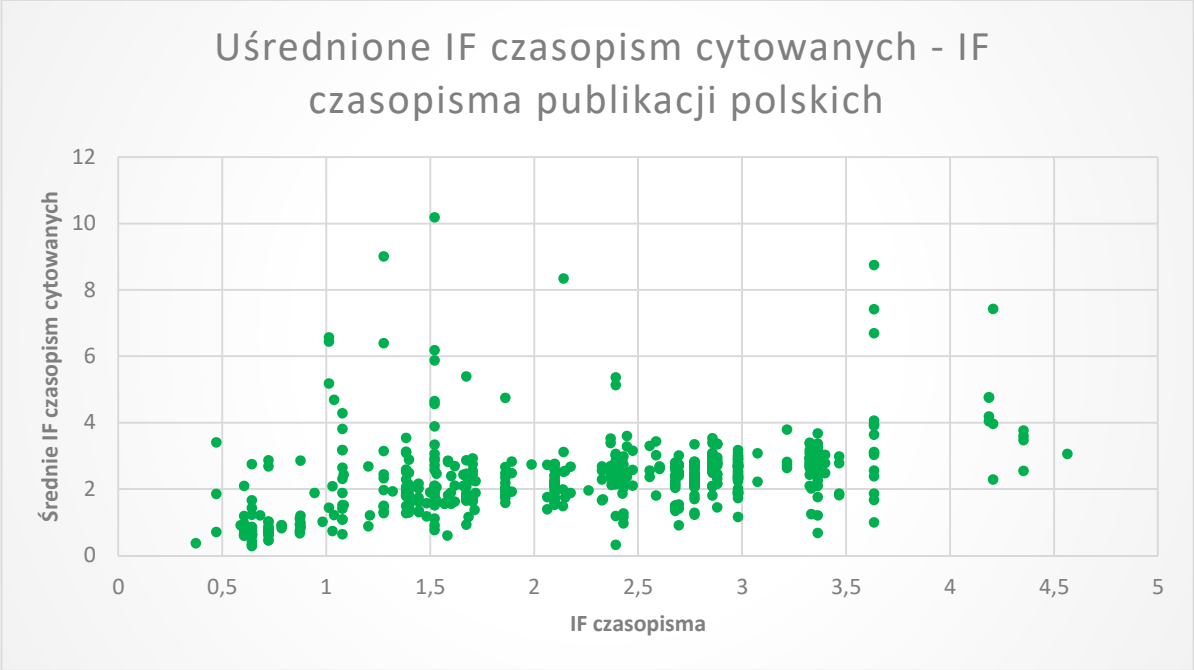


Rysunek 36 Wykres zależności maksymalnego h-indeksu autorów od SJR czasopisma publikacji

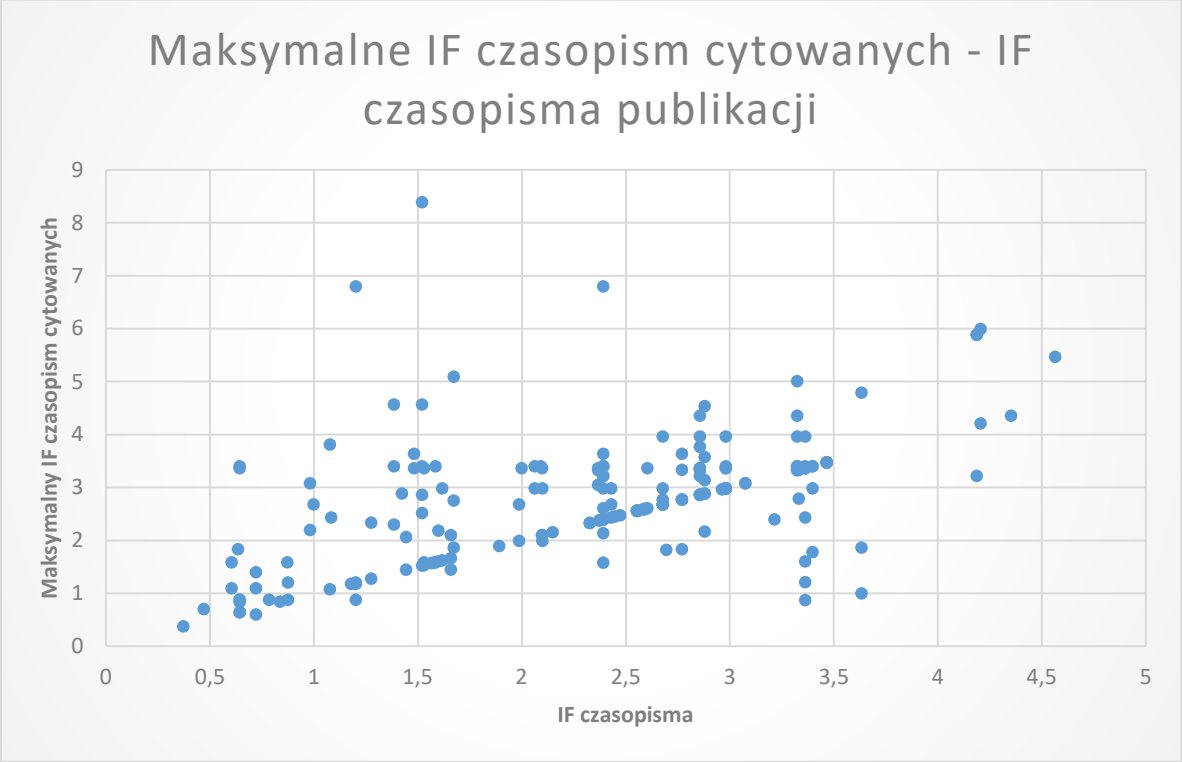


Rysunek 37 Wykres zależności h-indeksu średniego autorów od SJR czasopisma publikacji

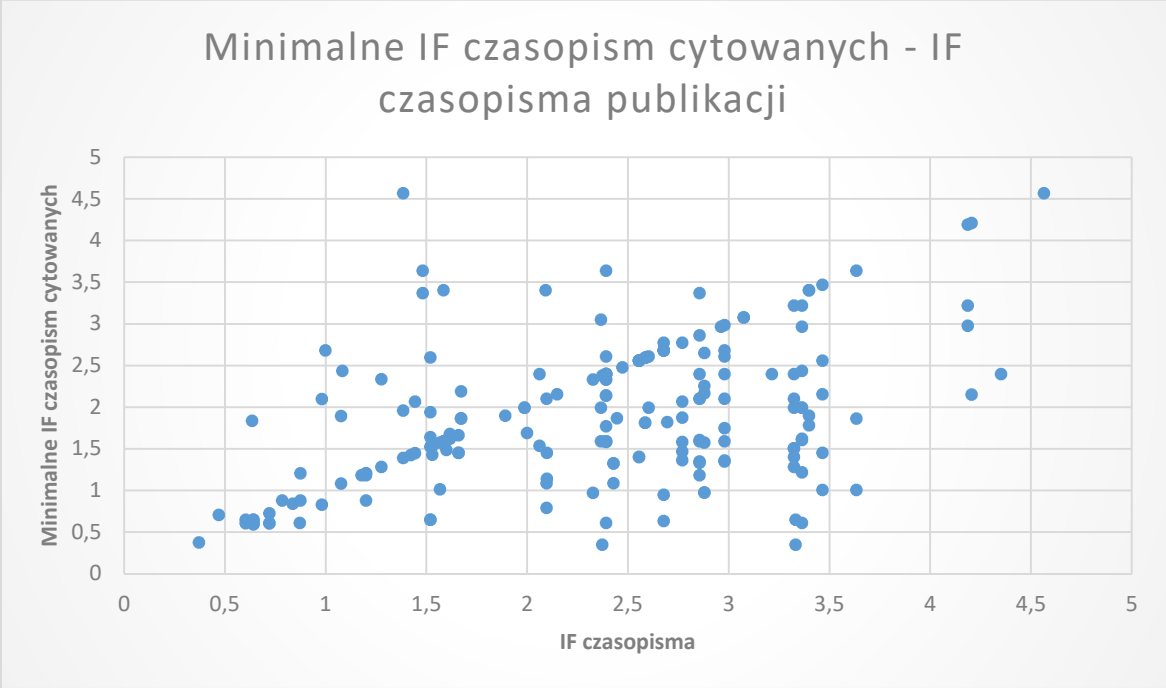
B.2 Statystyki cytowań artykułu



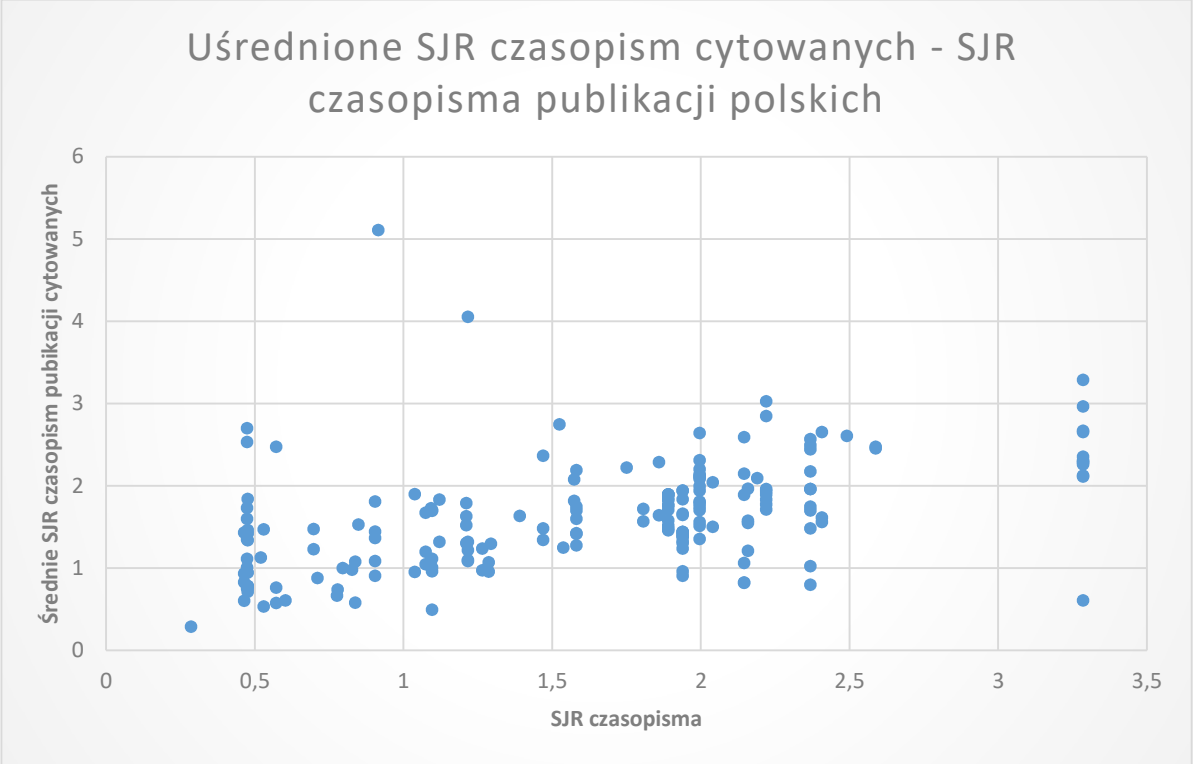
Rysunek 38 Wykres zależności uśrednionego IF czasopism cytowanych publikacji do czasopisma publikacji



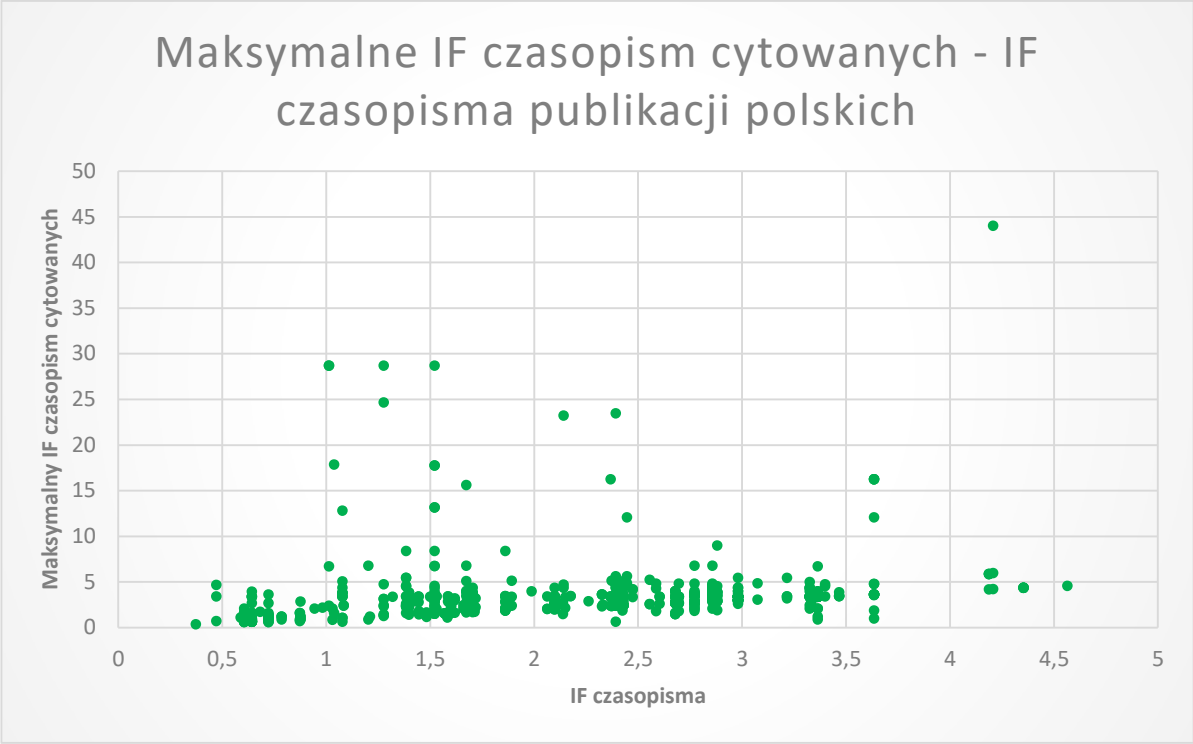
Rysunek 39 Wykres zależności maksymalnego IF czasopisma cytowanej publikacji do czasopisma publikacji



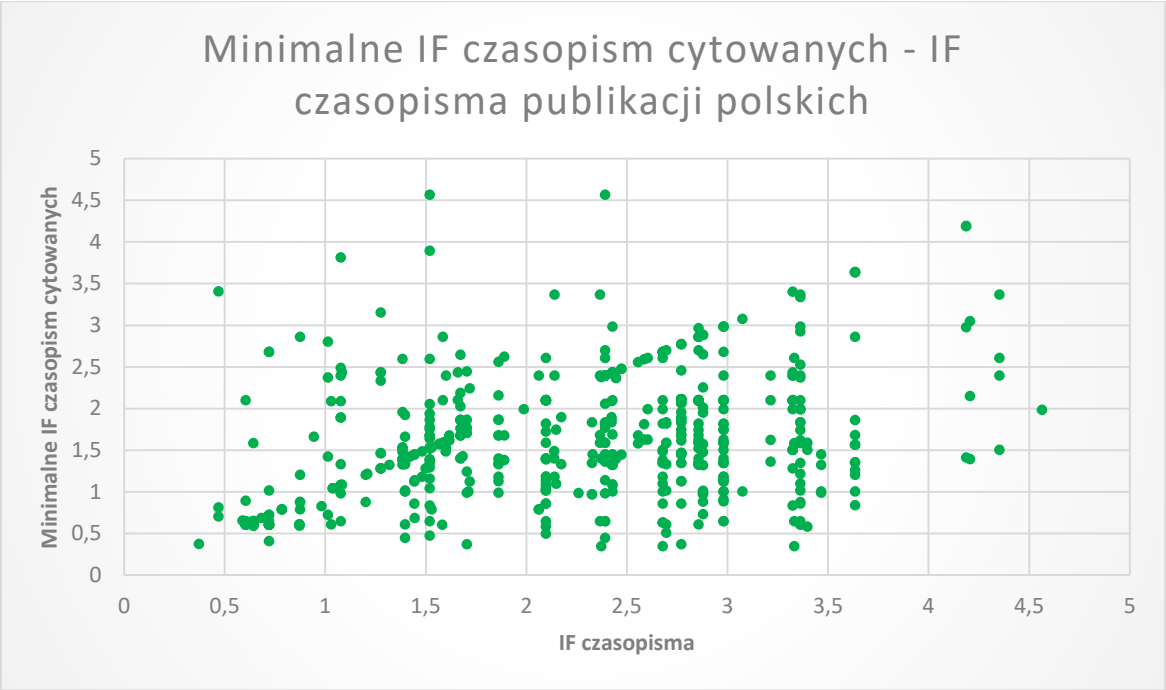
Rysunek 40 Wykres zależności minimalnego IF czasopisma cytowanej publikacji do czasopisma publikacji



Rysunek 41 Wykres zależności uśrednionego SJR czasopism cytowanych publikacji do czasopisma publikacji

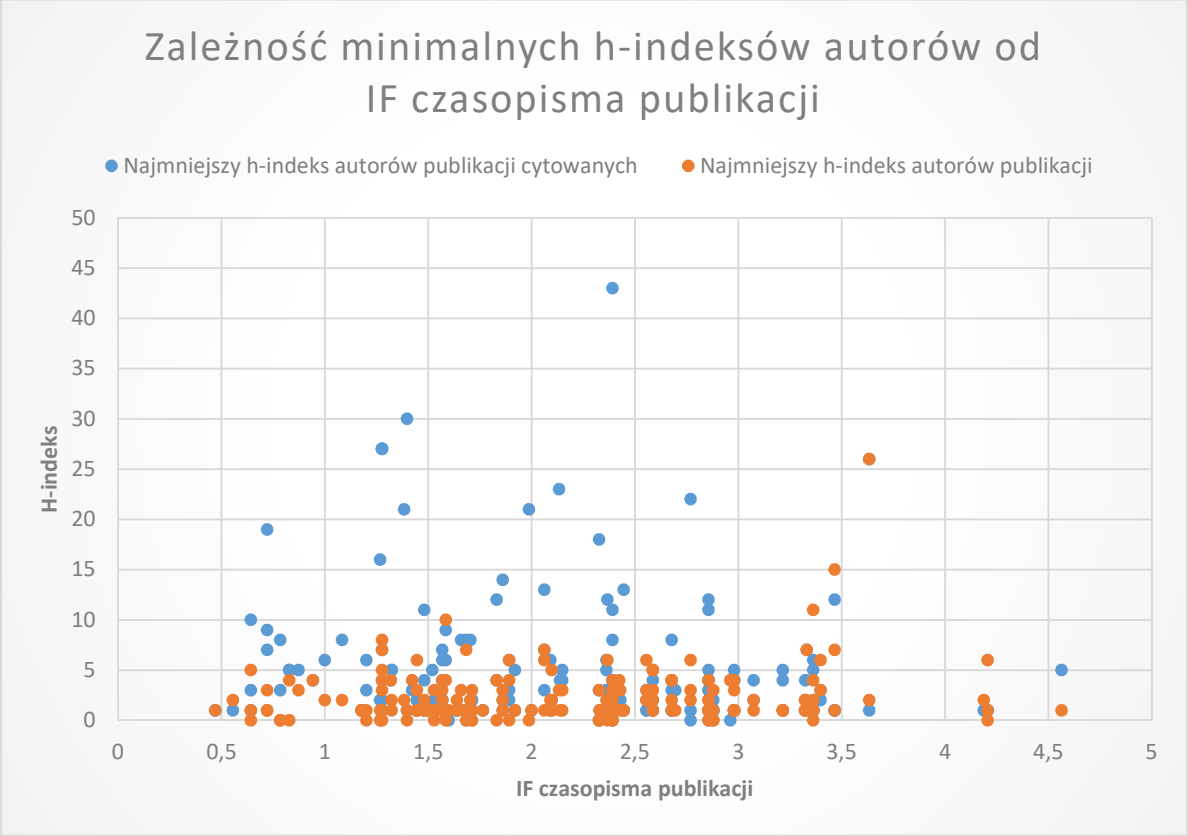


Rysunek 42 Wykres zależności maksymalnego IF czasopisma cytowanej publikacji do czasopisma publikacji

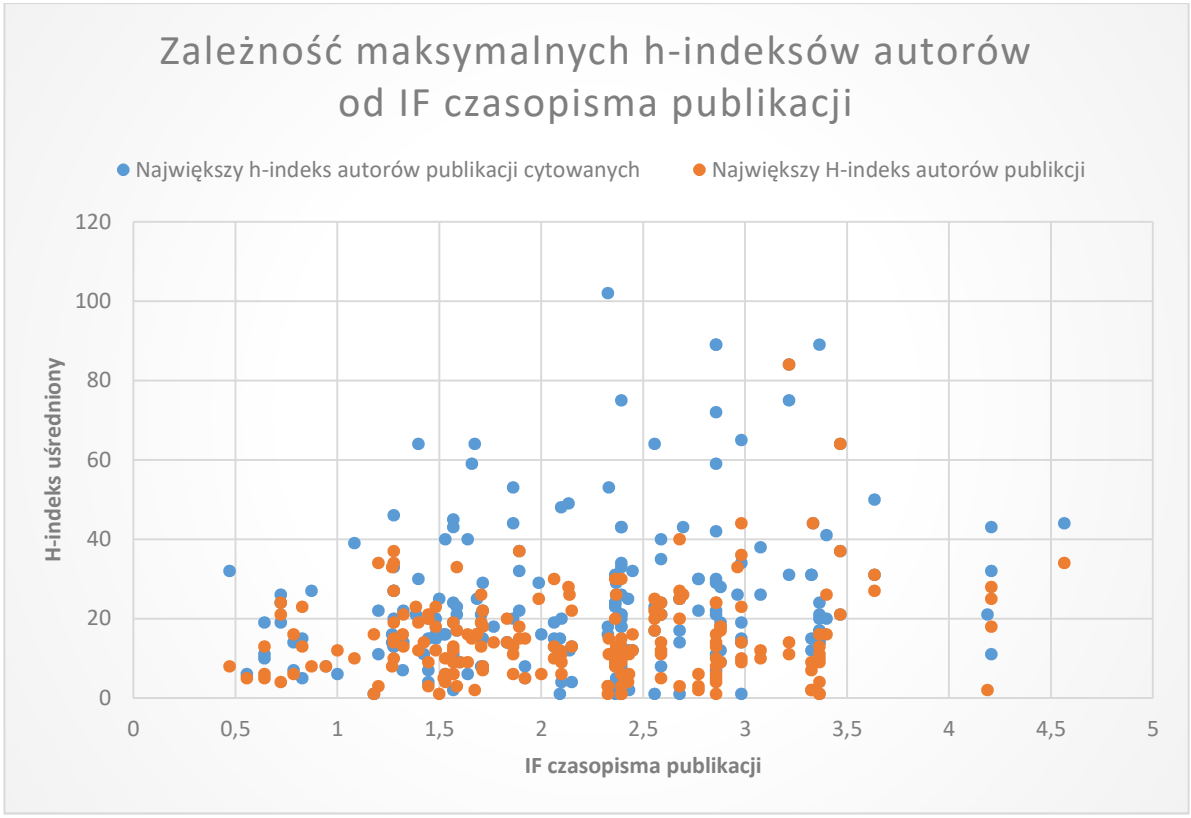


Rysunek 43 Wykres zależności minimalnego IF czasopisma cytowanej publikacji do czasopisma publikacji

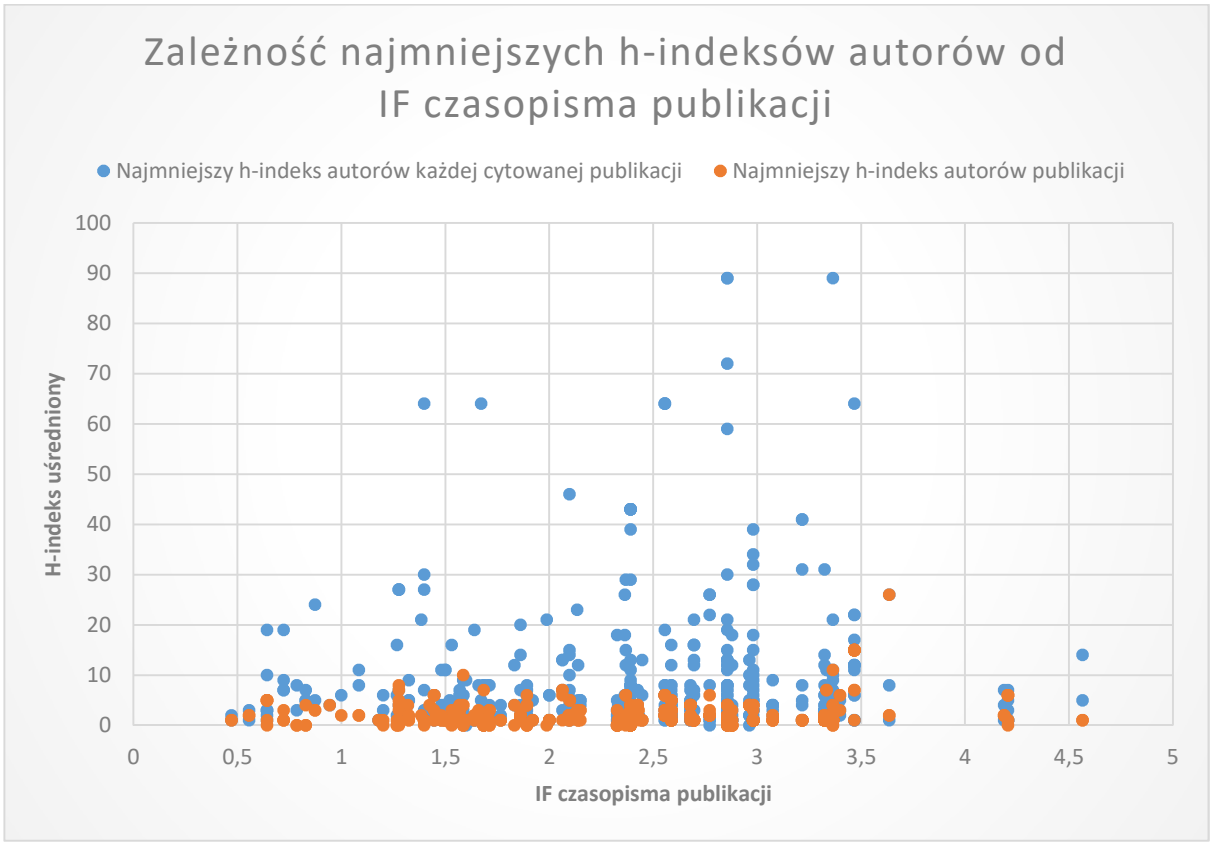
B.3 Statystyki autorów cytowanych publikacji



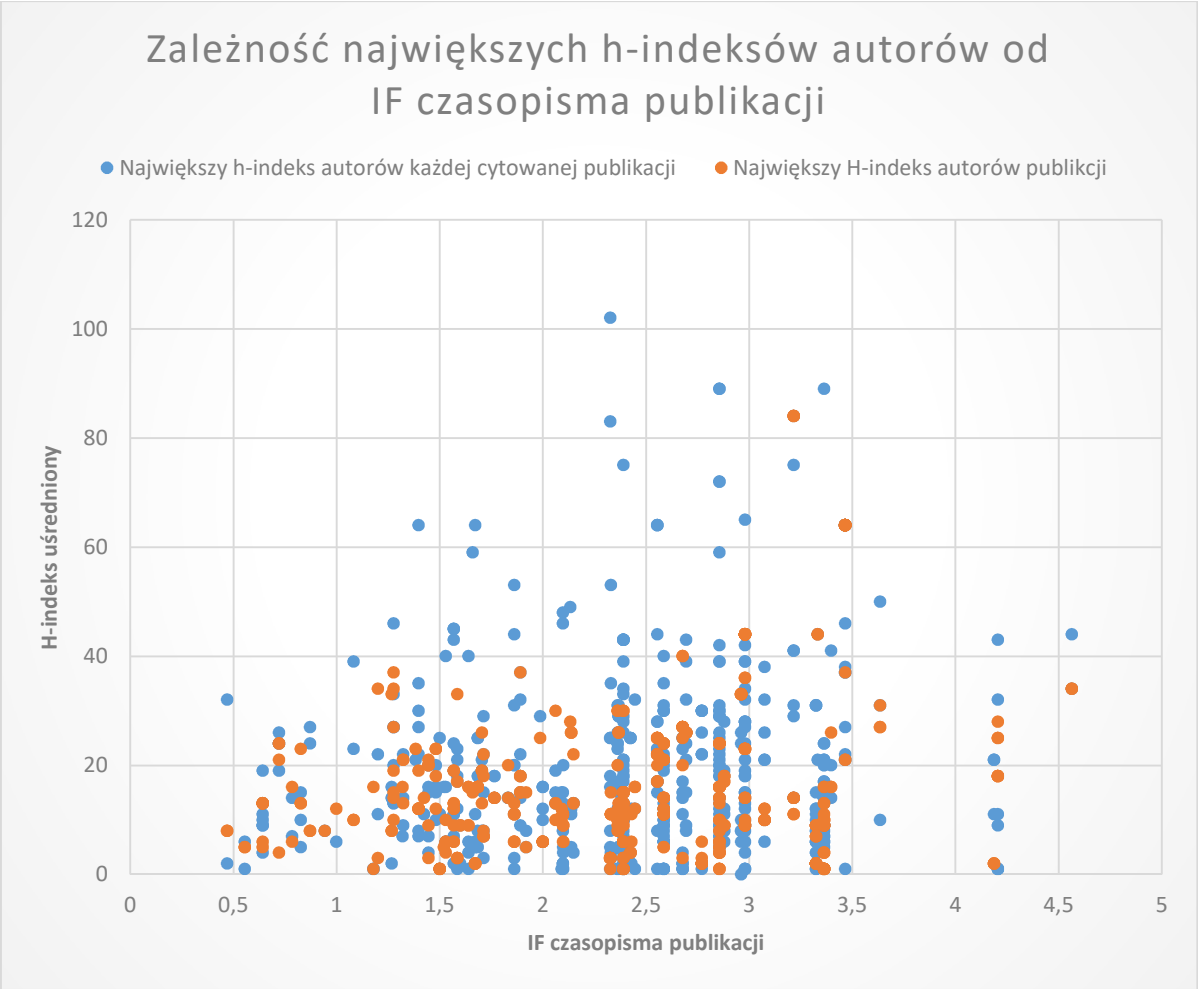
Rysunek 44 Wykres zależności minimalnych h-indeksów autorów od IF czasopisma publikacji



Rysunek 45 Wykres zależności maksymalnych h-indeksów autorów od IF czasopisma publikacji

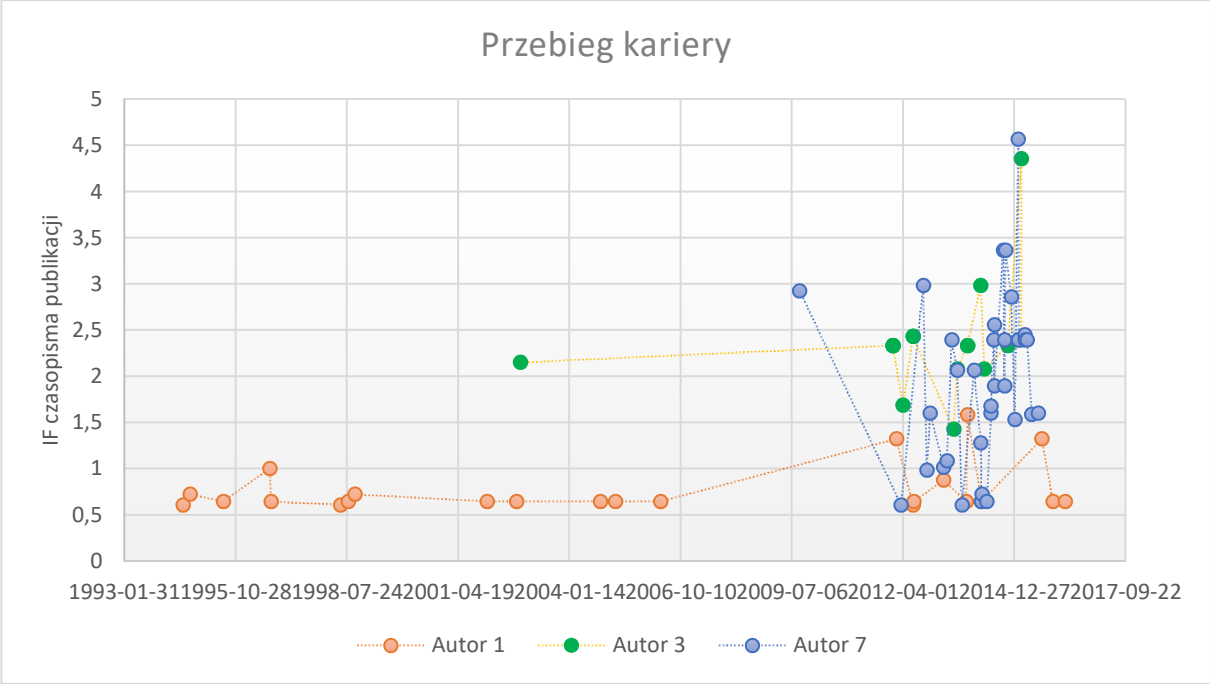


Rysunek 46 Wykres zależności najmniejszych h-indeksów autorów od IF czasopisma publikacji

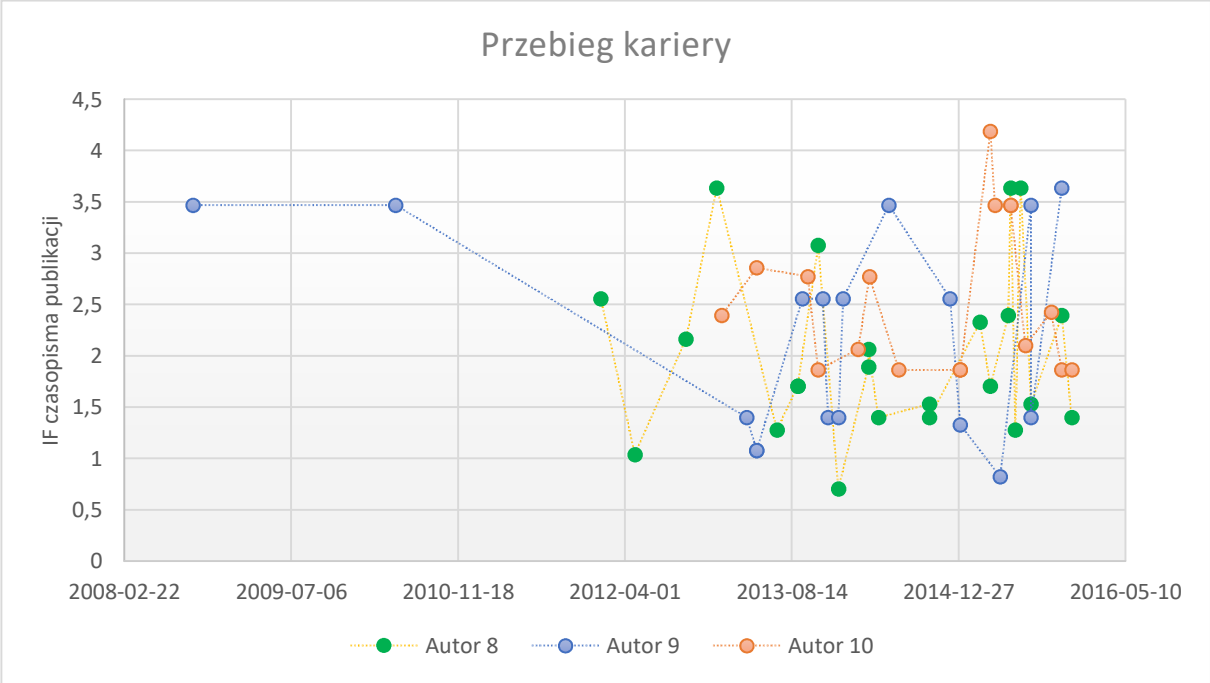


Rysunek 47 Wykres zależności największych h-indeksów autorów od IF czasopisma publikacji

B.4 Przebieg kariery



Rysunek 48 Przebieg kariery autorów 1,3,7



Rysunek 49 Przebieg kariery autorów 8,9,10