論文の被引用数に影響する要因に関する統計学的研究

小野寺夏生 onodera@slis.tsukuba.ac.jp[*1]　山崎　茂明 shige@asu.aasa.ac.jp[*2]
芳鐘　冬樹 fuyuki@niad.ac.jp[*3]　岩澤まり子 miwasawa@slis.tsukuba.ac.jp[*1]
辻　慶太 keita@slis.tsukuba.ac.jp[*1]　緑川　信之 midorika@slis.tsukuba.ac.jp[*1]
天野　晃 kamano@affrc.go.jp[*1]　大谷　裕 y-ootani@mnc.toho-u.ac.jp[*1]
城山　泰彦 kiyama@juntendo.ac.jp[*1]　児玉　閲 kodamat@ks.kyorin-u.ac.jp[*1]
角田　裕之 tsunoda@shokei-gakuen.ac.jp[*1]　山﨑　静香 yshizu@sc.itc.keio.ac.jp[*1]
*1 筑波大学大学院図書館情報メディア研究科
*2 愛知淑徳大学文学部　*3 大学評価・学位授与機構

抄録：　同一雑誌、同一年(2000)発表の原著論文の被引用数の変化が、論文の計量書誌的要素からどの程度説明できるか検討した。6 つの分野から各 4 誌を選び、それぞれから原著論文 50-60 件を無作為抽出した。9 種の説明変数を用いて、分野別及び雑誌別にステップワイズ法による重回帰分析を行った。多くの分野で、プライス指数、参考文献数、論文長が有意な説明変数として選択された。回帰の自由度調整済み決定係数は 0.1 ～ 0.5 であった。

１．研究の背景と目的

研究評価を行う際の参考データとして、論文の被引用数に基づく指標がよく用いられる。被引用数は論文の影響度を示す統計的測度としては適切であると考えられるし、今のところ、論文の利用度や影響度を相当の客観性をもって測り得る他の測度は存在しない。しかし、引用データを研究評価に用いるには、その特性をよく知り、十分な注意を払う必要がある [1)]。たとえば、論文が掲載された雑誌のインパクトファクターを論文の評価指標に転用する例がしばしば見られるが、雑誌中の論文の被引用数分布は極めて歪度が大きく、その平均値であるインパクトファクターは代表値として適切ではない [2)]。

更に、単純に論文間の被引用数を比較することには問題がある。論文の被引用数は、分野、掲載誌の発行国、論文の種類（原著論文、短報、レビュー等）、使用言語等、多くの要因に影響されるからである。また、引用を影響度の測度として取り扱うとすれば、自己引用の問題について注意しなければならない。

論文の被引用数といろいろな要因との関係を検討した研究が多数報告されている。しかし、その多くはある単一の要因に着目しているか、複数の要因をそれぞれ独立に見ているので、ある要因が被引用数と相関があっても、それが他の要因との交絡である可能性を否定できない。

重回帰分析は、個々の説明変数による効果を分離することができるので、考え得る多くの要因の中のどれが被引用数に対する説明力が高いかを推察することができる。この方法を用いた研究はいくつかあり [3-7)]、それぞれ興味深い結果が得られているが、1 つの分野だけを扱っている、考慮した変数の種類が限られている、論文がある関心の下に選定されている等により、結論の一般性に乏しい。

我々は、科学研究費補助金（基盤研究（Ｂ））により、6 つの分野から各 4 種の雑誌を選び、同一雑誌、同一時期に発表の原著論文の被引用数が、論文の内容以外の要素にどのように依存するかに注目した研究を進めている。重回帰分析を用いて、論文の被引用数に対する潜在的要因の影響をそれぞれ分離し、各要因の寄与を検討する。各分野における論文被引用数の標準的パターンの知識が得られれば、研究評価等への引用データの利用に、より精密なベースラインを与えることができると考えられる。

被引用数を説明する要因として、15 種類の説明変数を考えているが、今回は、このうち 9 変数（後述）を用いて、分野別及び雑誌別に重回帰分析を行った結果を報告する。

２．研究の対象と方法

２.１ 対象とする論文

物性物理学、無機・核化学、電気・電子工学、生化学及び分子生物学、生理学、胃腸系臨床医学の 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。以下ではこれらの分野をそれぞれ P, C, E, B, S, G で表す。これらの雑誌を**表 1** に示す。

表１　選択した分野と雑誌

分野	分野記号	Journal Title	出版者と出版国[*1]	論文数[*2]	インパクトファクター[*3]
物性物理学	P	Eur. Phys. J. B	Springer(独)	791	1.43
		J. Phys.-Condes. Matter	IOP Publishing Ltd(英)	2505	2.05
		Physica B	Elsevier Science(蘭)	2872	0.68
		Phys. Rev. B	Am Phys Soc(米)	10168	3.08
無機・核化学	C	Inorg. Chem.	Am Chem Soc(米)	2164	3.45
		Inorg. Chim. Acta	Elsevier Science(スイス)	1156	1.55
		J. Chem. Soc.-Dalton Trans.	Roy Soc Chem(英)	1412	2.93
		Transit. Met. Chem.	Springer(蘭)	329	0.86
電気・電子工学	E	IEE Proc.-Circuit Device Syst.	IEE(英)	135	0.50
		IEEE Trans. Circuits Syst. I-Fundam. Theor. Appl.	IEEE(米)	447	0.93
		IEEE Trans. Microw. Theory Tech.	IEEE(米)	711	1.54
		Signal Process.	Elsevier Science(蘭)	348	0.59
生化学・分子生物学	B	Eur. J. Biochem.	Blackwell Publishing(英)	1203	3.26
		J. Biochem. (Tokyo)	Jap Biochem Soc(日)	473	2.29
		J. Biol. Chem.	Am Soc Biochem(米)	12959	6.36
		J. Mol. Biol.	Elsevier Science(米)	1747	5.54
生理学	S	Jpn. J. Physiol.	Center Acad Publ Japan(日)	126	0.81
		J. Gen. Physiol.	Rockefeller Univ Press(米)	191	5.11
		J. Physiol.-London	Blackwell Publishing(英)	1339	4.35
		Pflugers Arch.	Springer(独)	419	2.26
胃腸系臨床医学	G	Am. J. Gastroenterol.	Blackwell Publishing(英)	819	4.72
		Gastroenterology	W B Saunders(米)	677	13.09
		Gut	B M J Publishing Group(英)	659	6.60
		J. Gastroenterol.	Springer Tokyo(日)	402	1.21

*1) 2000 年当時　　*2) 2002-2003 の 2 年間の掲載論文数(articles+reviews)　　*3) 2004 年

２.２ 検討する要因（説明変数）と目的変数

論文の被引用数を説明する変数として次の 9 つを用いた：(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（規格化したページ数）(Length)。

論文の被引用数分布は極めて歪度が大きいことから、目的変数は(被引用数+1)の対数(log(C+1))とした（被引用数 0 の論文があるため 1 を加える)。ここで C は、論文が 2006 年 10 月までに受けた引用数(C)である。自己引用を含んだ被引用数と除いた被引用数を取得しているが、今回は前者を用いた。

２.３ データの取得

2006 年 7 月に、Thomson Scientific のデータベース Web of Science から、24 誌の 2000 年における原著論文(articles)のデータをダウンロードした後、各誌から 50-60 論文を無作為抽出して調査用標本とした（2 ページ以下の論文、会議プロシーディング論文は除外した)。これらの標本論文の書誌データと、その参考文献及び引用文献のデータを、Thomson Scientific から購入し、これから上記の説明変数及び目的変数のデータを取得した（多くの変数については、変数値取得のための加工が必要であった)。図の数、表の数、数式の数は、別途原論文に当たって取得した。

２.４ 統計解析

6 つの分野別及び 24 の雑誌別に、２.２に示した説明変数と目的変数に対して線形重回帰分析を行った(ステップワイズ法による変数選択)。事前の分析として、各変数の分布状態の確認と変数間の相関分析を行った。統計プログラムには Excel と SPSS Base 16.0 を用いた。

３．結果

3.1　変数の分布形

被引用数 C は極めて歪んだ分布を示し、分野別に見ると歪度 1.56(分野 S)から 3.27(分野 E)であった。対数変換後の log(C+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。無被引用論文(C = 0)の比率は、全体で 6.4%、分野別では 1.3%(分野 B)から 16.6%（分野 E）の範囲であった。

目的変数もほとんどが正の歪度を示したが、C ほど極端ではなかったので、対数変換は行わなかった。Price, Figs, Length は比較的歪度が小さい。分野 G の論文はすべて Eqs = 0 であった。

3.2　相関分析

有意な相関が多かった変数の組み合わせを**表 2** に示す(ここに示す相関はすべて正)。分野と雑誌の標本サイズの違いを考慮して、分野(n = 227 〜 240)では p 〈 0.001、雑誌(n = 51 〜 60)では p 〈 0.05)を有意とした。これらはそれぞれ、ほぼ r 〉 0.22、r 〉 0.26 に相当する。log(C+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 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 変数の間に多数の相関が見られた。特に、Refs-Length 間、Figs-Length 間の r は全分野で 0.46 〜 0.77 と高かった。

表 2　有意な相関が多い変数の組み合わせ

変数の組み合わせ	相関のある分野(p,0.001)	相関のある雑誌(p,0.05)
log(C+1) − Authors	2	4
log(C+1) − Refs	5	7
log(C+1) − Price	5	15
log(C+1) − Figures	2	6
log(C+1) − Length	4	9
Authors − Insts	6	22
Authors − Countries	5	11
Insts − Countries	6	22
Refs − Figures	5	7
Refs − Length	5	24
Figures − Length	6	23
Tables − Length	2	14
Eqs − Length	3	10

以上から、log(C+1)に対する分野共通の有効な説明変数として、まず Price が候補となる。次に Refs, Figs, Length のグループと Authors, Insts, Countries のグループが考えられるが、これらグループ内同士で相関があるので、重回帰分析で変数選択をすると、それぞれのグループから 1 つないし 2 つが選ばれるのではないかと予想される。

3.3　重回帰分析

(1) 分野ごとの重回帰分析

6 つの分野それぞれに対して行った重回帰分析結果の概要を**表 3** に示す。自由度調整済み決定係数 Rc^2 は最良 0.51 で、あまり高くはない。

表 3　分野ごとの重回帰分析結果の概要

分野		P	C	E	B	S	G
n		230	227	229	240	236	233
自由度調整済決定係数		0.196	0.301	0.075	0.415	0.512	0.289
標準誤差		0.398	0.341	0.398	0.301	0.314	0.429
偏回帰係数（t値）	Insts			0.0912 (2.43)			
	Countries						0.180 (2.64)
	Refs	0.00891 (4.42)	0.00319 (2.06)		0.00371 (2.44)	0.00409 (2.09)	0.01043 (5.19)
	Price	0.00835 (6.23)	0.00651 (4.43)	0.00385 (3.06)	0.01087 (10.46)	0.00991 (8.60)	0.00886 (6.26)
	Tables					−0.0371 (−2.34)	0.0451 (2.48)
	Eqs		−0.0136 (−2.52)				
	Length		0.0488 (4.55)	0.0178 (2.29)	0.0280 (3.87)	0.0502 (6.86)	

ステップワイズ法で選択された説明変数は 2 〜 4 種であった（変数投入基準 p 〈 0.05、変数除外基準 p 〉 0.1)。Price, Refs, Length が多くの分野で選択され、特に Price はほとんどの分野で最も説明力が高かった。偏回帰係数から、プライスの 10%上昇及び参考文献数の 10 件の上昇はともに被引用数を 10-25%引き上げる。Insts, Countries, Eqs は 1 分野のみで選択されたが、説明力はそれほど高くない。Tables は 2 分野で採択されたがそれぞれの符号が逆なので、意味のある選択とは言い難い。Authors, Figs はいくつかの分野で目的変数と有意な相関があっ

たが、重回帰では選択されなかった。これらと相関の高い他の変数が優先的に採り入れられたためと考えられる。

(2) 雑誌ごとの重回帰分析

雑誌ごとの重回帰分析の結果は、分野ごとのそれに比べて一般によくなかった。3 つの雑誌では有意な説明変数が 1 つも選択されなかった。回帰が有意であった 21 誌で、Rc^2 の範囲は 0.11 ～ 0.39 であり、選択された変数はほとんど 1 ～ 2 種（1 誌のみ 3 種）であった。しかし、雑誌ごとの集合は分野ごとの集合の 1/4 のサイズしかないので、変数選択の基準が同じであれば少数の変数しか選択されないことになる。そこで、その条件を緩めた（変数投入基準 $p < 0.1$、変数除外基準 $p > 0.2$）ところ、すべての雑誌で回帰は有意となった。最も説明力が高い変数はやはり Price で、16 誌で選択された。次いで Refs（10 誌）、Figs（7 誌）であるが、Figs のうち 2 誌は偏回帰係数が負である。

４．考察

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 20 ～ 50%程度まで説明することができた。同一雑誌内ではこれより低い説明力であった。今回用いた説明変数はすべて、論文の質や内容、あるいは著者の実績や名声の要素を含まない計量書誌的な量であることを考えれば、これ以上の説明力を望むのは無理かもしれない。今後、論文著者の実績を示す 6 つの量(過去の論文数、それらが得た被引用数、著者の活動期間等)を説明変数に加えることにより、説明力が向上するか検討する予定である。

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。どの分野でも Price は最も説明力の高い変数であり、Length と Refs 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。プライス指数が被引用数の説明に有効であることは、これまでほとんど報告されていない。

著者数が被引用数に関係があるとする過去の研究が多いが、今回の結果は否定的である。これまでの研究では、他の要素(分野、雑誌等)との交絡があった可能性がある。１．で挙げたいくつかの重回帰分析研究[3,5-7]では、著者数の説明力は微妙である。図、表、数式の数は、直観的に被引用数との関係は薄いと考えられるが、相関があることを報告した研究[8]があるので候補変数に加えた。単純相関では有意であった場合も見られたが、重回帰分析では積極的な結果は得られなかった。

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、次の 2 つの原因が考えられる：(a)標本サイズが大きく（4 倍に）なった、(b)異質の雑誌の混合により見かけの相関が生じた。しかし、次の 2 つのことから、(b)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。ほとんどの場合、分野内分散はその分野の雑誌内分散の 1 位と 2 位または 2 位と 3 位の間にある。

②分野内と雑誌内で、log(C+1)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。

謝辞 本研究は、科学研究費補助金（基盤研究（Ｂ））により行っているものである。

引用文献

1) 根岸正光；山崎茂明．研究評価．丸善，2001.

2) Seglen, P. O. The skewness of science. *J Am Soc Inf Sci.* 43(9), 628-638, 1992

3) Peters, H. P. F.; van Raan, A. F. J. On determinants of citation scores: A case study in chemical engineering. *J. Am. Soc. Inf. Sci.* 45(1), 39-49, 1994.

4) Callaham, M. et al. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-revied journals. *JAMA* 287 (21), 2847-2850, 2002.

5) Basu A.; Lewison, G. Going beyond journal classification for evaluation of research outputs. *Aslib Proc*. 57(3), 232-246, 2005.

6) van Dalen, H. P.; Henkens, K. Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics* 64(2), 209-233. 2005.

7) Bornmann, L.; Daniel, H-D. Selecting manuscripts for a high-impact journal through peer review: A citation analysis of Communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *J. Am. Soc. Inf. Sci.* 59(11), 1841-1852, 2008.

8) Snizek, W. E. et al. Textual and non-textual characteristics of scientific papers. Neglected science indicators. *Scientometrics* 20(1), 25-35, 1991.