Developing a predictive model of editor selectivity in a current awareness service of a large digital library

Thomas Krichel [a], Nisa Bakkalbasi [b]

[a] *Palmer School of Library and Information Science, Long Island University, 720 Northern Boulevard, Brookville, New York 11548, USA*
[b] *Kline Science Library, Yale University, PO Box 208111, New Haven, Connecticut 06520, USA*

Abstract

"NEP: New Economics Papers," the current awareness service for the RePEc (Research Papers in Economics) digital library, is made possible by volunteer editors who filter new additions to RePEc into subject-specific reports. The official purpose of current awareness service is to filter working papers by subject matter without any judgment of its academic quality. In this article binary logistic regression analysis estimates the probability of a paper being included in any of the subject reports as a function of a range of observable values. The analysis suggests that, contrary to their claims, editors use quality criteria: the series the paper is coming from and the reputation of the authors. The findings suggest that a current awareness service can issue quality signals.

1. Introduction

RePEc (Research Papers in Economics) is a large digital library for economics research. Its roots go back to 1993, when Thomas Krichel started to collect information about downloadable electronic working papers in economics (Walshe, 2001). Working papers are accounts of recent research results before formal publication. Most economics departments in universities, as well as many other institutions that are involved in economics research (e.g., central banks and intergovernmental organizations), publish working papers. At the time of writing this article, over 360 archives based at institutions that issue working papers contribute to RePEc. They collectively provide 100,000 bibliographic records about the papers that they have published. Service providers periodically harvest that date and aggregate it to produce services for users who are interested in economics research[1]. One of the RePEc services, "NEP: New Economics Papers", is a human-mediated current awareness[2] service that was founded by John S. Irons and Krichel in 1998. NEP primarily operates through electronic mail. Volunteer editors from all corners of the globe, most of whom are PhD students or junior university

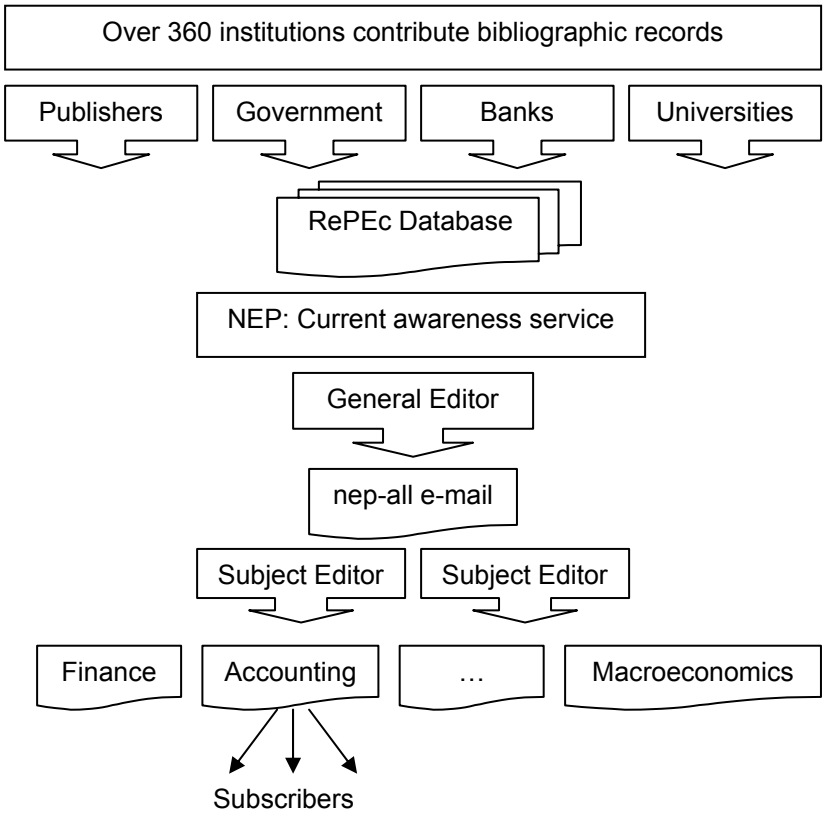
[1] The RePEc Web site at http://repec.org lists the service providers. All RePEc services are available to the public at no charge.
[2] The Dictionary of Library and Information Management (1997) defines current awareness service as "an organization or individual who notifies customers of the most up-to-date information in their field."

E-mail addresses: krichel@openlib.org (T. Krichel), nisa.bakkalbasi@yale.edu (N. Bakkalbasi)

faculty, operate NEP[3]. Applicants for editorship must demonstrate subject competence; usually this is done through their CVs. NEP administrative staff provides the operative training.

NEP has a simple, two-stage workflow. In the first stage, a general editor collects all new working paper data that have been submitted to RePEc in the previous week. He or she filters out records corresponding to papers that are new to RePEc, but are not new papers. Such records would typically come from new RePEc archives that add a whole back catalog of working papers to RePEc. The remaining records form a NEP report called nep-all. As its name suggests, it contains *all* the new working papers in RePEc from the previous week. Each issue of nep-all is circulated via e-mail to subject editors. This completes the first stage. In the second stage, subject editors filter every nep-all issue they receive to contain only papers in a certain subject area. When a subject editor has created a new subject-specific issue, it is circulated via e-mail to subscribers of the subject report. Figure 1 illustrates the NEP process.

Figure 1: NEP service process



Since its inception in 1998, NEP has grown in scale and scope. As RePEc has grown, so has the size of nep-all issues. This is scale growth[4]. On the other hand, over time, more subject reports have been created. This is scope growth. At the time of this

[3] NEP has a homepage at http://nep.repec.org.
[4] http://logec.repec.org shows monthly data for the evolution of the RePEc working paper stock.

writing, there are close to 60 distinct subject reports in NEP. Over 11,000 unique e-mail addresses have subscribed to at least one NEP report. Over 30,000 new papers have been included in nep-all.

Chu & Krichel (2003) find that NEP is an innovative service model for digital libraries. However, questions about its long-term sustainability, as a volunteer service, remain. Barrueco Cruz, Krichel, & Trinidad (2003) present a simple empirical assessment of the NEP service. One of the issues they look at is the subject coverage of NEP. Does NEP cover all the subjects that are found in RePEc? If it does, then each working paper in a nep-all issue appears in at least one subject report. Empirically, this conjecture can be examined by looking at the coverage ratio. This is the ratio of papers in nep-all that have been announced in at least one NEP subject report. As more subject reports have been added, the coverage ratio ought to improve over time, and, in the long run, reach 100 percent. Surprisingly, the data reported by Barrueco Cruz, Krichel, & Trinidad (2003) suggest that the coverage ratio has not been improving over time, and that it certainly remains well below full coverage. Currently, the coverage ratio stands at 78 percent.

Full coverage is a desirable goal. However, the implications of steps to achieve better coverage have to be carefully considered.. The first idea that comes to mind is that one can open more subject-specific reports. However each additional report raises coordination costs. Therefore it is prudent to formally investigate apparent reasons behind the lack of full coverage of NEP, before deciding that insufficient number of subject categories is the main reason. This article formally investigates subject editors' behavior, specifically to examine what makes a paper "announceable" in a NEP report. The remainder of the article is organized as follows. Section two presents a conceptual framework, whereas Section three describes the methodology for developing a predictive model and Section four discusses the data set. Section five presents the findings and Section six develops the conclusions and suggests future work.

2. Conceptual Framework

This article introduces two basic theories about editor behavior that aim to explain the static nature of the NEP coverage ratio. They are the "target theory" and the "quality theory," respectively.

The target theory starts with the observation (Barrueco Cruz, Krichel & Trinidad, 2003) that the size of nep-all issues has been highly volatile in the short run, and has been steadily growing in the long run. The theory suggests that, when composing an issue of a subject report, the editors have an implicit issue size in mind. Therefore, if the size of nep-all is large, they will take a narrow interpretation of the subject matter of the report (i.e., they will be choosier as to what papers they include). Thus, the target theory claims that the observed long-run static nature of the coverage ratio comes from the simultaneous effect of scale and scope growth of NEP. Scale growth, all other effects being equal, will reduce the coverage ratio. Scope growth, all other effects being equal, will increase the coverage ratio. The long-run static coverage ratio is the result of both effects canceling out each other.

The quality theory suggests that the subject editors filter for paper quality. There are two types of quality indicators. First, there is the descriptive quality of the record that describes a paper. Some papers are poorly described, they have a meaningless title,

and/or no abstract. Second, there is the substantive quality of the paper itself. The paper may be written by authors whom nobody has ever heard of, and/or who are based at institutions with an unenviable research reputation. Whether it is substantive or descriptive, the quality of a paper is likely to be important when it comes to its inclusion in any NEP report.

An e-mail discussion on the private mailing list used by the subject editors has revealed that editors have a uniform view of the quality theory: they reject it. They claim that they perform their work independent of quality considerations. They assert that their only concern is to disseminate new working papers based on the subject matter. Furthermore, they specifically insist that NEP cannot be regarded as a vehicle for a preliminary peer-review. Here are some sample comments

> *"I only filter papers on the basis of content and relevance to the topic of my list. This sometimes involves looking at the paper itself, although most of the times the decision whether to include a paper or not is based on the description given in the abstract. To me NEP is a dissemination service and as such editors are not there to make 'quality' judgments."*
>
> *Subject Editor, Economic Geography*

> *"As an editor of NEP-AFR (articles concerning Africa), I can say that in general I try not to filter based on "quality" as I am especially sensitive to disseminating papers that come from African scholars themselves and are often from research institutions or universities which may not have prestigious reputations (or authors who are unknown)."*
>
> *Subject Editor, Africa*

> *"There is no quality editing on my side (NEP-DGE) unless there is a report with an unusually large number of relevant items, which is rare."*
>
> *Subject Editor, Dynamic General Equilibrium*

If the general assessment of the editors is wrong, then NEP may be considered as a first stage in an alternative peer-review system. Such a system may be constructed as an extended service, sitting on top of NEP. To date, RePEc does not engage in peer-review other than vetting the providers of archives. However, in the long run, the idea of quality review through NEP could start to change that.

The debate between the two theories also has some short-run stakes for the running of NEP itself. If the target theory is correct, then opening additional specialized report categories should be considered as a way to improve the coverage of NEP. If the quality theory is correct, opening additional report categories will have no effect on coverage. This question of whether to open more reports or not has been one of the important motivations for the research conducted in this article.

3. Research Methodology

This article assesses both the target and the quality theories and explores if either one or both can be confirmed. A simple way to assess the target theory empirically is to see if

the coverage ratio declines with the size of a nep-all issue. Barrueco Cruz, Krichel & Trinidad (2003) have a cross-sectional plot of coverage ratio versus size of nep-all. The shape of the plot suggests that this seems to be the case. However, they only provide descriptive statistics. Even if inferential statistics were used, it would only look at one aspect of the selectivity issue. This article attempts to build an overall model that combines a set of variables that may influence the probability of a working paper being "announced" in any report. These variables are:

- Size of the nep-all issue in which the paper appeared;
- Length of a title of the paper;
- Presence/absence of an abstract to the paper;
- Inclusion of a paper in a series; and
- Prolificacy of the authors of the paper.

The first variable assesses the target theory. The others assess the quality theory. The length of the title and the presence of the abstract are indicators of the influence of the descriptive quality of the paper. The inclusion of the paper in a series and prolificacy of the author of the paper are indicators of the substantive quality of the paper. The idea is that some working paper series publish better papers than others and better papers are more likely to be announced. In a similar way, authors who write more papers are usually better known.

The statistical hypothesis is that a percent of the variance in the response variable (i.e., the presence/absence of a paper in any report) can be accounted by predictor variables. If the hypothesis turns out to be correct, a prediction equation will allow predicting the probability of a working paper being included in a NEP report. The most appropriate statistical method for analyzing this relationship is Binary Logistic Regression Analysis (BLRA). There are three reasons for this choice. First, the dependent variable is dichotomous that can suitably be coded with values of 0 and 1. Second, the independent variables are both quantitative and qualitative in nature. Last and most important, BLRA is a flexible technique. BLRA does not require any of the following assumptions commonly made for linear regression analysis to work:

- Linear relationship between the independent variables and the dependent variable;
- Homoscedasticity of the dependent variable for each level of independent variables;
- Normally distributed dependent variable; and
- Normally distributed error terms.

In a review of statistical techniques in library and information sciences, Bensman (2001) suggests that, because of the highly skewed probability distributions observed in library and information science, the researcher should look at the biomedical sciences for methodologies used to attack these issues. According to Hosmer & Lemeshow (2000), BRLA originated in the epidemiological research and it is now heavily used in biomedical research. Its use in library and information science has not been widespread.

4. Data Set

The data set has been extracted from historic email message archives that contain NEP report issues and bibliographic records for the papers referred in the report issues. The data, which go back to the inception of NEP in 1998, contain 32,892 records, with each

record corresponding to a paper that has appeared in nep-all. The response variable ANNOUNCED, takes the value 1 if the paper was announced, and 0 if not.

Table 1
The variables identified for exploration

Description	Values	Variable Name
announcement of paper	1 = yes, 0 = no	ANNOUNCED
nep-all size	3 to 803	SIZE
number of characters including space in the title	3 to 1945	TITLE
presence/absence of an abstract	1 = yes, 0 = no	ABSTRACT
average announcement ratio of series	0 to 5.5	SERIES
number of papers the lead author submitted to RePEc archives previously	1 to 284	AUTHOR

The candidate predictor variables are as follows. SIZE corresponds to the size of a nep-all issue and is easily tracked. TITLE is the length of the title and ABSTRACT is the presence/absence of an abstract. Both come from the bibliographic record of the working paper. In the very rare cases where the bibliographic information is not available, the record is dropped. SERIES is the number of different subject-specific reports a series appears in. It is a quantitative variable that measures the ratio of the total number of times working papers from a specific series have been announced in subject reports, divided by the total number of working papers from that series that appeared in nep-all. This gives an overall indication of how well-respected and visible a series is. AUTHOR, the measure of prolificacy of an author, is the most difficult variable to construct. There are at least three problems with constructing such a measure. First, RePEc does not cover the entire economics discipline. Second, it is not easy to know if two similar author names represent the same person. Third, since co-authorship is frequent in economics, one needs to decide how to aggregate the prolificacy of individual authors. To deal with the second problem, RePEc runs an author registration service (*http://authors.repec.org*), which collects records about the authors and the papers they have written. Authors contact the service to build their own electronic CVs. Such registration is voluntary, of course. Many papers in RePEc do not have identified authors. One measure that allows the retention of the most records is the "lead author prolificacy" (LAP). For each paper in the data set, the LAP is the number of papers in the RePEc database of the registered author with the largest number of papers. Still, due to a high number of unregistered authors, a significant number of records are missing. After removing these records, 10,652 records remain. Since author registration and appearance of papers in reports are independent events, the removal of a large number of records introduces no bias. The descriptive statistics of both the original data set and the smaller data set are similar. Table 3 shows that the averages and standard deviations of the other independent variables stay approximately the same after the removal of about two thirds of the records. The size of the remaining data is still amply sufficient to conduct the analysis. Table 2 shows a few sample records from the data set before the removal of the records with missing values. In Table 2, HANDLE corresponds to the unique identifier for each record in the data set.

Table 2
Data set sample

HANDLE	ANNOUNCED	SIZE	TITLE	ABSTRACT	SERIES	AUTHOR
RePEc:jku:econwp:2001_05	0	230	88	1	1.083	NA
RePEc:nbr:nberwo:9361	1	175	42	1	1.619	NA
RePEc:fip:fedfap:2002-02	1	433	54	1	1.519	95
RePEc:wop:wobaiy:2957	0	803	66	0	1.917	NA
RePEc:cbr:cbrwps:wp207	1	433	74	1	1.405	NA

All our calculations use the R language and environment (see http://www.r-project.org/). The computer code and the data set are available on request.

5. Findings

5.1. Exploratory Data Analysis

A frequency count of the response variable ANNOUNCED, shows that 2,373 papers are not announced and 8,279 papers are announced. This implies a coverage ratio of approximately 78 percent.

Table 3
Descriptive statistics for quantitative predictor variables

	SIZE	TITLE	SERIES	AUTHOR
Minimum	3.0	3.0	0.000	1.0
1st quartile	125.0	48.0	1.245	11.0
Median	202.0	63.0	1.480	26.0
Mean	240.4 [253.7[b]]	66.2 [68.4[b]]	1.450 [1.388[b]]	40.5
Standard deviation	160.5 [172.8[b]]	32.0 [30.6[b]]	0.421 [0.442[b]]	42.3
Coefficient of variation [a]	0.668	0.483	0.290	1.044
3rd quartile	306.0	82.0	1.619	55.0
Maximum	803.0	1945.0	5.500	284.0

[a] Coefficient of Variation = Standard Deviation / Mean
[b] Prior to removing the missing values

Table 3 displays the descriptive statistics for the quantitative predictor variables. It does not include the qualitative variable ABSTRACT. Initial intuition suggests constructing ABSTRACT as the number of characters in each abstract. However, proceeding in that way, we encounter a wide range of values [0, 11295], with the value 0 occurring very frequently. Conventional measures of central tendency and variance are meaningless in this context. Therefore we make ABSTRACT a categorical variable and encode it as 0 (no abstract) and 1 (has abstract) within each record. This is a common procedure used in statistical analysis to remove the large variation in a predictor variable while maintaining the functional relationship between the response variable and the predictor variable (Hosmer and Lemeshow, 2000).

For each quantitative predictor variable, there are different measures of central tendency (i.e., mean and median) and dispersion (i.e., standard deviation, coefficient of variation, range). Two quantitative predictor variables, TITLE and SERIES, have their mean

and median close to each other and their variation is small, implying a symmetrical distribution. For these two variables, the mean is an appropriate measure to determine their typical values for observation. Therefore, a typical working paper title contains 66 characters with a standard deviation of 32 and the average announcement ratio for a series is 1.45 with a standard deviation of 0.421.

For the variables AUTHOR and SIZE, there are seemingly significant differences between the mean and the median. The dispersion measures for those two variables indicate a high variation, due to the frequency of extreme values. To illustrate, there are many authors who have written three or four papers. Only one author (Nobel laureate Joseph E. Stiglitz) has 284 papers. Because he appears as a prolific author on so many papers, he introduces an upward distortion for the average number of papers that an author has written. The same scenario is valid for SIZE. The mean is highly influenced by extreme values. Therefore, for these two variables, the median qualifies as a more appropriate measure of central tendency than the mean. Therefore, a typical lead author has written 26 papers, and that the typical nep-all issue contains 202 working papers.

5.2. Outlier analysis

Visual examination of the descriptive statistics shows some questionable observations. For example, there appears a paper title with as few as three characters and another with as many as 1945 characters. One nep-all issue containing 803 working papers immediately raises eyebrows. To detect striking deviations as potential outliers, we carefully examine lists of data points with values greater than three standard deviations from the mean. According to the empirical rule (Freund & Wilson, 2003), the interval $(\bar{y} \pm 3s)$, where $\bar{y}$ is the mean of a variable and s is its standard deviation, contains virtually all the observations if the shape of the distribution is nearly bell-shaped. Although the empirical rule furnishes us with a practical way of obtaining potential outliers, it does not appear to work well for variables that are not bell-shaped.

In general, dealing with outliers is difficult and a matter of judgment. In some cases the outliers are legitimate but extraordinary occurrences, whereas in other cases, they are likely to be errors in the data. All suspicious observations require looking at the original record and correcting some records. For example most of the titles with less than 10 characters turn out to be acronyms for instructional data sets. In another case, a paper with 1945 characters in the title field turns out to have a poorly formatted bibliographic record, its abstract appears in the title field. These should not have appeared in nep-all issues at all. We drop the erroneous occurrences that cannot be corrected and keep the extraordinary occurrences. At the end of this process, 86 outliers are dropped. This allows 10,566 records to be used for building the binary logistics model.

5.3. Inferential Data Analysis

Pair-wise correlations among the four predictor variables show that there is no significant correlation among any of the pairs of the predictor variables. Lack of correlation among the predictor variables increases the confidence that there is a higher likelihood for each predictor variable to contribute to the final prediction equation independently.

Table 4
Pearson correlation test among the four quantitative predictor variables

	Correlation	t-statistic	p-value	95 % CI
SIZE - TITLE	0.007	0.790	0.429	[-0.011, 0.0268]
SIZE - SERIES	-0.107	-11.02	0.000	[-0.125, -0.0877]
SIZE - AUTHOR	0.015	1.522	0.138	[-0.043, -0.004]
TITLE - SERIES	0.015	1.590	0.112	[-0.004, 0.034]
TITLE - AUTHOR	-0.083	-8.563	0.000	[-0.102, -0.064]
SERIES - AUTHOR	0.015	1.522	0.128	[-0.004, 0.034]

The "Design Library of Modeling Functions" in R allows building the regression model.

5.3.1 Fitting and testing the model

The results of the logistic regression analysis are shown in Table 5.

Table 5
Estimated coefficients for multiple logistic regression model

	Coefficient	S.E.	Wald	P
Intercept	–1.1202	0.1268499	–8.83	0.0000
SIZE	–0.0008	0.0001454	–5.61	0.0000
TITLE	0.0038	0.0009651	3.89	0.0000
ABSTRACT	0.3067	0.0634233	4.84	0.0001
SERIES	1.4434	0.0696371	20.73	0.0000
AUTHOR	0.0025	0.0006381	3.87	0.0001
	Model χ_2	d.f.	p-value	
	690.94	5	0	

A likelihood ratio test for the overall significance of the five coefficients for the independent variables assesses if the model as a whole is significant

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_A : At least one coefficient is not equal to 0

where β_i is the coefficient for each predictor variable, respectively.

The likelihood ratio test statistic takes the value $G_M = 690.94$, where G_M is referred to as the Model χ^2. It rejects the null hypothesis at virtually any significance level. Therefore at least one of the five coefficients is different from zero and that, together, SIZE, TITLE, ABSTRACT, SERIES, and AUTHOR, are significant predictors of ANNOUNCED.

The Wald statistics for the individual coefficients tests the significance of the variables in the model.

$H_0 : \beta_i = 0$

$H_A : \beta i \neq 0$

The Wald test statistics W are the ratio of each coefficient to its standard error.

$W_i = \dfrac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)}$, where β_i is the coefficient for each predictor variable

Based on the evidence contained in the data, at a significance level of $\alpha = 0.05$, the Wald statistics reject the null hypothesis for each of the five coefficients and conclude that each of the predictor variables is significant. Table 5 contains the details.

Running different models using different subsets of the predictor variables reveals, after a thorough comparison of various models, that none of the candidate variables can be excluded from the final model. Let x be a vector representing values of the predictor variables. Then, the final logistic regression model, which gives the estimated logistic probability, is

$$(1) \qquad P(\text{ANNOUNCED} = 1 \mid x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}},$$

where the estimated logit is

$$(2) \qquad \hat{g}(x) = -1.1202 - 0.0008*\text{SIZE} + 0.0038*\text{TITLE} + 0.3067*\text{ABSTRACT} + 1.4434*\text{SERIES} + 0.0025*\text{AUTHOR}$$

5.3.2 Interpreting the fitted logistic regression model

Equation (1) gives a probability for the event occurring given all the values of the predictors. Equation (2) looks like a linear regression model as it is commonly understood. In such a linear regression equation, the coefficients are interpreted as the rate of change in the dependent variable associated with one-unit change in the respective independent variables. In the logistic regression model, however, the slope coefficient represents the change in the logit corresponding to a change of one unit in the independent variable. Therefore the relationship between the independent and the dependent variable is less intuitive. One commonly used measure of association is called the odds ratio, commonly abbreviated as OR. Roughly speaking, OR is a measure of the degree of association between each predictor variable and outcome. It is obtained by transforming the estimated coefficients. There are different ways of expressing odds ratio depending on the different types of predictor variables in the model, i.e., dichotomous, polychotomous, or continuous.

Table 6 contains the estimated OR values for the predictor variables. Interpretation of the categorical variable ABSTRACT is pretty straightforward. The OR for the ABSTRACT coefficient is 1.36 with a 95% confidence interval of [1.20, 1.54]. This suggests that in the presence of an abstract, a working paper is 1.36 times more likely to be included in at least one subject category than in the absence of an abstract. For the continuous variables, creating intervals allows the observation of the impact of "c" units of change in the independent variable as opposed to one-unit of change, which does not offer any practical inference for a continuous. The three intervals are the following:

1. 1st quartile
2. 2nd and 3rd quartiles combined
3. 4th quartile

Table 6
Estimated odds ratios and 95% confidence intervals for predictor variables

	Interval	Difference (c)	Odds Ratio (OR)	95% CI over OR
SIZE	3 – 125	122	0.91	[0.87, 0.94]
	125 – 306	181	0.86	[0.82, 0.91]
	306 – 803	497	0.67	[0.58, 0.77]
TITLE	4 – 48	44	1.18	[1.09, 1.28]
	48 – 82	34	1.14	[1.07, 1.21]
	82 – 218	136	1.67	[1.29, 2.15]
ABSTRACT	N/A	N/A	1.36	[1.20, 1.54]
SERIES	0 – 1.25	1.25	6.08	[5.12, 7.21]
	1.25 – 1.62	0.39	1.70	[1.62, 1.79]
	1.62 – 5.5	3.88	270.91	[159.51, 460.12]
AUTHOR	1 – 11	10	1.02	[1.01, 1.04]
	11 – 54	43	1.70	[1.62, 1.79]
	54 – 284	230	1.76	[1.32, 2.35]

The estimated OR for SIZE suggests that, in the first interval, where an increase of 122 papers occurs, the odds of a paper being announced in at least one subject category increases 0.91 times. In other words, an estimated OR of approximately 1 indicates that a working paper with a 122 increase in SIZE is equally likely to be announced or not to be announced. As it is shown in Table 6, the odds ratio reduces as the difference in SIZE increases for the next two intervals. The significant drop in the odds ratio for the third interval, where the increase in SIZE is 497, indicates that working papers from large nep-all issues are less likely to be included in subject-specific reports. The OR results for the TITLE variable show that the odds for being "announced" increase as the title gets longer. More specifically, an increase of 136 characters in the title increases the odds of a working paper being included in a subject-specific report 1.67 times. Similarly, the OR estimates for SERIES and AUTHOR suggest that, as the respective "c" units of change increases, there are significant corresponding increases in the likelihood of a working paper being included in at least one subject-specific report.

6. Conclusions

100 percent coverage appears as a desirable goal for a current awareness service. This article provides quantitative evidence that, a setup like NEP in which each individual editor acts autonomously as to what papers to include in a report, it is not likely to achieve 100 percent coverage. There is quantitative evidence that editors also filter for quality of papers. Both the substantive quality of the paper (i.e., the fame of the author and the appearance in a working paper series of great renown) as well as the quality of the descriptive record, (i.e., the presence of an abstract and the length of the title) are influencing the selection decision. The quality theory about editor behavior therefore cannot be dismissed. No matter how many reports there are, some papers will remain unannounced.

This article pioneered the use of BLRA to the study of editor selectivity in a current awareness service. The BLRA technique is suitable for building a quantitative

model of editor selectivity. It is likely that the technique can also be used to make prediction about individual subject reports. Future research intends to examine each new nep-all issue automatically and develop a forecast for each subject editor regarding the inclusion of a given working paper in a specific subject area. This will make it easier for the editors to scrutinize the new working papers. The ultimate aim would be a recursive system where each new forecast is based on the evidence of the previous editorial judgment. Such a system will undoubtedly make the work of the subject editors easier, and keep NEP on a path of sustainability.

References

Barrueco Cruz, J. M., Krichel, T., & Trinidad, J. C. (2003). *Organizing current awareness in a large digital library*. Presented at the 2003 Conference on Users in Electronic Information Environments in Espoo, Finland, September 8-9, 2003, http://openlib.org/home/ krichel/papers/espoo.pdf.

Bensman, S. J. (2000). Probability distributions in Library and Information Science: A historical and practitioner viewpoint. *Journal of the American Society for Information Science and Technology,* 51, 16-833.

Chu, H., & Krichel T. (2003). NEP: Current awareness service of the RePEc Digital Library. *D-Lib Magazine, 9*(12). http://www.dlib.org/ dlib/december03/chu/12chu.html

Dictionary of Library and Information Management, Peter Collin Publishing (1997). Retrieved 01 March 2005, from xreferplus. http://www.xreferplus.com/entry.jsp?xrefid=1039558&secid=.2.-

Freund, R. J. & Wilson, W. J. (2003). Statistical Methods. (2nd ed.). New York: Academic Press.

Hosmer, A. W., & Lemeshow S. (2000). *Applied logistic regression*. New York: Wiley

Walshe, E. (2001) *Creating an academic self-documentation system through digital library interoperability: the RePEc model*", The New Review of Information Networking, 7, 43-58.

Acknowledgements

We are grateful to Stephen J. Bensman, Louisiana State University, and Sune Karlsson, Stockholm School of Economics, for helpful comments on an earlier version of this paper.

Thomas Krichel is an Assistant Professor in Palmer School of Library and Information at Long Island University. He studied Economics and Social Sciences at the universities of Toulouse, Paris, Exeter and Leicester. Between February 1993 and April 2001 he lectured in the Department of Economics at the University of Surrey. In 1993 he founded NetEc, a consortium of internet projects for academic economists. In 1997, he founded the RePEc data set to document Economics. His main area of work is the development open digital libraries for scholarly communication. His homepage is http://openlib.org/home/krichel.

Nisa Bakkalbasi is the General Science Librarian in Kline Science Library at Yale University. Her primary responsibilities include reference service, instruction program coordinator, preservation and conservation liaison, and Web management. She holds an M.L.I.S. from Long Island University and an M.S. in Applied Statistics from University of Alabama. Prior to joining Yale University, Nisa worked as a Science/Electronic Resources Librarian at Purchase College, State University of New York.