様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月３１日現在

研究種目：基盤研究（Ｂ）
研究期間：2006-2008
課題番号：18300080
研究課題名（和文）　論文の引用を支配する要因に関する統計学的研究

研究課題名（英文）　A statistical study on the factors affecting citations received by scientific articles.

研究代表者
小野寺　夏生（ONODERA NATSUO）
筑波大学・大学院図書館情報メディア研究科・教授
研究者番号：00302430

研究成果の概要：
学術雑誌論文の被引用数はその論文の影響度を示す指標とされるが、論文の質や内容以外の多くの要因に影響される。6 つの分野から抽出した 1,395 の論文を用いて、どのような要因がどの程度の影響力を持つかを重回帰分析によって検討した。その結果、どの分野でも、参考文献の数、特にその中の近年の文献に対する比率が高い論文ほど、高い被引用数を得る傾向があることを見出した。これは、研究評価に被引用数を利用する際の重要な知見となる。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	6,400,000	1,920,000	8,320,000
２００７年度	4,300,000	1,290,000	5,590,000
２００８年度	2,100,000	630,000	2,730,000
年度			
年度			
総　計	12,800,000	3,840,000	16,640,000

研究分野：計量書誌学
科研費の分科・細目：情報図書館学
キーワード：引用分析、多変量解析、要因分析、計量書誌学、論文、学術雑誌、研究評価

１．研究開始当初の背景

研究評価を行う際の参考データとして、論文の被引用数がよく用いられる。しかし、被引用数は、分野、論文の種類、使用言語等に影響され、更に、同じ雑誌に同時期に発表された同種の論文の間でも著しい差があることが知られている。論文の被引用数とそれに影響する要因の関係を調べた研究は多いが、それらは次の意味で不十分である：(1)各要因を個別に扱っているため他の要因との交絡効果が含まれている可能性がある、(2)総合的な要因分析を行っている少数の研究は特定の分野のみを扱っている。そこで、いくつかの分野から計画的に採取した複数の標本を用いて、どの要因がどの程度被引用数に寄与するかを統計的に検討したいと考えた。

２．研究の目的

同一雑誌、同一時期に発表された原著論文の被引用数に影響を与える要因（論文の内容や品質以外の要素）を明らかにするため、重回帰分析を用いて、これら要因の影響をそれぞれ分離し、各要因の寄与を検討する。これをいくつかの分野について行うことにより、各分野における論文被引用数に対するベースラインを与えることを目標とする。

３．研究の方法

(1) 対象とする論文

物性物理学、無機・核化学、電気・電子工

学、生化学・分子生物学、生理学、消化器病学の 6 分野から各 4 誌（すべて英文誌）を選び、それぞれから 2000 年発表の原著論文 50-60 件（計 1395 件）を無作為抽出して、調査対象論文とした。

(2) 検討する要因（説明変数）と目的変数

目的変数は、(1)で示した標本論文が 2006 年 9 月までに得た被引用数である。自己引用を含んだ場合(TC_Total)と除いた場合(TC_NotSelf)の 2 通りを設定した。

論文の被引用数を説明する要因として次の 12 変数を選んだ : (a)著者数(Authors)、(b)著者所属機関数(Insts)、(c)著者所属国数(Countries)、(d)参考文献数(Refs)、(e)プライス指数（参考文献中最近 5 年間に発表されたものの百分率)(Price)、(f)図の数(Figs)、(g)表の数(Tables)、(h)数式の数(Eqs)、(i)論文長(Length)、(j)著者の過去論文数(Publ)、(k)著者が過去に得た被引用数(Impact)、(l)著者の過去活動期間(Age)。j-l の 3 つの説明変数は各論文の第一著者に対して求めたが、24 誌中 6 誌(各分野 1 誌)については、更に論文中最高位の論文数、被引用数、活動期間を持つ著者の値も採った。

(3) データの取得

Thomson Scientific から購入した調査対象論文のデータから、(2)で述べた変数データのうち、被引用数と説明変数 a-e 及び i のデータを得た。いくつかの変数については変数値取得のための加工を行った。説明変数 f-h は原論文に当たって取得した。説明変数 j-l を取得するため Web of Science で著者名検索を行い、得られた回答から同名の別著者の論文を除去するための判別式を作成して、該当の論文を抽出した。このデータから説明変数 j, l を定め、更に、Thomson Scientific からそれらの引用データを購入して説明変数 k を定めた。

(4) 統計解析

6 つの分野別及び 24 の雑誌別に線形重回帰分析と負の 2 項重回帰分析を行った(ステップワイズ法による変数選択)。被引用数分布は極めて歪度が大きいので、線形重回帰の目的変数は(被引用数+1)の対数（log(TC_Total+1) または log(TC_NotSelf+1)）とした。統計処理には Excel、SPSS、及び R を用いた。

４．研究成果

先述のように、被引用数は自己引用を含めた場合（TC_Total）と除いた場合（TC_NotSelf）を検討したが、全ての結果において、本研究の目的に関する限りどちらを用いても大きな違いはなかった。従って以下では前者を用いた結果を示し、変数名も単に TC と記す。

（1）予備的な分析

①変数の分布形：被引用数 TC は極めて歪んだ分布を示し、分野別により歪度 1.56～3.27 であった。対数変換後の log（TC+1）の歪度は−0.47～+0.43 の範囲に収まり、ほぼ正規型の分布になった。目的変数もほとんどが正の歪度を示したが、TC ほど極端ではなかったので対数変換は行わなかった。

②相関分析：log（TC+1）との相関が高い説明変数は Refs、Price、Length であり、これらは概ねどの分野でも相関係数 r が 0.3 から 0.6 の範囲であった。説明変数同士では、Authors-Insts-Countries の 3 変数の間、及び Refs-Figs-Length の 3 変数の間に多くの分野で相関が見られた。以上のことから、被引用数に対する有効な説明変数として、まず Price が候補となる。Refs-Figs-Length のグループと Authors-Insts-Countries のグループでは、重回帰分析で選択される変数が減ると予想される。

（2）重回帰分析の結果

次の 2 通りの標本に対して線形重回帰分析（LR）と負の 2 項重回帰分析（NBR）を行った。

（A）分野別に 6 つの標本（n=227～240）を作る。論文が掲載された雑誌をダミー変数とすることにより、異なる雑誌の影響を考慮する。変数 Publ、Impact、Age には第一著者の値を使用する。

（B）雑誌別に 24 の標本（n=51～60）を作る。6 誌（各分野 1 誌）の変数 Publ、Impact、Age は共著者中の最大値を用い、他の 18 誌は第一著者の値を使用する。

その結果、（A）の方が有意な変数の選択に関して系統的な傾向が得られた。（B）の標本サイズが十分でなかったと考えられる。以下では（A）についての結果を述べることとし、6 つの分野それぞれに対する LR と NBR の結果の概要を表 1 に示す。

①モデルの当てはまりの良さ：表 1 に、自由

度調整済み決定係数（LR では R_c^2、NBR では $R_{DEV,c}^2$）と相対残差平方和（Rel SE）を示した。これについては(3)で述べる。

表 1　LR と NBR の結果

分野	物性物理	無機化学	電気工学	生化学	生理学	消化器
n	230	227	229	240	236	233
		LR の結果概要				
R_c^2	0.262	0.320	0.271	0.475	0.607	0.497
Rel SE	468.5	332.0	1068.6	172.8	251.4	385.3
Authors	△			△		△
Insts	●	○				
Countries		▲				△
Refs	◎		○	○	○	△
Price	◎	◎	○	◎	◎	◎
Figures			○		◎	△
Tables		▲				◎
Eqs		●				
Length		◎		○		▲
Publ			○			
Impact	○	▲				
Age						
1A	◎	△	◎	◎	◎	◎
1B	◎		◎	◎	◎	◎
1C	○		△		◎	◎
		NBR の結果概要				
$R_{DEV,c}^2$	0.277	0.314	0.279	0.451	0.541	0.486
Rel SE	181.2	142.1	290.8	104.6	122.4	161.5
Authors	○			○		
Insts	●					
Countries						
Refs	◎	△	◎	○	○	
Price	◎	○	◎	◎	◎	◎
Figures		△			◎	
Tables		▲				◎
Eqs		●				
Length		○		○		
Publ		●				
Impact	○					
Age		△			▲	
1A	◎	○	◎	◎	◎	◎
1B	◎		◎	◎	◎	◎
1C	△				◎	◎

偏回帰係数が正　△:10%有意、　○:5%有意、　◎:0.1%有意
偏回帰係数が負　▲:10%有意、　●:5%有意

②有効な説明要因：表 1 に、偏回帰係数が有意とされた変数を示す。まず Price が全分野で LR、NBR ともに選択された。次いで Refs が 4 つの分野で共通に選択され、他の 2 分野でも LR、NBR のいずれかで選択された。この 2 つが被引用数に対する最も説明力の高い要因であることが実証された。この 2 要因に比べるとやや弱いが、Authors、Length、Figs も、多少の説明力がある。これらは全て偏回帰係数の符号が正である（被引用数に対し正の効果を持つ）。他の変数は、選択された分野や回帰法が限られたり、分野により符号が逆であったりで、本研究では有意な説明要因とは認められなかった。やや意外であったのは、被引用論文著者の過去業績に関する変数（Publ、Impact、Age）が有効な要因とされなかったことである。

③実測値と予測値の一致の程度：例として、無機・核化学分野の NBR における両者の関係を図 1 に示した。相関係数は 0.60 で、6 分野の中で中程度の一致である。

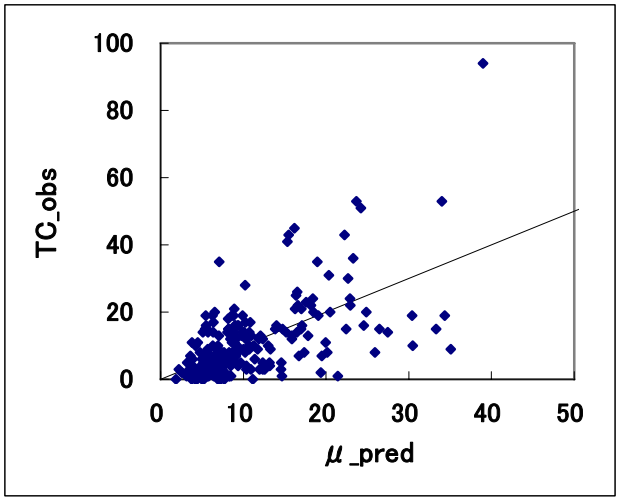

図 1　NB 回帰による被引用数予測
（無機・核化学分野の例）

(3) LR と NBR の当てはまりの良さの比較

表 1 には、当てはまりの良さについて 2 つの指標を示した。しかし、NBR に対する自由度調整済み決定係数（デビアンスから導かれる $R_{DEV,c}^2$）は、LR でよく用いられる R_c^2 とは定義が異なる。そこで、LR と NBR の適合性を比較するため、次式で定義される相対残差平方和（Rel SE）を用いた。

$$\text{REl SE} = \Sigma_i [(y_i - \mu_i)/\mu_i]^2$$

y_i、μ_i はそれぞれ、目的変数の観測値と予測値（期待値）である。NBR においても $\log(y)$ を目的変数とする LR においても、残差は μ_i に比例して大きくなりやすいため、より合理的に比較できるよう、μ_i で規格化して残差の均等化を図った。表 1 の通り、どの分野においても、NBR の方が LR よりも当てはまりが

よいことが示された。被引用数ごとの論文数の観測値と予測値の分布を調べたところ、これについてもNBRの方がLRより実際の分布をよく再現し、特に低被引用数における当てはまりがよいことが判った。

(4) 被引用論文著者の過去検索論文から同名異著者の論文の除去について

説明変数 Publ、Impact、Age のデータを得るには、対象の論文（ソース論文）の著者の過去発表論文を同定する必要がある。このために、1395 の対象論文の 2595 著者（6 誌については全著者、他の 18 誌については第一著者）について著者名検索を行ったところ、62.9 万の文献が検索された。この中には、検索対象著者の論文（真論文）の他、同名同イニシアルの異著者の論文（偽論文）が大量に混入している。以下に、偽論文を半自動的に除去するアルゴリズムの手順を述べる。本研究ではこの部分に最も時間を要した。

ソース論文との間に名の一致する著者を 2 名以上含む論文（複数著者一致論文）は真論文の可能性が高いと言えるので、それらとそれ以外の論文を区別した。

①複数著者一致論文以外の検索論文（59.0 万件）の処理

(i) ソース論文と著者アドレスがほとんど一致しない論文、及びソース論文雑誌と検索論文雑誌の間に引用関係がない論文を偽論文と見なして除去した。ここでは多少真論文を逃がすことは犠牲にして、候補論文を絞り込むことに重点を置いた。

(ii) (i)を通過した 6.7 万論文から 2400 論文を抽出し、それぞれのサンプル検索論文の真偽を各 2 名の判定者が判定した。この結果に基づき、サンプル検索論文が偽論文である確率 p を予測するロジスティック重回帰モデルを分野別に導出した。説明変数は次の 4 つとした：(a)所属機関アドレス類似度、(b)雑誌間の引用関係強度、(c)検索論文とソース論文の間の経過期間、(d)著者所属国によるダミー変数（ソース著者が日中韓台の場合 1）。サンプル論文を訓練群と検証群に分け、どの分野も 90%の正解率で真偽判定を行えることを実証した。

(iii) (ii)で得た回帰モデルを一次フィルタリングを通過した全論文に適用し、5.2 万件が真論文と判別された。

②複数著者一致論文論文（3.9 万件）の処理

ソース論文によって状況が大きく異なっていたので、$p \geqq 0.5$ の検索論文を 20 以上含むソース著者についてのみ、それらの検索論文を個別に見て真偽を判定した。それ以外のソース著者ではすべてを真論文と見なした。その結果 3.8 万件が真論文と判別された。

結論として、ここで提案した判別法は、主題分野や所属機関・所属国が広範にわたる大量の文献から一定の精度で「真論文」を選別したいという場合には、有効であると考えられる。しかし、十分に偽論文除去ができなかった少数のソース著者が残存している。

(5) まとめ－得られた成果の位置付け、効果

同一分野内の論文の被引用数の違いを、2～4 個の選択説明変数により 25～60%程度まで説明することができた。今回用いた説明変数はすべて計量書誌学的な量であることを考えれば、この程度の説明力に留まるのは当然と言える。これらの要因で説明できない誤差の部分が、その論文の内容や品質に関する重要な情報を含む可能性があり、今後その検討が重要である。

LR モデルと NBR モデルの比較では、後者の方が観測値に対する当てはまりがよいことが示された。LR の相対残差平方和を標本サイズ n で割ると概ね 0.5 なので、信頼限界をその倍とすると、μ_i を倍半分以内で予測でき

る。回帰前の被引用数の分散が非常に大きいことを考えれば、予測幅をかなり縮めることができたと言える。

選択された説明変数は、分野を越えてある程度の共通性があり、標準的論文の被引用数に対して一貫性のある予測モデルが得られる可能性を示した。被引用数を論文評価に用いるとき、このようなモデルはそのベースラインとなり得る。どの分野でも Price が最も説明力の高い変数であることが確認されたが、プライス指数の有効性は本研究で初めて見出されたことである。Refs、Length、Figs は、互いにやや強い相関を持ちながらも、両者が有効な説明変数になる場合が多かった。一方、著者の過去業績については、その有効性を明確に示すことはできなかった。これは過去の研究の結果とは異なるが、同名異著者の論文が除去しきれなかったことの影響かもしれない（(4)を参照）。

分野ごとの回帰の説明力が雑誌ごとのそれに比べ高かったことには、標本サイズの問題の他、異質の雑誌の混合により見かけの相関が生じた可能性があるが、次の 2 つのことからその可能性は低いと考えられる。

①どの変数をとっても、分野内の分散と雑誌内の分散に大きな差はない。

②分野内と雑誌内で、目的変数と主要な説明変数の間の相関係数に系統的な差はない。

従って、雑誌の標本サイズを大きくすれば、より明確な回帰結果が得られる可能性がある。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計４件）

①芳鐘冬樹，辻慶太，小野寺夏生．論文引用に影響を与える要因－負の二項重回帰による検討－．2009 年日本図書館情報学会春季研究集会発表要項．飯能，2009-5-23. p.63-66.
②小野寺夏生，芳鐘冬樹，岩澤まり子，辻慶太，緑川信之ほか．著者検索で得られた大量の論文から同名異人著者を除去する方法．2009 年日本図書館情報学会春季研究集会発表要項．飯能，2009-5-23. p.51-54.
③小野寺夏生，山崎茂明，芳鐘冬樹，岩澤まり子，辻慶太，緑川信之ほか．論文の被引用数に影響する要因に関する統計学的研究．第 56 回日本図書館情報学会研究大会発表要綱．奈良，2008-11-15/16, p.41-44.
④小野寺夏生，岩澤まり子ほか．自然科学分野の雑誌における国別被引用率の比較．第 54 回日本図書館情報学会研究大会発表要綱．北九州，2006-10-21/22, p.21-24.

６．研究組織

(1)研究代表者
小野寺　夏生（ONODERA NATSUO）
筑波大学・大学院図書館情報メディア研究科・教授
研究者番号：00302430
(2)研究分担者
緑川　信之（MIDORIKAWA NOBUYUKI）
筑波大学・大学院図書館情報メディア研究科・教授
研究者番号：70166073
岩澤　まり子（IWASAWA MARIKO）
筑波大学・大学院図書館情報メディア研究科・教授
研究者番号：20292568
辻　慶太（TUJI KEITA）
筑波大学・大学院図書館情報メディア研究科・准教授
研究者番号：30333545
(3)連携研究者
山崎　茂明（YAMAZAKI SHIGEAKI）
愛知淑徳大学・文学部・教授
研究者番号：40246450
芳鐘　冬樹（YOSHIKANE FUYUKI）
大学評価・学位授与機構・助教
研究者番号：30353428