Original Article

Subscription to digital libraries and corresponding journal impact: A value-based approach to demand for digital research data

Soumitra K. Mallick

is Professor of Economics & Finance & Control at the Indian Institute of Social Welfare & Business Management in Kolkata, India. He is a B.Com from St Xavier's College, Kolkata; Chartered Accountant; and a PhD from the Economics Department of NYU, NY, USA. IISWBM is the first management institute of India and is affiliated to Calcutta University for the degrees.

Thomas Krichel

is Assistant Professor of Library & Information Science at the Palmer School of Library & Information Sciences. He holds an M.S. from the University of Paris, Sorbonne, Paris; M.A. from the University of Exeter, Devon UK; PhD from the University of Surrey, Surrey, UK. Palmer School of Library & Information Sciences is in the C. W. Post campus of Long Island University, Brookville, NY, USA. He is also a faculty member at the Division of Information Systems, Faculty of Information Technologies, Novosibirsk State University, Novosibirsk, Russian Federation.

Marco Novarese

is a researcher in Economics at the Centre for Cognitive Economics, University of Piemonte Orientale, Alessandria, Italy and Assistant Professor in Cognitive and Experimental Economics at the Faculty of Law in the same University. The Laboratory aims at developing the application of Cognitive Science to Economics. He holds a PhD from Université de Nice – Sophia Antipolis, Nice, France.

ABSTRACT This article finds an explanation for the subscription to those various fields in the *Research Paper in Economics* (RePE) e-library, which form the RePEc reports. Subscription to a particular report is regressed on the value as to which papers in that field create as measured by the impact factor of the relevant field journal and the ranking. Subscription has been equated to demand for the fields by analysing demand in cardinal measurement and ordinal measurement. We find that both measurements give rise to demand for digital research data, which is explainable by subsequent value created in the form of publication and citations in high-quality journals for the respective fields.

Journal of Digital Asset Management (2010) 6, 336–343. doi:10.1057/dam.2010.38

Keywords: digital library; RePEc e-library; demand; impact factor; subsequent value created

INTRODUCTION

Subscription to digital libraries is a matter of considerable evolving interest in the digital era. Digital libraries substitute for physical libraries by freeing-up space and making accessibility potentially global. They also provide for online sorting and ranking of their contents based on various criterion, and because of their potential lack of space requirements can be upgraded very

fast both in terms of quantity as well as quality, for example by providing more specialised categories. A specific e-library that forms the subject matter of this research is the website <http://repec.org/>.¹ 'Research Papers in Economics (RePEc) is a collaborative effort of hundreds of volunteers in 63 countries to enhance the dissemination of research in economics. The heart of the project is a decentralised database of working papers, journal

Correspondence:

Soumitra K. Mallick
IISWBM, Management
House, College Square
West, 3, Madan Mohan
Sen Street, Kolkata,
West Bengal 700 073,
India
E-mail: smallick@iiswbm.edu

articles and software components. All RePEc materials are freely available'.²

An important service provided by RePEc is New Economics Papers (NEP). 'NEP is an announcement service which filters information on new additions to RePEc into edited reports. The goal is to provide subscribers with up-to-date information about the research literature'. The bibliographic data available here are contributed at no charge by editors and made available freely. NEP organises working papers in economics into over 80 fields like macro-economics (NEP-MAC), insurance economics (NEP-IAS), sports economics (NEP-SPO) and so on. Every week all the working papers from participating institutions in different countries are loaded on to the RePEc archive, which is then circulated by the report editors to the subscribers according to the field of subscription. Other than working papers some journal articles are also circulated through the archive and the reports.

During the process of circulation, an article may be circulated in more than one report if it contains relevant material. Also, any subscriber may be a subscriber to more than one report. The number of subscribers for any report is a fixed number, at a given point-in-time, where the identity of the subscriber is contained in the archive but identities may overlap across fields. There is contained in the over 80 fields the NEP-ALL report, which is made up of all the papers in that particular issue and has its own set of subscribers. However, all the NEP-ALL papers are allocated to at least one of the fields, unless they are ambiguous. Hence, the chance of their being ambiguities in the subscription to the various specialised reports is negligible. If subscription implies preference for the working papers or papers in any particular field, then preferences do not have externalities across fields and it can be safely assumed that the subscription to any one field is independent of the subscription to any other field. This property is important in comparing subscriptions between fields in its underlying interpretation as the 'demand for that particular field papers' (to be used interchangeably as 'demand for a particular field').

The purpose of this article, therefore, is to find an explanation for the subscription to the various fields in this e-library which forms the

RePEc reports. Various categories of explanations can be derived for this newly evolving 'digital asset' demand. One can relate the subscription to a particular report in terms of its price because subscription is being interpreted as the demand and one can define an implicit price in terms of the time it takes to go through the papers in a report, which is not only a reflection of its quantity but also its quality. But this article is not concerned with the price implication of subscription. Rather, we try to start off by taking a dynamic explanation, which originates from the fact that subscription to reports is followed by the papers getting published in various field journals after circulation and subsequent peer review and the references to the papers in subsequent research determines the value they have created to the subscribers to the papers in working paper form circulated through the various reports. Hence, subscription to a particular report is sought to be regressed on the value which papers in that field create as measured by, for example, the impact factor of the relevant field journal or the one with the highest value.

The utilitarian explanation to subscription to digital libraries like RePEc, especially, in the nature of research working paper digital libraries is, in our opinion, new. Economic theory has a considerable amount of literature on demand for information, in general, in a market setting, where goods and services are being traded³⁻⁵ and for R&D expenditures by firms.⁶ In the former it has been discussed how information can be considered to be a collection of sets that are being traded, thus highlighting the diversity which is essential for information being a commodity, and in the latter it has been discussed how positive externalities created by spillovers into the productivity of labour and capital gives rise to the demand for R&D, both, however, going on to show that pricing through competitive markets of information and R&D is a problem. The nature of digital libraries such as RePEc lies in being a composite of both providing information on research in the differentiated fields dynamically and providing the results on research in various fields which can be put into practise in production, thus the above theories should apply in developing the demand function for digital libraries, simultaneously. However, this article

does not take such a route. Instead we study subscription to RePEc from the dynamic utilitarian viewpoint as has been discussed.

The utilitarian approach of this article does not consider the microfoundations to learning by using any utility function explicitly,^{4,7,8} but consists in relating the number of subscribers of the reports to the value-based ranking of the corresponding field journal in 2008 ranked by simple impact factors. What is of primary importance in order to do this association between the two sub sections of the RePEc data set is the element of our model that 'demand for a particular field' = 'demand for that particular field papers'. Where there are a number of field journals we consider the journal with the highest impact factor.

In the section below we describe the data set. In the next section we estimate the model with the 'simple impact factor', which is defined as the ratio of citations by the number of items in the journal. This gives the value of the journal to the subscribers in terms of its research potential and thereby gives the utility to the researchers for learning from the reports. In the section thereafter we estimate the model of subscription in terms of the journal ranking and impact factor and check its fit in comparison to considering the variables separately. After which we conclude.

DATA SET

The data for the number of subscribers as well as the impact factor and the number of citations, including the name of journals is taken from the RePEc database <http://repec.org/>. The data on the specific reports and the number of subscribers is contained in the NEP section of the website. The data on the journals in the various fields, including their simple impact factors and the number of citations are contained in the rankings subsection of the website. The simple impact factor is calculated as the ratio of citations/number of items in the journal.

There are 82 fields as of June 2008. They are as follows:

- (1) Business, Economic & Financial History;
- (2) Education; (3) Sports & Economics;
- (4) Unemployment, Inequality & Poverty;
- (5) Public Economics; (6) Financial Markets;
- (7) All new papers; (8) Econometrics; (9) Central & South America; (10) Social Norms & Social

- Capital; (11) Labour Economics;
- (12) Development; (13) Econometric Time Series; (14) Macroeconomics; (15) Industrial Competition; (16) Monetary Economics;
- (17) Entrepreneurship; (18) Industrial Organization; (19) Game Theory; (20) Cognitive & Behavioural Economics; (21) Dynamic General Equilibrium; (22) International Finance;
- (23) Microeconomics; (24) Corporate Finance;
- (25) Environmental Economics; (26) Central Banking; (27) European Economics;
- (28) Transition Economics; (29) Innovation;
- (30) Experimental Economics; (31) Health Economics; (32) Public Finance; (33) Energy Economics; (34) Computational Economics;
- (35) Accounting & Auditing; (36) Technology & Industrial Dynamics; (37) Network Economics;
- (38) Banking; (39) Efficiency & Productivity;
- (40) Economic Geography; (41) Evolutionary Economics; (42) History & Philosophy of Economics; (43) Discrete Choice Models;
- (44) Agricultural Economics; (45) Law & Economics; (46) Positive Political Economics;
- (47) International Trade; (48) Risk Management;
- (49) Regulation; (50) Urban & Real Estate Economics; (51) Human Capital & Human Resource Management; (52) Collective Decision-Making; (53) Business Economics;
- (54) Forecasting; (55) Resource Economics;
- (56) Post Keynesian Economics; (57) China;
- (58) Microfinance; (59) Economics of Strategic Management; (60) Information & Communication Technologies; (61) Cultural Economics; (62) Africa; (63) Marketing;
- (64) South East Asia; (65) Economics of Happiness; (66) Sociology of Economics;
- (67) Insurance Economics; (68) Neuroeconomics;
- (69) Financial Development & Growth;
- (70) Knowledge Management & Knowledge Economy; (71) Economics of Ageing; (72) Utility Models & Prospect Theory; (73) Tourism Economics; (74) Market Microstructure;
- (75) Central & Western Asia; (76) Economics of Human Migration; (77) Intellectual Property Rights; (78) Open Macroeconomics;
- (79) Project, Program & Portfolio Management;
- (80) Confederation of Independent States;
- (81) Contract Theory & Applications; and
- (82) Operations Research.

There is a maximum of 3926 subscribers for Business, Economic & Financial History

(NEP–HIS) and a minimum of 22 subscribers for Operations Research (NEP–ORE).

Corresponding to these series there are 67 journals which can specifically be related to these fields as field journals with the maximum simple impact factor amongst all the relevant journals for any given field. These include the journal Accounting, Business & Financial History with impact factor of 0.16, which is arrived at by multiplying the number of adjusted citations 34 divided by the number of items in the year 2007, which is 217. The highest impact factor in 2007 is that of *Econometrica*, which has a simple impact factor of 20.09. This data is of course either ‘open’ which means the data is for the entire set of journals contained in the NEP subsection or it is ‘closed’, which means it is only among the set of journals which conform to the model of this article, that is, the best impact factor field journal for each of the series in NEP for which there are journals.⁹

MODELS

The basic model

This model can be written as follows:

$$S = a_1 + a_2 \star \text{Factor} + \tilde{\varepsilon} \quad (1)$$

where, S =number of subscribers; a_1 and a_2 are the coefficients; and Factor represents the simple impact factor, $\tilde{\varepsilon}$ is the random error term with mean 0 and standard deviation $\sigma_{\text{BM}} > 0$. This model explains the volume of subscription in terms of the simple impact factor. This is a static linear Generalised Least Squares (GLS) model where the error is a simple 0 noise error as usual.

THE MULTIFACTOR MODEL

The model can be written as

$$S = b_1 + b_2 \star \text{Jnl.Rk.} + b_3 \star \text{Factor} \\ + b_4 \star \text{Adj.Cit.} + b_5 \star \text{Items} \\ + b_6 \star \text{Citations} + \tilde{\gamma}$$

Where, S =number of subscribers, and b_i and s are the coefficients on the corresponding value terms as follows:

1. Jnl. Rk. = Closed Journal Rank for the journals which correspond to the series and

are of the highest simple impact factor for each of the fields for the year 2007.

2. Factor = Simple Impact Factor for the journals corresponding to the closed set of fields which is defined as the number of citations of the journals during the year/ total number of papers (items), however no normalisation has been done amongst the closed set of simple impact factors, which is an initial approximation for the year 2007.
3. Adj. Cit. is the adjusted citations = Citations $\star$ Factor and gives the quality adjusted value of the citations and is the real value variable in its linear form, that is, in the simple compounding form for 2007. This variable converts number of citations in different field journals in equivalent units in terms of the simple impact factors. An alternative measure of impact value assessment has been done elsewhere in the literature in terms of number of pages.¹⁰
4. Items = Number of papers in the relevant field journal during the year 2007. Here also instead of the number of papers, Kalaitzidakis *et al*¹⁰ has used the number of pages. This reflects the number of papers in the relevant field.
5. Citations = Number of citations in refereed journals within 2007 for each of the journals, within the closed set of journals.

$\tilde{\gamma}$ is the random error term with covariance matrix $\sigma_{\text{MF}} > 0$ and has dimensions 6×6 . As this is a GLS specification we will test for the heteroskedasticity in the model by using the F -test (the justification of the procedure has been provided by Mallick *et al*.¹¹

RESULTS

The basic model results

The estimated basic equation obtains as:

$$\hat{S} = 524.73 + 9.39 \star \text{Factor}^{\wedge}$$

(8.29) (0.84)

where, $\hat{S}$ is the number of subscribers and Factor is the simple impact factor.

The t -statistic on the Factor variable is significant at a two-tailed probability of 0.40, the t -statistic on the constant is however, insignificant.

The regression statistics for the fit of the model is presented in Table 1.

The analysis of variance (ANOVA) statistics on the model autocorrelation is presented in Table 2.

The plot of the residuals is given in Figure 1.

Although the *R* squared is very low and so is the multiple *R* squared the *t*-statistics and the *F*-statistics are significant and the plot of the residuals is enough to suggest that there is some explanatory power of the simple impact factor in explaining the subscription to the RePEc working paper series. Hence, a value-based model for subscriptions require a multifactor in-depth model.

The multifactor model has been considered in two forms. One is by taking number of subscribers as the exogenous variable (the Cardinal model), the other is by taking the rank of the series as the exogenous variable (the Ordinal model). Both these variables capture the nature of subscription to the various series and can be used as dummies in the ultimate analysis in explaining the demand for the respective fields. The first would measure the cardinal properties of demand for the field, whereas the second would measure the ordinal properties of demand for the field, ranking them in relation to each other. This ranking is closed in terms of the series which have already been identified and research is categorised, and

subscription and hence demand is identified, in terms of these series. Of course, as the number of series increases, or existing ones are closed down, the model will have to become dynamic and the ranking will have to be open to accommodate for these new entries or exits.

The multifactor model results

The Cardinal model (*dependant variable: number of subscribers*)

The estimated equation obtains as :

$$S = 580.13 - 0.22 \star \text{Jnl. Rk.} + 13.74 \star \text{Factor} - 1.03 \star \text{adjcit} - 0.04 \star \text{items} + 1.02 \star \text{allcit}$$

(4.16) (-0.41)
(0.49) (-0.68)
(-0.88) (0.68)

S = number of subscribers, *Jnl. Rk.* = journal rank, *Factor* = simple impact factor, *adjcit* = adjusted citations = all citations * *Factor*, *items* = number of items and *allcit* = all citations

The *t*-statistics on all the variables are significant at the 38 per cent level (*Jnl.Rk.* at 69 per cent and *Factor* at 68 per cent level). The regression statistics are presented in Table 3.

Table 1: Regression statistics for the fit of the basic model

Regression statistics	
Multiple <i>R</i>	0.102494
<i>R</i> square	0.010505
Adjusted <i>R</i> square	-0.00426
Standard error	407.1072
Observations	69

Table 3: Regression statistics for the cardinal model

Regression statistics	
Multiple <i>R</i>	0.203663
<i>R</i> square	0.041479
Adjusted <i>R</i> square	-0.03459
Standard error	413.2091
Observations	69

Table 2: ANOVA statistics for the basic model autocorrelation

ANOVA	df	SS	MS	F	Significance F
Regression	1	117888.7	117888.7	0.711303	0.402013
Residual	67	11104328	165736.2	—	—
Total	68	11222217	—	—	—

Figure 1: Plot of the residuals for basic model.

The ANOVA statistics on model heteroskedasticity are presented in Table 4.

The plot of the line fit with respect to the Factor variable is presented in Figure 2.

As this is a multifactor model the residuals and model fits can be generated with respect to each of the variables' marginal models. We present here the fitted graph only with respect to the Factor variable. The fit is reasonably good considering that we are only considering the journal with the highest simple impact factor for each field as an initial model. As there are two or three journals in many of the fields these statistics will definitely improve, it is conjectured, with the addition of other journals in the field and including their adjusted citations. The adjusted *F*-statistic is also reasonably high suggesting low heteroskedasticity of the model although a number of citation variables for the journals have been included.

The coefficient on the Factor variable is positive which implies that the number of subscribers is positively related with the simple impact factor of the corresponding journal; however, the number of subscribers is negatively related with the journal rank and the adjusted citations according to the definition in this article, which is the definition being used by RePEc. Hence, the demand for research output through the mechanism of subscription to RePEc lists is positively related to the corresponding journal quality and hence value as measured by the simple impact factor but

negatively related to value as measured by journal rank, adjusted citations and all citations. Also, a comparison of the single factor model with the multifactor model reveals that the coefficient on the factor variable increases by including the other factors like journal rank and adjusted citations.

This may mean that there are spillover effects in assessing the quality of a journal as measured by simple impact factor, adjusted citations, journal rank and all citations.

Of course in the definition of simple impact factor there is the number of citations of the journal during the year, term, which is also included in the adjusted citations and all citations terms, so there will be some distribution of weights, but why the signs change and the magnitude of the weight on the factor term increases is not clear. One possible explanation for this could be found in the observation that the value for any field as is contained in the publication of the field papers in the top field journals depends in a non-linear way on the citations of papers in the journals, and hence the response of the subscription to the fields depends on the factor and citations are in opposite ways. This requires separate modelling of the variables involved. The spillover effects of R&D on productivity⁶ would have similar implications in deriving the demand and hence price of R&D output as the nonlinear nature of value variables on the demand for fields in this article, but we cannot confirm a direct correspondence.

Table 4: ANOVA statistics on Cardinal model heteroskedasticity

ANOVA	df	SS	MS	F	Significance F
Regression	5	465483.8	93096.76	0.545249	0.741273
Residual	63	10756733	170741.8	—	—
Total	68	11222217	—	—	—

Figure 2: Plot of the residuals for the Cardinal model.

The Ordinal model (*dependant variable: rank of series*)

The estimated equation obtains as :

$$\begin{aligned}\hat{R} = & 29.55 + 0.04 \star \text{Jnl. Rk.} - 0.406 \star \text{Factor} \\ & + 0.039 \star \text{adjcit} + 0.002 \star \text{items} \\ & - 0.038 \star \text{allcit}\end{aligned}$$

(4.39) (1.57) (-0.302)
(0.532) (0.797)
(-0.53)

$\hat{R}$ = rank (closed) of series, Jnl. Rk. = journal rank, Factor = simple impact factor, adjcit = adjusted citations = all citations $\star$ Factor, items = number of items and allcit = all citations.

The *t*-statistic on all the variables are significant for all except Jnl.Rk. at 40 per cent level with Factor being significant at 76 per cent level. The regression statistics are presented in Table 5.

The ANOVA statistics on model heteroskedasticity are presented in Table 6.

The plot of the line fit with respect to the Factor variable is presented in Figure 3.

As this is a multifactor model, the residuals and model fits can be generated with respect to

each of the variables' marginal models. We present here the fitted graph only with respect to the Factor variable. The fit is reasonably good considering that we are only considering the journal with the highest simple impact factor for each field as an initial model. As there are two or three journals in many of the fields these statistics will definitely improve, it is conjectured, with the addition of other journals in the field and including their adjusted citations. The adjusted *F*-statistic is also reasonably high suggesting low heteroskedasticity of the model although a number of citation variables for the journals have been included.

The coefficient on the Factor variable is negative which implies that the ranking (closed) of the field in terms of the number of subscribers is negatively related with the simple impact factor of the corresponding journal; however, the rank is positively related with the journal rank, total number of items and adjusted citations according to the definition in this paper, which is the definition being used by RePEc. Hence, the ordinal ranking of demand for research output through the mechanism of subscription to RePEc fields is negatively related to the corresponding journal quality and hence value as measured by the simple impact factor but positively related to value as measured by journal rank, adjusted citations and number of items. Also, a comparison of the single factor model with the multifactor model reveals that the coefficient on the factor variable decreases and becomes negative by including the ordinal rank variable. This may mean that the ordinal

Table 5: Regression statistics for the Ordinal model

Regression statistics	
Multiple <i>R</i>	0.290461
<i>R</i> square	0.084367
Adjusted <i>R</i> square	0.011698
Standard error	19.94471
Observations	69

Table 6: ANOVA statistics for the ordinal model

ANOVA	df	SS	MS	<i>F</i>	Significance <i>F</i>
Regression	5	2309.138	461.8276	1.160979	0.338309
Residual	63	25060.86	397.7915	—	—
Total	68	27370	—	—	—

Figure 3: Plot of residuals for the Ordinal model.

measurement of RePEc series behaves in an opposite manner to the Cardinal measure. In calculating shadow prices for R&D expenditures for aggregation with other types of capital expenditures problems arise.⁶ This article identifies and therefore solves partially, another set of problems which has to do with the cardinal measurement *vis-a-vis* the ordinal measurement of R&D investment, done digitally. Ordinal measurement would treat demand for various R&D categories corresponding to various series in e-libraries according to relative ranking instead of absolute levels.

CONCLUSION

This article has discussed the nature of subscription to digital libraries consisting of primary research data in the form of working papers, and so on, organised into series by specialisations. Subscription has been equated to demand for the fields by analysing demand in Cardinal measurement as well as Ordinal measurement. Cardinal measurement has been identified by the number of subscribers, whereas Ordinal measurement has been identified by the rank of fields. This article finds that both measurements give rise to demand for digital research data, which is explainable by subsequent value created in the form of publication and citations in high quality journals for the respective fields.

Hence digital circulation of research papers, as measured by circulation through the RePEc digital library, gives rise to demand for research data which is explainable in terms of subsequent value to the user. However, the nature of the relationship is variable. First, the relations as have been identified here are weak in all cases. Second, a multifactor model which includes a number of value variables in terms of journal impact factor, number of citations, journal rank and total number of items performs better than a single factor model with only the factor variable. Third, the relation of the number of subscribers to the series (fields) to the impact factor of the relevant highest impact field journal is opposite to the relation of the rank of the series to the relevant impact factor. In terms of optimal policy, however, it has been identified that both the measures of demand for

the fields are sensitive to the subsequent value created in terms of citations of journal papers and hence the corresponding simple impact factors.

ACKNOWLEDGEMENTS

Mallick thanks the IISWBM Computer Centre for technical support. The authors thank B. Batiz-Lazo, L. Fernandes Suoto and the NEP team for helpful comments and for permission to draw upon prior work. The usual disclaimer applies.

REFERENCES AND NOTE

- 1 Batiz-Lazo, B. and Krichel, T. (2005) On-line Distribution of Working Papers through NEP: A Brief Business History. Economic History 0505002, EconWPA.
- 2 <http://repec.org/>
- 3 Allen, B. (1990) Costly acquisition of (differentiated) information. Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning about Knowledge. Pacific Grove, CA, USA, March 1990, 169–184.
- 4 Mallick, S.K. (1993) Bounded rationality & Arrow-Debreu economies. Unpublished PhD dissertation, Dept. of Economics, New York University, NY, USA.
- 5 Gossner, O., Hernandez, P. and Neyman, A. (2006) Optimal use of communication resources. *Econometrica* 74(6): 1603–1636.
- 6 Nadiri, M.I. and Mamuneas, T. (1994) Infrastructure and Public R&D Investments, and the Growth of Factor Productivity in US Manufacturing Industries. NBER w.p. no. W4845.
- 7 Mallick, S.K. (2006) Bounded rationality & incomplete markets. In: N. Yannelis (ed.) *Economic Theory*, Monograph Series, New York, NY: Springer Verlag, submitted.
- 8 Novarese, M. and Lanteri, A. (2007) Individual Learning: Theory Formation, and Feedback in a Complex Task. MPRA Paper 3049, University Library of Munich, Germany.
- 9 Hrbacek, K. and Jech, T. (1999) *Introduction to Set Theory*, 3rd edn. New York, NY: Marcel Dekker.
- 10 Kalaitzidakis, P., Mamuneas, T. and Stengos, T. (2003) Rankings of academic journals and institutions in economics. *Journal of the European Economic Association* 1(6): 1346–1366.
- 11 Mallick, S.K., Sarkar, A., Roy, K.K., Chakraborty, A. and Duttachaudhuri, T. (2007) A mathematical statistical pricing model for emerging stock markets. *Journal of Asset Management* 7(5): 335–346.