2C27　書誌計量学的手法によるピアレビューの支援可能性の再検討

○林　隆之（大学評価・学位授与機構）

1．はじめに

研究評価における学問的質の評価の方法としては、評価対象分野の専門家に判断を仰ぐピアレビューが主流であることは各国に共通する。しかし、ピアレビューにも様々な問題があることは指摘されている(e.g. Chubin 1990, Kostoff 1994)。ピアレビューは基本的に評価者（レビューアー）の主観的判断に委ねられるために、意識的・無意識的にバイアスが入る。例えば、保守的な傾向、若い研究者や新参者の過小評価、ハロー効果、個人的・組織的なえこひいきである。さらに、分野ごとの専門家による評価であるため、分野間の比較や学際分野の判断が困難である。評価対象の数が多い場合には、全ての分野を網羅する評価者を揃えることも難しい。また、一人が評価を行う数は限られるため、少ないサンプル数の中での比較とならざるを得ない。このような問題のため、ピアレビューは伝統的な単一分野内部のみで評価を行う場合には比較的有効であるが、評価対象の数が多く複数の分野にまたがる場合には極めて脆弱であり、評価者の構成に評価結果が左右される可能性が高い。そのため、ピアレビューの質を向上するには、被評価者側からの報告だけを資料として評価者が半ば印象により評価を下すのではなく、他の評価手法から生成される定量的・定性的情報をも参考として上述の問題を補って評価を行うことが必要とされる。

ピアレビューを補足する手法の中でも学問的質が評価基準である場合には書誌計量学的手法（ビブリオメトリクス）がその代表例として挙げられる。書誌計量学的手法は論文数や被引用数ならびに引用や単語の共出現現象を基に研究開発活動の特徴を分析するものであり、特に論文数は研究活動の生産性の高さを示し、被引用数は他の研究者からその論文への「投票」としてのインパクトの大きさを示すと考えられるために、欧米諸国では評価に頻繁に用いられてきた。しかし日本ではこれまで書誌計量学的手法の利用については両方向からの反応がある。一つは、日本では欧米諸国と比べて体系的に用いられていないという定量的手法の欠如に対する評価者側からの批判である。しかし一方で、このような書誌に関する指標の意味は不確定であり問題点も多いため評価に使うべきでないという拒絶的な反応も研究者側からは強い。さらには、一部の大学ランキングや医学分野での人事における利用では、書誌計量学的手法に内在する問題を考慮せずに安易な形で分析を行い、その結果のみが一人歩きすることもある。

このような日本の現状を鑑み、本論では書誌計量学的手法が日本という非英語圏の国でどの程度、学問的質の評価に対して利用可能であるかを再検討する。事例として大学評価・学位授与機構(NIAD)が2000～01年度に行った理学分野の大学の研究評価を取り上げ、実際にピアレビューを補足する情報を形成しうるか考察する。

2．書誌計量学的手法とピアレビューとの整合性の先行研究

書誌計量学的手法が有効であるためには、それがピアレビュー結果とある程度同等のものを少コストで提示できる必要があり、さらにはピアレビューの誤判断を防ぐための情報を提供できることが必要である。書誌計量学的手法とピアレビューの結果の整合性の検証は、既に1960年代から焦点を置かれてきた (Narin 1976)。通常、ピアレビュー結果は詳細には公表されないことが多いため、先行研究の多くでは公表されている大学ランキングや大学評価の結果との比較による検討が行われている。米国ではCartter(1966)やRoose and Andersen (1969)による教員への評判調査の結果との比較が行われており(e.g. Anderson et al 1978)、英国では大学評価であるResearch Assessment Exerciseの結果との整合性が検討されている(Irvin 1989, Zhu et al. 1991)。他方、オランダでは大学協会(VSNU)が行う大学評価において、幾つかの分野でライデン大学 CWTS が書誌計量学的分析を受託して行っており、その結果が評価者に提示されるとともに整合性の分析もなされている (Rinia et al. 1998, 2001)。これら先行研究はいずれも両者の間で整合性が高いことを認めており、書誌計量学的手法の利用の根拠となっている。だが、これらは大学という機関全体レベルあるいはその内部の研究グループを対象としたものである。機関レベルではその規模が論文数やスター研究者の多さにつながることから、両者の整合性が高いことは比較的容易に期待できる。他方で、組織を構成する研究者の評価や研究プロジェクト選定の評価では個人レベルでの研究の質が中心となり、どの程度書誌計量学的手法が整合性を持ちうるかは明らかではない。本論では個人レベルでの整合性に焦点を置く。

3．方法

3.1 事例分析対象

書誌計量学的手法に限らず特定の評価手法を用いる際には、評価対象の規模・レベル、評価単位、対象期間、評価項目などにあわせて、テーラーメイドで方法の設定・修正を行う必要がある。そのため、まず分析対象を設定する。本分析ではNIADが2000～01年度に行った理学分野の大学の研究評価を事例対象とする。本評価で対象となったのは5大学1共同利用機関である。本評価は基本的には学部・研究科を単位に研究体制・支援体制や方策などを評価するものであるが、研究成果については個人レベルでの評価（「研究業績の判定」）を積み上げることで組織レベルの評価とした。個人レベルの評価では、理学分野を数理・情報科学、物理学、化学、生物科学、地球科学、天文・宇宙科学の6領域に区分し、各研究者（教員）はここから1つ、ないし複数の領域を選択し、最近5年間の研究内容や主要業績リストおよび業績5点以内を提出した。領域ごとの専門家から構成される各部会において、これら提出情報を基に最低2名の評価者が各研究者の研究業績の評価（「判定」）を行った[1]。この評価では学問的質（「研究水準」）と社会的貢献の2項目が評価され、学問的質については評価は4段階（および研究評価の対象には当たらない「該当せず」）に区分された（個人別の結果は非公開である）。この4段階の基準は各領域ごとに文章で設定されている。

3.2 書誌計量学的手法の精緻化

上記のピアレビュー結果との整合性を分析するため、書誌計量学的手法を精緻化する必要がある。書誌計量学的手法にも様々な方法論上の欠点がある(Martin and Irvin 1983, Schubert, 1996)。一つは分野ごとに指標の平均値が異なることであり、ピアレビューと同様に分野を超えた比較はできないとされる。また、データベース自体に収録雑誌・分野や言語の偏りがあり、データベース上の入力誤りや表記揺れもある。また自己引用や論文の分割投稿の問題もある。これまで様々な改善の試みはあるが、実際の評価実務に定型的に使用しうる品質として、上述のオランダ・ライデン大学CWTSの補正(Van Rann 1996)を参照しつつ、以下のように技術的改善を行った。

・被引用数の標準化： NIADの評価では被評価者が提出する業績リストの形式は緩やかにしか規定されず、何をどの程度記載するかは被評価者に委ねられた。そのため単純にアウトプットの多さを比較することは適さない。また評価者は提出された5編に実際に目を通して研究の質の評価を行ったことから、本論では質を反映していると考えられる被引用数の測定を中心に分析を行う。だが上述の通り被引用数の平均値は分野によって異なる。NIADの評価におけるピアレビューでも領域ごとに独立に評価が行われたが、評価結果を参照する社会一般の側からすれば、領域ごとに評価基準が異なれば解釈の誤りにつながる。また、各領域内部でも複数の研究分野があり、それらの間の比較可能性も担保することが求められる。そのため、書誌計量学的手法において何らかの標準化を行い、異なる研究分野間でも比較を可能性とする必要がある。本分析では*SCI*で用いられている約170の分野分類を基に標準化を行った。*SCI*では各ジャーナルについて1つ以上の分野分類が付与されているため、対象の論文（の掲載されたジャーナル）の分野分類と少なくとも一つの分野分類を持つジャーナルの全論文を参照範囲とした[2]。ただし、分野分類が複数付されているジャーナルの論文では分数カウントを行った。そのため分野分類の重なりが大きいジャーナルの論文ほど加重カウントされる。具体的には*SCI*(CD-ROM版)に毎年約80万件収録されている全論文の被引用数を計測し、分野ごとにその平均と分布を算出してこれを基に標準化を行った。標準化の方法としては、当該分野の全論文の平均被引用数との比を指標とする方法、および当該分野での被引用数のランキングの位置（被引用数が当該分野で上位何%に入るか）を指標とする方法を候補とした。また、1年前に出た論文は被引用数が少ないことを鑑み、各論文でなくジャーナルの平均被引用数(IFに相当)の分野平均との比を指標とする方法も候補とした。

・出版年、文書形式の区分： 出版年が古い論文ほど引用される期間が長くなるため、上述の比較は同一年に出版されたものの間のみとした。また文書形式もArticle, Review, Letterを区分する場合としない場合の双方を検討した。

・自己引用の除去： 同一著者名を含む論文からの引用を自己引用と機械的に推定し、*SCI*上の全論文について、自己引用を除去する場合としない場合の双方の被引用数を測定して分析した。

・その他の補正： 著者名のミドルネーム有無の表記揺れや引用論文の記述の表記揺れを補正した。

このような方法を用いることにより、被引用数の指標について次のような分析上の選択肢を設定できる。すなわち、①指標として、論文の被引用数の分野平均値との比、分野内のランキングの上位%、ジャーナルの平均被引用数の分野平均との比、②自己引用を含む、含まない、③文書形式の区分を行う、行わないである。さらにNIADの評価では

---

[1] 評価の際に評価員には雑誌のインパクトファクター(IF)の一覧を参考資料の一つとして配布している。また、各評価員が独自にSCIを検索することは妨げていない。そのため、評価員の評価が書誌計量学的データと全く独立に行われたとは言えない。

[2] Scienceやnatureなどの「multidiscipline（学際分野）」という分野分類が付されているジャーナルについては、個々の論文について、その論文の参考文献リストに記された論文のジャーナルの分類を集計して上位3つの分類を選択し、当該論文の分野とした。

業績５編が実際に提出されたが、本分析では指標化する④分析対象として、提出された５編、それに拠らずに被引用数上位５編、業績リストの全ての合計の選択肢を設定した。分析ではこれら選択肢の様々な組み合わせを試行した。

４．結果

4.1 研究アウトプットの *SCI* 収録割合

書誌計量学分析で対象となるのは、研究アウトプットの内で「ジャーナル論文」のみである。すなわち、書籍や報告書や学会発表・講演は含まれない。さらに、引用分析に用いる *SCI* では、収録されているものの殆どが英文誌である。そのため、まずは書誌計量学分析が研究アウトプットのどの程度の割合を分析しているかを明らかにする必要がある。表１は、各部会に提出された研究者の業績リストの全体的傾向を示したものである。表からは、報告書や学会発表等をも含む全業績リストの中で「英文ジャーナル論文」の割合は、物理や化学では８割を超える一方、地球科学、数理・情報では半数程度でしかないことがわかる。さらに、「英文ジャーナル論文」の内で *SCI* に収録されている割合も数理・情報および地球科学で低い。そのため、この２分野では提出された研究アウトプットの内で *SCI* による書誌計量学分析の対象となるのは30%以下に過ぎず、分析には限界があることを認識しておく必要がある。

表１　部会ごとの *SCI* での検索ヒット率

部会	研究者数	*SCI* で検索された全論文数	英文ジャーナル論文		*SCI* 上の論文	
			業績リストに占める割合	ジャーナル論文に占める割合	英文ジャーナル論文の中での検索率	業績リストに占める割合
数理・情報	127	280	56.8%	87.4%	40.4%	22.9%
物理	170	2,316	84.9%	94.9%	87.3%	74.1%
化学	146	2,358	85.2%	95.2%	91.7%	78.1%
生物科学	132	1,073	78.1%	91.3%	79.6%	62.1%
地球科学	128	560	48.1%	63.7%	59.9%	28.8%
天文・宇宙	183	1,333	68.8%	91.8%	83.6%	57.5%

4.2 ピアレビューとの整合性

上述の多種の選択肢による測定について、ピアレビュー結果との整合性を Spearman の順位相関により測定した。相関が高かったのは、数理・情報を除き、「根拠資料によらずに被引用数上位５編を選択し、その被引用数ランキング上位%[3]の合計を、自己引用を含む形で指標化した場合」であった。文書形式の区分では違いは生じなかった。被引用数上位%が適していた理由は、分野によっては被引用数の分野全体の平均値が１以下であるために、それとの比では数回しか引用されていない論文が過大評価されるためと考えられる。また、自己引用を含む場合の方が相関が高い理由は、自己引用が多いことは当該研究者が論文を多く産出していることを背景としており、ピアレビューでもそれが評価されたことと整合したと考えられる。また、根拠資料によらないで被引用数上位５編を選択することも、論文産出数が多い場合にはより被引用数の高いものが計算に組み入れられる可能性が高くなるためである。一方、数理・情報領域では各論文の被引用数ではなく掲載されたジャーナルの平均被引用数の分野平均との比を用いた場合が相関が高かった。論文ごとの被引用数の相関が低い理由は、この領域では被引用数が０回の論文が多く、差が出なかったためである。

表２ ピアレビュー結果との相関

部会	引用数	アウトプット数
数理・情報	0.259*	0.445**
物理	0.591**	0.556**
化学	0.611**	0.578**
生物科学	0.659**	0.705**
地球科学	0.387**	0.339**
天文・宇宙	0.416**	0.459**

他方、被引用数ではなく、アウトプット数との相関も表には示している[4]。数理・情報ではアウトプット数の方が被引用数より相関が高いが、他分野ではほとんど差が出ていない。そもそも、被引用数上位%とアウトプット数の相関自体が高く、これは論文生産性の高い研究者は被引用数の高い論文を産出していることを示唆している[5]。

4.3 差異の原因

被引用数の分析結果についてもピアレビュー結果と同じ割合で評点１～４にグループ分けを行った。その結果、両者で２段階以上の差がついたものは各分野で数%に過ぎなかった（表３）。これから両分析の間で大きく異なる結果は出ていないと言える。ピアレビューが２段階低く評価しているケースにおいて、評価者が記入シートに付していたコメントを示すと（表４）、提出書類に研究内容の説明が殆どないことや関与の割合が不明などの「記述不足」が最も多いが、他方で「筆頭論文がない」ことを厳しく評価したり、「レベルが低い」という主

表３　ピアレビューと書誌計量学的分析との相違

部会	同じ	１段階異なる	２段階以上異なる
数理・情報	48%	47%	4%
物理	54%	44%	1%
化学	57%	38%	4%
生物科学	60%	37%	3%
地球科学	65%	27%	8%
天文	58%	41%	1%

[3] 下記の計算では指標を(100%-上位%)と変換して、100 が最も引用回数が高く 0 が低いというように向きを変更している。

[4] アウトプットには論文や報告書などあり、どの種類を合計に入れるか選択肢があるが、表では最も相関が高い場合を示している。数理・情報、物理、天文では全アウトプット数が最も相関が高く、地球科学では英文・邦文ジャーナル論文とプロシーディングスの合計、生物科学、化学では SCI 上にある論文の数であった。

[5] なお、論文を引用数の上位%で重み付けして合計した値の相関は上記の指標の相関よりも低かった。

観的なコメントのみの場合、さらには記入シートには特に否定的なコメントは付されていない場合もあり、論文の被引用数からは把握できない明確な理由を評価者が有していたとは言い切れない。

また先述の表３ではピアレビューと書誌計量学的分析の間で一段階のみの差異が生じたものは全体では40%程度にも上ることを示している。だが、他方でNIADのピアレビューの過程において、同一の被評価者に対して二人の評価者が最初につけた判断（最終的に合議により判断を行う前の段階での判断）で一段階の差異があったものも35%と同じ程度であった。すなわち、一段階の差異は評価者や分析方法の違いによって容易に起こりやすいものであると言える。さらには評価者の間で当初２段階以上の差異がついたものも３%存在しており、それらを詳細に見ると、概して厳しめの評点をつける評価者が存在するなど、初期には評価者の間で評価の基準が明確には共有しきれていなかったことも伺える。それは評価者間だけでなく、領域間での差異にも表れている。書誌計量学的分析を基に、被引用数上位%という指標においてどの範囲がピアレビューで各評点１～４に区分されているかを分析することができる。その結果の図１からは、(「英文ジャーナル論文」形式のアウトプットが多くない領域では値が低くなることには注意が必要であるが)、領域によっては他よりも上位の評点を付け易い傾向を有していたことが分かる。

表４　ピアレビューの方が２段階低く評価したものに付されていたコメント

コメント	件数
研究内容の説明がない。論文への関与の割合が不明。	5
筆頭論文や単著がない	4
研究数や研究のレベルが低い	4
（その他）	5

（各被評価者に対して評価員は２人ずつ）

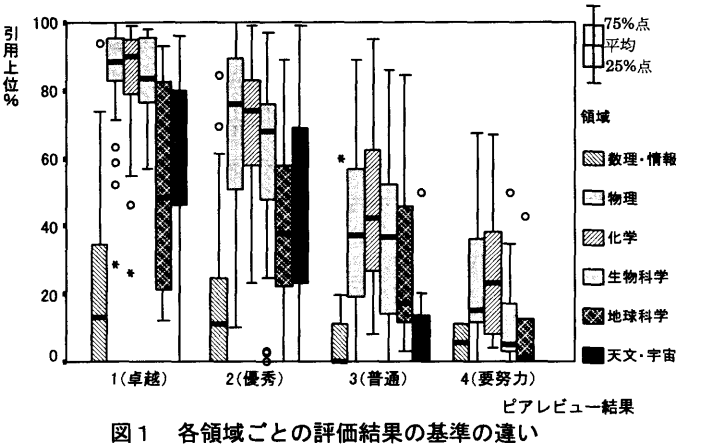

図１　各領域ごとの評価結果の基準の違い
（*SCI*に１本以上論文がある研究者のみ）

5. 結論 ～ 書誌計量学的手法の有効性と限界

本分析では、数理・情報を除く各分野については、ピアレビューと書誌計量学的分析との間にある程度の整合性が認められた。これから、*SCI*データベースで論文が十分検索できる領域では、書誌計量学的分析が評価者の負担軽減のための情報を生みうることは示唆された。特に両者で２段階の差がつくことはまれであるため、評価者の思いこみ等によって大きく誤った評価結果を生むことを抑止する情報にはなる。同時に、書誌計量学的手法は領域間や評価者間で、評点をつけるための標準的な基準を共有することを支援することができる。書誌計量学的手法により、評価者は同時に評価を行っている少数のサンプル内の比較ではなく、世界全体の論文の中でのベンチマークを行うことができる。「世界全体の論文の中で当該論文の引用回数が上位どのくらいに位置しているか」という標準的な指標を設定することにより、異なる分野の評価者の間でも、大きな評価基準の相違が生じないように支援することを可能とする。

しかしながら、本分析でも示されたように、評価者間あるいは書誌計量学的分析との間では判断が一段階異なることは頻繁に起こりうることであり、書誌計量学的手法やその他手法の支援によっても一段階の差が生じなくできるとは期待できない。このことは、一般的に、研究評価の結果を基に資金配分を行う際に一段階の評点の違いで大きく資金額が異なるような方法は正当化しにくいことを示唆する。一段階の誤差が許容できるほどの緩やかな結びつきを有する資金配分を複数設けることにより、極端なリスクは回避されるべきであると言える。

他方、評価は資金配分だけでなく機関の改善の促進という目的もある。書誌計量学的手法はその他の指標（研究費や共同研究数などの各種の指標）と組み合わせることにより、当該機関の研究活動の特徴や競争力を有する分野の情報を少コストで提示できる可能性を有する。書誌計量学的手法が個人の研究アウトプットだけでなく、機関レベルでの改善と戦略性の促進という目的をいかに支援できるかは別に検討する必要がある。

【主な参考文献】
Chubin and Hackett (1990), *Peerless Science*, SUNY Press
Irvin, J. (1989),"Evaluation of scientific institutions", *The Evaluation of Scientific Research*, John Wiley &Sons, pp.141-168
Martin, B.R. and Irvine J. (1983), "Assessing basic research", *Research Policy* 12, pp.61-90
Narin F. et al.(1976), *Evaluating Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*, CHI
Rinia E.J. et al. (1998), "Comparative analysis of a set of bibliometric indicators and central peer review criteria", *Research Policy* 27, pp.95-107
VanRaan, A.F.J.(1996), "Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises", *Scientometrics*, 36, 397-420