Modeling Journal Bibliometrics to Predict Downloads and Inform Purchase Decisions at University Research Libraries

Daniel M. Coughlin and Bernard J. Jansen
College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802. E-mail: dmc186@psu.edu; jjansen@ist.psu.edu

University libraries provide access to thousands of online journals and other content, spending millions of dollars annually on these electronic resources. Providing access to these online resources is costly, and it is difficult both to analyze the value of this content to the institution and to discern those journals that comparatively provide more value. In this research, we examine 1,510 journals from a large research university library, representing more than 40% of the university's annual subscription cost for electronic resources at the time of the study. We utilize a web analytics approach for the creation of a linear regression model to predict usage among these journals. We categorize metrics into two classes: global (journal focused) and local (institution dependent). Using 275 journals for our training set, our analysis shows that a combination of global and local metrics creates the strongest model for predicting full-text downloads. Our linear regression model has an accuracy of more than 80% in predicting downloads for the 1,235 journals in our test set. The implications of the findings are that university libraries that use local metrics have better insight into the value of a journal and therefore more efficient cost content management.

Received September 2, 2014; revised March 9, 2015; accepted March 10, 2015

 DOI: 10.1002/asi.23549

Introduction

University libraries at large research institutions have increased the percentage of their budget allocated to electronic resources (Kyrillidou, M., Morris, S., & Roebuck, 2013). As physical storage space has become less of an obstacle, the number of electronic titles provided can exceed 100,000, and the annual dollars spent can exceed tens of millions (Furlough, 2012). The increase in spending and the limitation in budget have created an environment within university collection management that requires an understanding of the value of the electronic content. Leveraging various statistics and metrics, libraries attempt to improve their ability to measure the usage of resources; however, most of these measurements are descriptive (Coughlin, Campbell, & Jansen, 2013). If libraries could combine the descriptive approach with other statistics to create a model that would predict value, universities could be better equipped to enter negotiations with resource providers, understand how much they would be willing to pay for a resource, debate financing, and defend budgets. Developing this predictive ability for university libraries is the focus of this research.

Publishers provide journal metrics based on citations to indicate some level of value for the journal. We call these metrics *global* in the sense that they measure for the entire distribution of a journal across many institutions. The assumption these metrics make is that what is valued globally is reflective of what is valued locally at an individual institution. This unfounded assumption that global reflects local is an oversimplification of which resources will provide value to an institution.

In addition to global metrics, we propose the development of *local* metrics, which are measures concerning online content dependent only on the institution being evaluated. To our knowledge, there has been limited use of local

metrics on any systemic scale beyond simple counts, such as downloads. This dearth of local metrics exists for a reason—it has been extremely complicated to determine value at the local level for online content. Historically, the library has been a steward for aggregating content from many sources to create a collection of knowledge. Many of the numbers that exist to create local metrics are stored in disparate systems (i.e., Web of Science, *Journal Citation Reports* [*JCR*], Journal Reports 1 [JR1], financial databases) and in different formats (i.e., spreadsheets, SQL databases, PDF documents). A similar problem exists to measure the value within these resources because the information is from multiple sources and formats and requires a great deal of curation to understand the value a journal has at a local institution. The difficulty, scale, and multiple digital formats have created a complex environment for aggregating these data locally, and this issue currently remains unsolved.

The goal of this research is to advance the knowledge base within collection management and create a model to predict journal downloads at an individual university library using a combination of global and local metrics in order to identify the value of journals prior to a purchase or other content management decisions. This paper examines global and local metrics to identify those that have a strong correlation with local article download numbers. By providing a prediction for downloads, we can provide a range of value for that content that the library may then use to make decisions such as the price the institution may be willing to pay. A more comprehensive understanding of the metrics that are strongly linked to local downloads will provide insight to the measures that should be more widely considered when evaluating journals and determining their value at an institution.

Literature Review

Importance of Evaluating the Value of Journals

Prior literature has numerous discussions about the importance of evaluating the significance of electronic journals (cf. Gallagher, Bauer, & Dollar, 2005; Metz & Cosgriff, 2000). Metz and Cosgriff (2000) discuss a process of identifying titles of value to a library community using surveys, which could inform collection development decisions. Gallagher et al. (2005) employed data from multiple sources to inform decisions about which print journals were targets for cancellation. Although there may be secondary motivations (e.g., library budgeting and collections prioritization, marketing of library value to their parent institution), the fundamental reason for this evaluation at a research library is to make sure the needs of the library's researchers are being met. There are a number of methods of extracting meaning from the global metrics provided by commercial companies to libraries, and algorithms using authorship, citations, or download numbers to create metrics to attempt to measure journal value. Effective collection management practices require the ability to look beyond just the numbers presented by these metrics in order to define the meaning to fully understand the strengths and weaknesses of these metrics.

Evaluating Journal Value Via Content, Cost, and Usage

Content. The need to have a solid foundation of metrics to assist in collection management has grown increasingly important in the information age in order to quantify the content to which an institution has or should have access. Not only has the amount of information increased, but the accessibility and availability of scientific journals has increased. There are more peer-reviewed articles, unpublished manuscripts, conference papers, and so on, and this deluge of documents emphasizes the need for key performance indicators (KPI) to substantiate the importance of online resources and to prevent the overabundance of information (Oosthuizen & Fenton, 2014). KPIs "measure performance based on articulated goals" (Jansen, 2009, p. 3). There is an apprehension that lacking higher-quality standards in electronic resources leads to a proliferation of inferior work, as untrained researchers use, cite, and mimic lesser works (Bartsch & Tydlacka, 2003).

Today's digital ecosystem relies on outside vendors and publishers to help provide metrics for evaluation. Currently, there is no standard method or set of metrics existing for libraries to systematically or effectively make strategic recommendations on electronic resource decisions within collection management (Lakos, 2007), although there are products that assist in measuring an institution's content collection impact (e.g., Elsevier's SciVal, Thomson Reuters Essential Science Indicators).

Cost. Analysis of electronic resources related to cost considers the matter of budgets, the source of the money to fund resources, price increases, and the merit for the price increases (Miller-Francisco, 2003). The evaluation of electronic resources and the significance of providing access to them has also created alternative methods for access, such as Interlibrary loan (ILL) on demand rather than an annual subscription fee (Leon & Kress, 2012). Additionally, the necessity of reducing and effectively managing content is compelling because of the conditions surrounding library budgets and the scrutiny that they are under.

Usage. Aside from content and cost, a third attribute of evaluation for collection management is usage. Usage is usually measured by successful full-text downloads of an article from a specific journal and/or provider. Although not a panacea for assessment, downloads remain a popular metric because of their simplicity. Therefore, they are a good surrogate for usage. Additionally, downloads can be combined with other metrics, like cost or content, to create more valuable KPIs that can further assist in regulating journals within collection management. Having a broad understanding of the journals researchers within an institution are accessing via downloads indicates a perceived level of value

DOI: 10.1002/asi

and can have a significant impact on the collection management process (Medeiros, 2007).

However, there are some potential biases that exist when evaluating journals solely on full-text downloads. For instance, some academic domains reference articles in their publications more often than others. Subjects that require more references for published articles will naturally receive more article downloads in order to produce those references. This can skew the number of downloads when comparing journals from different subject areas. New standards and metrics are being considered to improve some of the inherent problems with usage. One COUNTER standard, released March 2014, is Usage Factor. Taking the median usage for a journal over a 2-year period and normalizing that usage based on the total number of published items online at that time calculates Usage Factor. Using the median, instead of the mean, limits outliers caused by a single popular article, which can create an apparent spike in usage (Pesch, 2012). Nevertheless, downloads, have been an important benchmark to measure demand and the value of electronic resources (Metz, 1992).

Evaluating Journal Value Via Global Metrics

Bibliometrics are used to measure and analyze primarily citations, which in turn facilitate the evaluation of scholars (cf. Minasny, Hartemink, & McBratney, 2007) and journals. There is an inherent assumption in bibliometrics that a citation is a sign of meaningful work; a citation indicates that an article has not only been read but also considered useful enough to cite. However, citations can be problematic due to normalization issues. These problems can include conferences or journals requiring (silently or otherwise) citations from their own publication in order for articles to be published, and/or journals with a large number of articles likely to have more citations in a year than a journal with fewer articles. Journals among different fields have different citation rates. Additional problems regarding citations are that they do not tell the whole story. For example, someone may read an article that leads to a different article that is eventually cited. However, the discovery of the initial article, that was not cited, has no value attributed to it when using citation alone as the metric of value. Furthermore, the context of a citation is not available when judging its impact. For instance, an article may be cited as a way to show what is wrong with regard to a particular method or analysis and yet the citation is considered of equal value with an article that is cited as exemplar. By evaluating these metrics further and by analyzing previous research where these metrics have been used, we can understand their existing strengths and weaknesses and how they can be leveraged together to create more meaningful metrics.

The results of these evaluations and prior works make it clear that global metrics are useful, and these global metrics could provide further benefits if combined with institutional (i.e., local) metrics (Arnold & Fowler, 2011; Coughlin et al., 2013; Garfield, 2006). Journal rankings suppress some widely accepted and useful measurements for performance, which indicates the need for a more broadly focused investigation to inform management of these resources (Rafols, Leydesdorff, O'Hare, Nightingale, & Stirling, 2012). It is necessary to create a reliable, open, and interoperable network of this information to promote new and enhanced metrics for evaluating scientific journals (Lane, 2010). We believe that enhanced evaluation, based on combining global and local metrics, will help collection management better evaluate journals both at the necessary scale for today's digital age and with the efficiency that is required in a time of rising costs. Based on this review of the literature and obvious need, we present the following research objectives.

Research Objective

The research objective is to predict the usage of a journal at an institution using global and local metrics to inform institutional decisions concerning the value of an electronic resource and guide collection management decisions.

This research objective quantitatively measures the correlation between journal metrics and the number of downloads for that particular journal at a given institution. If there is a correlation between these metrics and full-text downloads, then this correlation will allow library collection managers to predict the value of a particular journal prior to purchase based on an acceptable cost-per-download model.

The global measures such as impact factor, eigenfactor, article influence score, and so on, come from *JCR*s, and because of this, it is important to understand which of these metrics has the strongest correlation to local downloads. Determining local metrics can require a great deal of work. Therefore, a full understanding of the correlation between local metrics and future journal downloads can help inform how much effort should be put into generating these local metrics and which ones should be generated.

By creating regression models to define the correlation between full-text downloads and other independent variables, it is possible to see the independent variable's impact in this correlation. Specifically, we create a model that includes both global metrics (i.e., total citations, impact factor, total articles, etc.) and local metrics (e.g., internal citation metric—number of citations by local institution authors for articles published in a given year, and annual subscription cost). Thus, can we create statistically significant stronger models by combining metrics from multiple sources that have influence both globally and locally than just creating models from one data source? We posit the strongest models are using both global metrics and local metrics.

The research institution being analyzed had purchased 3,400 line items in 2012 that cost more than $10 million and of those 3,400 line items, there are 1,510 items that represent journals with all the global metrics we are using

DOI: 10.1002/asi

for evaluation. For example, not every journal has a record in a *JCR* and subsequently does not have an impact factor, article influence score, and so on, so, we could not evaluate every journal purchased in 2012. However, we have a significant portion of the annual financial spending and full-text downloads to make these data meaningful. The 1,510 journals with full data available represent over 40% of the annual subscription fees for this library and more than 1.5 million downloads in 2012. To create our regression models, we used a random sample of 275 of the 1,510 journals for the training data. Once the regression model was created, we applied this model to remaining 1,235 journals to see how well this model predicted downloads on a real-world test data set.

Methods, Reports, Data Preparation, and Analysis

Methods

The approach used here is web analytics, which is the measurement, collection, analysis, and reporting of Internet data for the purposes of understanding and optimizing web usage (Jansen & Rieh, 2010). Although web analytics cannot directly measure motivation or satisfaction, it provides a method to evaluate online behaviors, correlate usage with other available data, and standardize information across data sets and can inform financial decisions (Ortiz-Cordova & Jansen, 2012). One of the contributions from this research is the correlation of many advanced global metrics with local metrics to predict downloads of journal articles at a local institution. Understanding the relationship these advanced metrics have on downloads will assist in determining the demand of these journals at an institution and will aid collection management practices in understanding the local value of journals (Carroll, 2009).

The statistical model used in this research is multiple linear regression to create the strongest model for predicting downloads at an institution with the smallest number of independent variables. Multiple linear regression is a statistical analysis that allows multiple factors (i.e., independent variables), which in this case are local citation counts, Eigenfactor, impact factor, article influence score, cited half-life, immediacy index, 5-year impact factor, total articles, and global citations, to be considered to estimate the effect of the independent variables on the dependent variable, in this case, full-text downloads (Sykes, 1993).

Data, Metrics, and Reports

The significance of journal evaluation has led to the creation of several reports and scoring metrics to assist in understanding journal rankings. In this study, data are coalesced from four different types of reports to further examine journal value: *Journal Citation Reports* (*JCR*), Web of Science, Journal Report 1 (JR1), and institutional financial data. *JCR*s are part of ISI Web of Knowledge offering an organized document that contains meaningful metrics (see Table 1) for more than 10,000 journals from both science (8,000 journals) and social science (2,900 journals).

TABLE 1. Listing of the various metrics provided in a *Journal Citation Report* and what they measure to provide value (provided by Thomson Reuters, accessed at http://wokinfo.com/media/pdf/qrc/jcrqrc.pdf).

Metric name	Measurement
Total cites	Total number of times that each journal has been cited by all journals included in the database within the current JCR year.
Impact factor	The frequency an average article from a journal is cited in a particular year.
Five-year impact factor	The average number of times articles from the journal published in the last 5 years has been cited in the JCR year.
Immediacy index	The frequency the average article from a journal is cited within the same year it is published.
Article counts	The number of articles published in a journal in a particular year.
Cited half-life	Identifies the number of years from the current year that account for half of the cited references from articles published by a journal in the current year.
Eigenfactor score	Uses the current JCR year citations to citable items from the five previous years. Eigenfactor assigns a greater weight to citations coming from influential journals, allowing these journals to exert greater influence in determination of the rank of any journal they reference. Eigenfactor does not count self-citations, the sum of all Eigenfactor scores is 100; each journal's Eigenfactor score is a percentage of this total.

The JR1 is a standard format for reporting journal downloads created by Counting Online Usage of Networked Electronic Resources (COUNTER), which is a nonprofit consortium with an international steering committee of specialists from the library and publishing fields. Since 2003, this organization has been making it possible to record and report usage stats to allow for dissemination of this information in a consistent and systematic way (Shepherd, 2012). COUNTER is responsible for provisioning the standards of many important reports for various electronic resource types (i.e., databases, journals), specifically, successful full-text downloads for a journal. Thomson Reuters provides information on journal citations; filtering source articles by a particular institution creates a report for a given year on the articles written by that institution. The local institution provides cost data. The purpose of these data is to provide the local cost metric, and other relevant metadata (ISSN, purchase order ID, database, etc.) for each journal at the institution.

Data Preparation

Analysis for this research required coalescing data from *JCR*, JR1, Thomson Reuters citation data, and local (institution-specific) data and various metrics that are

DOI: 10.1002/asi

TABLE 2. List of the metrics, a categorization of global or local for this particular institution, and the data source that supplies these metrics, used as independent variables to model the correlation to downloads.

Metric name	Metric type	Data source	Definition
Full-text downloads	Local	Journal Report 1 (JR1)	Spreadsheet listing the successful full-text downloads for each journal by provider for the institution.
Cost	Local	Finance reports	Spreadsheet listing prices for 3,400 + electronic resources purchased (for this research in 2012).
Local citations	Local	Web of Science	Spreadsheet listing published articles for the institution and corresponding cited articles.
Total cites	Global	*Journal Citation Reports* (*JCR*)	Spreadsheet listing the journal name and the corresponding global metrics (i.e., impact factor, eigenfactor, article influence score, etc.).
Impact factor (IF)	Global	*Journal Citation Reports* (*JCR*)	
Five-year impact factor	Global	*Journal Citation Reports* (*JCR*)	
Immediacy index	Global	*Journal Citation Reports* (*JCR*)	
Article counts	Global	*Journal Citation Reports* (*JCR*)	
Cited half-life	Global	*Journal Citation Reports* (*JCR*)	
Eigenfactor score	Global	*Journal Citation Reports* (*JCR*)	
Article influence score	Global	*Journal Citation Reports* (*JCR*)	

TABLE 3. Listing of the linear regression model measures separately for each metric and local institution downloads.

Metric	Number of journals	S	R^2	R^2 (adjusted)	R^2 (predictability)
Article influence score	275	1,332	3%	3%	-.3%
Cited half-life	275	1,351	.3%	-.1%	−1%
Eigenfactor	**275**	**895**	**56%**	**56%**	**54%**
Immediacy index	275	890	3%	3%	−1%
Impact factor	275	804	4%	4%	2%
Five-year impact factor	275	805	4%	4%	2%
Total articles	275	1,110	33%	32%	30%
Global citations	**275**	**879**	**58%**	**58%**	**55%**
Local citations	**275**	**827**	**63%**	**62%**	**60%**

Note. Bold figures represent the three independent variables with the strongest correlation to full-text download.

provided or derived from these data sources. Data were migrated from the multiple sources (see Table 2) into a database using a Ruby Rake task, a script written in the Ruby programming language. Each spreadsheet was parsed separately and linked into the database by ISSN or (if ISSN was not provided) by journal title.

The collection and coalescence of these data from disparate sources eventually provided us with a database of information to begin our analysis. The database now has local categorization for the journals based on funding sources in 2012, download data to represent successful full-text downloads for these journals in 2012, citation data depicting the number of times authors from this institution cited these journals in their published articles in 2012, and finally their corresponding global metrics in 2012, such as total cites, impact factor, 5-year impact factor, immediacy index, articles, cited half-life, Eigenfactor, and article influence. The database contained 1,510 journal line items that included the global metrics and the local metrics for evaluation (such as subscription cost, full-text downloads, and local citations). We exported these data to an Excel Spreadsheet, imported the data into Minitab, and used the random selection tool in Minitab to randomly select 275 rows from our data and begin our regression analysis. Minitab 16.2 was used for all statistical calculations.

Results

To get an understanding of the relationship between the individual metrics (both global and local indicated in Table 2) and total downloads at an institution, we ran separate single linear regression models that compare total downloads and the respective metrics listed in Table 3 (i.e., article influence score, cited half-life, eigenfactor, etc.) on the 275 journal training data set. Table 3 indicates S, the average distance that the observed values of total downloads fall from the regression line and R^2, which informs us how well the model explains the variance of the response (i.e., downloads). Table 3 also presents R^2 (adjusted), a modified version of R^2 to account for how much the

DOI: 10.1002/asi

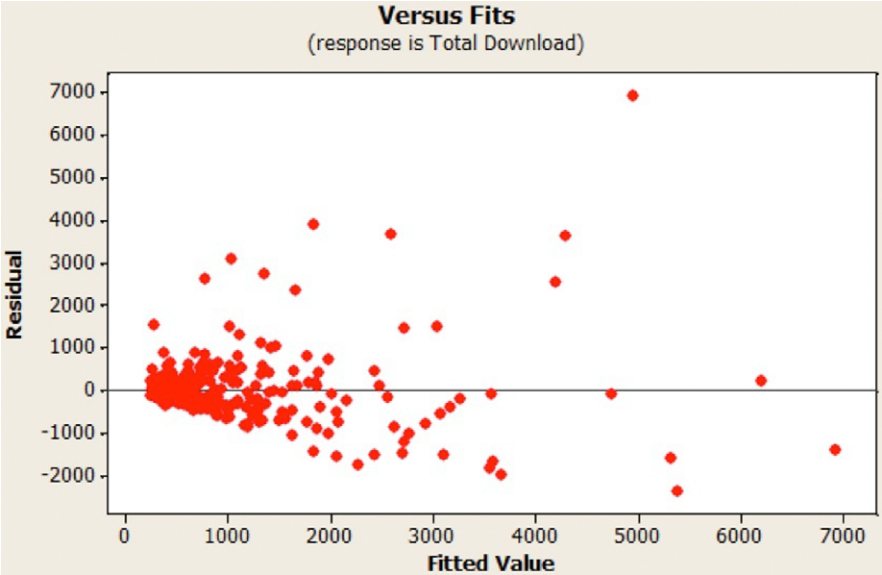

FIG. 1. Residual plot for global citations, displaying random error and centered on zero, which is consistent for an assumption of a regression model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

predictor variable improves the model based on the number of predictors in the model. Finally, Table 3 shows R^2 (predictability), which is how well a regression model predicts responses for new observations rather than just the original data set (Frost, 2013).

Table 3 illustrates the various strengths of each model (i.e., R^2, R^2 [adjusted] and R^2 [predictability]). Each metric listed in column 1 represents a separate model, and the strength of that model (indicated by the three R^2 values) indicates the varying strength of the correlation between that particular single metric and downloads. For example, local citations, global citations, and Eigenfactor all create regression models with an R^2 over 56% and even have predictability more than 54%, with local citations having the highest predictability at more than 60%. However, some metrics create weak models. In this instance, cited half-life and immediacy index have a R^2 of 0.3% and 3.0%, respectively, and a negative ability to predict downloads based on these metrics alone.

The assumptions for the residual plots in both the global citations (see Figure 1) and the local citations (see Figure 2) display random and unpredictable error that is consistent with an assumption of a regression model (i.e., one should not be able to predict the error for any given observation). The residuals are centered on zero, and the model is correct, on average, for fitted values.

Now, if we were to create a model based on combined multiple metrics using multiple linear regression, which metrics would be meaningful to add to this model? Using the training data set of 275 journals, we can run a linear regression model with all the metrics at one time to create an overall regression model. The normal probability plot (see Figure 3) of the residuals is approximately linear, and this supports the condition that the error terms are normally distributed.

The summary of the regression model using all metrics available (see Table 4) indicates a model that is stronger than any single regression model in explaining variance (R^2), how much the additional predictor variables improve the model (R^2 adjusted), and how well this model predicts downloads based on new observations (R^2 predictability). The regression equation is a mathematical formula that can be used to predict the outcome—in this case downloads—based on the predictor variables.

However, while this model is stronger than the individual models, it still does not tell us which of these metrics are statistically significant meaningful additions to our regression model. Looking at the coefficient table (see Table 4), there is a listing of the metrics, associated p value, and the coefficient (the multiplier) for that metric.

We consider a p value of .05 or less to be statistically significant. Using this p value, the following metrics are significant: Eigenfactor, five-year impact factor, total articles, and local citations, and impact factor.

A regression model using only the statistically meaningful metrics is listed in Table 5 (local citations, total articles, impact factor, five-year impact factor, eigenfactor). The summary of this model shows that the model lost little to no strength, based on R^2 values, by eliminating four metrics (article influence score, cited half-life, immediacy index, global citations). Based on an analysis of these regression models, the simplest equation that also had the highest R^2 value for predictability is presented in Table 5. Four of these metrics are global and provided by *JCR* (total articles,

DOI: 10.1002/asi

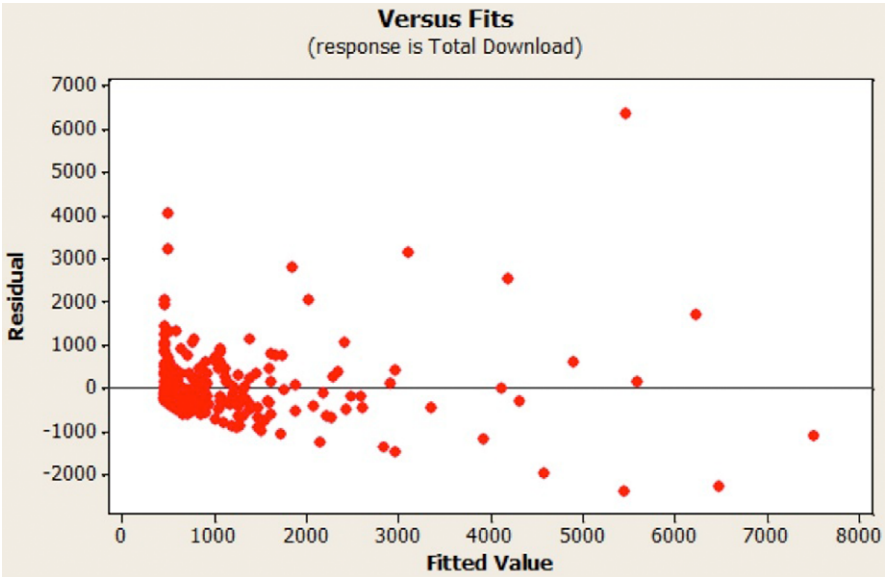

FIG. 2. Residual plot for local citations, displaying random error and centered on zero, which is consistent for an assumption of a regression model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

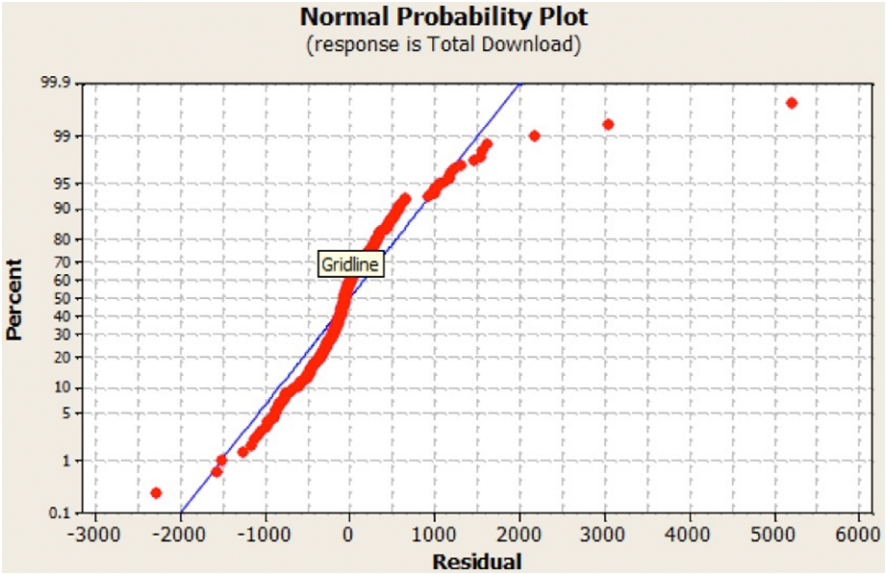

FIG. 3. Probability plot with regression line for all metrics at one time, which supports the condition that the error terms are normally distributed. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

impact factor, five-year impact factor, and eigenfactor), and one metric is local (local citations) and has an R^2 predictability value of 74%.

Based on the regression model presented in Table 5 we then used the developed equation on a test set to predict a range of downloads for 1,235 journals. We considered the model accurate if the predicted number of downloads, from the regression equations, fell within plus or minus of the S value (660), the average distance the observed value falls from the regression line. The results of this model's accuracy for prediction are displayed in Table 6, as well as the number of downloads and cost those journals accounted for in 2012 for this institution.

The R^2 predictability value for this model was 74%. When applied to a real-world data set, this model was able to successfully predict downloads within range 81% of the time; so our model performed better than expected. In our initial research objective, we stated that a successful model would predict a range of downloads within 70% of the journals tested, and these results exceed those expectations

DOI: 10.1002/asi

TABLE 4. Summary of the regression model created using all the metrics available.

Field	Value
S	658
R^2	77%
R^2 (adjusted)	76%
R^2 (predictability)	73%
Regression equation	89.205 + 15.8377 Citations—0.00495926 Total Cites—160.003 Impact Factor + 254.964 Five-Year Impact Factor -139.04 Immediacy Index + 0.846012 Total Articles—14.9955 Cited Half Life + 19,474.8 Eigenfactor—223.299 Article Influence Score

TABLE 5. Listing of the metrics used and their statistical significance for meaningful addition to a linear regression model for the sample data 275 journals.

Metric	*P*-value	Coefficient
Article influence score	.062	−223
Cited half-life	.288	−15
Eigenfactor	**.000**	**19,475**
Immediacy index	.233	−139
Impact factor	**.035**	**−160**
Five-year impact factor	**.006**	**255**
Total articles	**.000**	**.8**
Global citations	.606	0
Local citations	**.000**	**16**

Note. Bold figures represent metrics with a statistically significant *p* value.

TABLE 6. Summary of the regression model created using only the statistically significant metrics.

Field	Value
S	660
R^2	77%
R^2 (adjusted)	76%
R^2 (predictability)	74%
Regression equation	16.6861 + 16.1885 Citations—121.229 Impact Factor +111.129 Five Year Impact Factor + 1.00529 Total Articles + 14,925.5 Eigenfactor

of a successful model. There were eight journals that could be considered outliers for the number of downloads that this model was unable to accurately predict. Those eight journals represented 322,183 (44%) of all full-text downloads for the 235 journals that were not predicted within range. So if we removed these outliers, the predictability of our model would be even higher. This handful of journals, which are the outliers, may require future analysis.

A usage prediction model can be used to determine journal value or create a reasonable price range for libraries. Based on the metrics, and regression model developed in this research, we can create a predictable download range, where the range is the plus/minus average observed error in predicting downloads. Although we acknowledge other views of usage and data-based methods for calculation of usage exists (Carroll & Cummings, 2010), by using this downloads prediction range we can create an acceptable price range based on the average cost-per-download in 2012 (the 1,235 journals analyzed had a $2.47 average cost-per-download). Using this range of payment, we then find the outliers that are priced above this price range. Those that have an actual price above the predicted price range can be singled out for price adjustment. Then we can compare the actual price and the minimum and maximum price range to create a range of potential savings (Table 7).

The regression model developed predicts *The Journal of Applied Polymer Science* should have a range of 3,430–4,750 downloads. Based on this range of downloads and the average cost per journal, the library can calculate the range they would be willing to pay in order for this journal to maintain an average cost-per-use of $2.47. Based on these numbers, the library may be willing to pay in the range of $8,471–$11,731. We create the min price offer figure by multiplying the low number (3,430) in the predicted download range and average cost-per-download. We create the max price offer figure by multiplying the high number (4,750) in the predicted download range and average cost-per-download. This calculation creates dollar figures that allow the library to approach providers with when negotiating purchase deals and contracts with an understanding of the value of these journals at their home institution. In this example alone, the institution could have saved between $8,629 (min potential savings)–$11,889 (max potential savings) on the purchase price of this single journal or explored alternative methods of acquiring this journal (Table 8).

Using this model to calculate a range in downloads and subsequent range in spending, we can derive the number of journals that fall outside of this range on the expensive side and the potential savings that could be gained if this range was used as a basis for payment. Of the 1,235 journals analyzed, 264 (21%) were above the high price ceiling in the range created by predicted downloads and average cost-per-download. By focusing on cost savings or alternative methods of acquiring these journals, the institution could see savings in the range of $496,183–$1,356,929 per year. The range in savings is based on using either the lowest dollar point in the range the library should pay compared with the actual cost the institution paid or the highest dollar point in the range the institution should be willing to pay with the actual cost paid.

Discussion and Implications

*JCR*s have several metrics designed to provide insight into how valuable a journal will be at an institution. We consider these global metrics because they are providing information on the journal's totality (i.e., how often a journal was cited by any journal from an author at any institution) rather than providing data on a journal's impact

DOI: 10.1002/asi

TABLE 7. The number of journals tested for accurate range of prediction for downloads and those within range of prediction and those outside of the range, corresponding downloads, and cost.

	Number of journals	Overall number of downloads	Cost of journals
Journals within range of predicted downloads	1,000 (81%)	538,905 (43%)	$2,253,780 (72%)
Journals out of range of predicted downloads	235 (19%)	727,553 (57%)	$870,765 (28%)
Total	1,235	1,266,458	$3,124,545

at a specific institution. Metrics dealing with a particular institution are local metrics. The comparison of global metrics alone with the combination of global and local metrics is significant because it illustrates the distinction between global and local metrics, and the importance for libraries to calculate local metrics. The strongest single metric to model local downloads of a journal is local citation, how often a journal is cited by authors within that institution. The ability to predict local downloads is a key metric in determining the value of a journal; COUNTER continues to push the benefits of usage data, with new metrics such as usage factor, for evaluation of electronic resources.

We also show the predictive capability with both a single-regression model and multiple-linear regression model that used local citation rate as a meaningful variable in predicting full-text downloads. There is a strong correlation between these local numbers (citations and downloads), but what was surprising is the ineffectiveness of many other global metrics at predicting downloads at an institution. For example, many global metrics are calculated with the effort to normalize citations for a particular journal based on self-citations or normalization based on the number of articles that are published in a year. Yet the raw number of global citations still produces a stronger model to correlate the number of downloads at a local institution than either of these metrics (e.g., impact factor, five-year impact factor) that have been normalized for equity.

However, these metrics do not need to live in a vacuum, and there may be advantages that can be gained by combining these metrics to create stronger models. Creating a simple model (i.e., few metrics, or independent variables) is still a goal, but creating a strong model is the ultimate goal, and thus we only add metrics to our regression model that are considered significant. For example, global citations alone created a relatively strong model to correlate journal downloads. However, when creating a regression model with multiple independent variables, global citations became less meaningful and local citations remained essential. This indicates the importance for institutions to create local metrics to rely on for predicting journal value. Eigenfactor also remained a meaningful metric in this predictive model for local full-text downloads, illustrating that the best way to analyze these resources is to combine both global and local metrics. Although there is the possibility that a local metric may change over time, we would expect any changes to be gradual, similar to the changes in global metrics. Conversely, some of these metrics are not helpful in predicting downloads either alone or in a multiple regression model, and they have little value relative to correlating downloads at a local institution.

The ability to create a strong model that can accurately predict downloads within a range provides the capability to create an acceptable price-per-download (PPD) that an institution may be willing to pay. Based on this model, PPD, and the ability to predict downloads, a library can enter a contract negotiation with an estimated range of how much they would be willing to pay for a journal or a dollar figure on the value of that journal to the institution. From a collection management view, this provides administrators with powerful metrics to leverage when considering purchasing journals. For example, if a library pays $10,000 for a journal, and articles from that journal are downloaded 7,000 times a year, then it makes sense to pay for that journal on an annual basis rather than on a per-article basis.

The model developed to create the ability to predict value has inherent weaknesses. *JCR*s only exist for a little over 10,000 scientific journals, and this is only a part of the journals that the libraries provide access to. This creates an issue because not all journals can be evaluated with the model, unless the data can be obtained from other sources. However, this is yet another layer of abstraction required to create a prediction model that relies on third-party resources. Additionally, there is a large overhead involved in aggregating these data from various resources, writing scripts to parse the data, and data modeling to successfully organize the data for evaluation.

However, the benefit relative to the overhead involved in programming, data modeling, and aggregating the data is that this process can easily be repeated each year after the initial investment; some tasks done individually, prior to automation, typically take weeks. The volume of journals in a library collection to manage can be more than 100,000, so this model creates an ability to quickly take the number of journals to pay particular attention to a more manageable number by finding outliers based on a PPD and thus providing the ability to make a financial impact. Additionally, this technique can be applied to packages or new additions to packages in addition to individual journals. This model also creates local metrics that may not otherwise exist and illustrates the correlation these local metrics have with journal value at an institution. Finally, this model aggregates all the data sources and simplifies the ability to provide a number of people to access one place to analyze data and reference for

DOI: 10.1002/asi

TABLE 8. A subset of journals that have a higher actual price than predicted price range based on predicted downloads. Their predicted download range for each journal and the range of price to offer and potential savings based on the difference between price offer and actual price.

Name	Actual price	Predicted download range	Min price offer	Max price offer	Min potential savings	Max potential savings
Journal of Applied Polymer Science	$20,361	3,430–4,750	$8,471	$11,731	$8,629	$11,889
Chemical Physics Letters	$16,501	1,245–2,565	$3,073	$6,334	$10,167	$13,427
Thin Solid Films	$15,600	3,351–4,671	$8,276	$11,536	$4,063	$7,323
Journal of Cellular Physiology	$15,399	1,340–2,660	$3,310	$6,571	$8,827	$12,088
Journal of Cellular Biochemistry	$14,828	1,168–2,488	$2,884	$6,145	$8,683	$11,944
Journal of Crystal Growth	$14,446	513–1,833	$1,267	$4,527	$9,918	$13,179
Surface Science	$11,214	2,535–3,855	$6,261	$9,521	$1,692	$4,953
Molecular Microbiology	$11,181	860–2,180	$2,123	$5,384	$5,797	$9,057
Neuroscience	$10,800	3,033–4,353	$7,491	$10,751	$48	$3,309
Journal of Neuroscience Research	$9,769	574–1,894	$1,418	$4,678	$5,090	$8,350
Surface & Coatings Technology	$9,226	1,274–2,594	$3,147	$6,407	$2,819	$6,079
Developmental Dynamics	$8,457	599–1,919	$1,478	$4,739	$3,718	$6,978
Statistics in Medicine	$8,131	1,676–2,996	$4,140	$7,401	$730	$3,990
Electrophoresis	$8,108	892–2,212	$2,203	$5,464	$2,644	$5,904
Cell and Tissue Research	$7,786	62–1,382	$152	$3,413	$4,373	$7,633
The Science of the Total Environment	$7,577	1,580–2,900	$3,902	$7,163	$413	$3,674
Nuclear Engineering and Design	$7,441	245–1,565	$605	$3,866	$3,575	$6,835
International Journal of Heat and Mass Transfer	$7,329	1,379–2,699	$3,406	$6,667	$662	$3,922
Marine Biology	$7,003	808–2,128	$1,996	$5,257	$1,746	$5,006

further discussion regarding collection management. A contribution of this application would be to not qualitatively make final decisions with regard to collection management, but rather highlight the electronic resources that are worthy of discussion and further evaluation.

Focusing on methods that allow for more time to spend on data analysis than aggregation will provide opportunities for future work. One area for future work is to improve on predictability aspects by identifying journals that are likely to be outliers in usage or cost. Another is how the multiplicity of journal access sources impacts the findings from this model. Duplication of content is, of course, a likely area for libraries to save additional dollars on electronic resources. Additionally, we believe that the approach used in this research could be combined with other data sources, such as survey data (Akbulut, 2015), to further enlighten the impact of journals at the local level.

Conclusion

Collection management for electronic resources is an extremely complex job. There are tens of thousands of journals, a multitude of providers, and various platforms to manage; these factors all contribute to the complexity. The information needed to analyze these resources resides in a number of systems (*JCR*, Web of Science, JR1, etc.), and many of these systems are remote and have various formats to access them (typically spreadsheets, PDFs, and more spreadsheets). The outside data from third-party vendors have value, but the real value is when these global data can be combined with local numbers to create a true understanding of what resources are providing value to the institution and to better serve the mission of the library and—more specifically in this case—collection management departments.

Some of the bibliometrics provided by *JCR* attempt to normalize citation numbers to reduce the impact of a journal's size. However, journal price appears tied to journal size and, because of this dependency, some of the metrics provided by *JCR* are not significant in predicting journal price, as prior work has noted a lack of correlation between the price of a journal and impact (Bensman, 1996, 2012). COUNTER's Usage Factor makes a similar attempt to normalize download data, by using the median usage over 2 years instead of the mean. This could impact the usage of full-text downloads as a measure for journal quality and therefore price. However, our research indicates normalized global metrics do not have as much worth in predicting value at an institution as metrics created locally. In fact, the most powerful metric correlating to download is a local metric—the number of citations a journal received at the institution in that year.

One aspect of this research that perhaps was not emphasized is the amount of work needed to bring all of these data together. There are plenty of existing data format standards that would allow the coalescence of these disparate systems in a more real-time and seamless manner. Mechanisms should be created to allow this information to flow more

DOI: 10.1002/asi

freely. Evaluation of electronic resources typically provides many numbers to consider. Each of these numbers tells part of the story but leave much to guess work. However, when these numbers are combined they can create a model that can financially benefit libraries. Providers of these data should be willing to help make these data less complex and create an environment where fiscal responsibility is something providers facilitate rather than hinder.

References

Akbulut, Y. (2015). Predictors of inconsistent responding in web surveys. Internet Research, 25(1), 131–147.

Arnold, D.N., & Fowler, K.K. (2011). Nefarious numbers. Notices of the American Mathematical Society. American Mathematical Society, 58(3), 434–437.

Bartsch, R.A., & Tydlacka, B.L. (2003). Student perceptions (and the reality) of percentage of journal articles found through full-text databases. Research Strategies, 19(2), 128–134.

Bensman, S.J. (1996). The strucfure of the library market for scientific journals: The case of chemistry. Library Resources & Technical Services, 40(2), 145–170.

Bensman, S.J. (2012). The Impact Factor: Its place in Garfield's thought, in science evaluation, and in library collection management. Scientometrics, 92(2), 263–275.

Carroll, D. (2009). Procedures for creating a serials decision database. Retrieved from https://research.wsulibs.wsu.edu:8443/xmlui/handle/2376/2277

Carroll, D., & Cummings, J. (2010). Data driven collection assessment using a serial decision database. Serials Review, 36(4), 227–239.

Coughlin, D.M., Campbell, M.C., & Jansen, B.J. (2013). Measuring the value of library content collections. Proceedings of the American Society for Information Science and Technology, 50(1), 1–13.

Frost, J. (2013). Applied regression analysis: How to present and use the results to avoid costly mistakes, part 1—Adventures in Statistics | Minitab. The Minitab Blog. Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/applied-regression-analysis-how-to-present-and-use-the-results-to-avoid-costly-mistakes-part-1

Furlough, M.J. (2012). Opening access to research: From concepts to actions [Powerpoint slides]. Retrieved from http://scholarsphere.psu.edu/files/37720c723

Gallagher, J., Bauer, K., & Dollar, D.M. (2005). Evidence-based librarianship: Utilizing data from all available sources to make judicious print cancellation decisions. Library Collections, Acquisitions & Technical Services, 29(2), 169–179.

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. JAMA: The Journal of the American Medical Association, 295(1), 90–93.

Jansen, B.J. (2009). Understanding user-web interactions via web analytics. Synthesis Lectures on Information Concepts, Retrieval, and Services, 1(1), 1–102.

Jansen, B.J., & Rieh, S.Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. Journal of the American Society for Information Science and Technology, 61(8), 1517–1534.

Kyrillidou, M., Morris, S. & Roebuck, G. (2007). ARL Statistics 2011-2012, Association of Research Libraries, Washington, DC, available at: http://publications.arl.org/ARL-Statistics-2011-2012/45 (accessed 22 March 2015).

Lakos, A. (2007). Evidence-based library management: The leadership challenge. Libraries and the Academy, 7(4), 431–450.

Lane, J. (2010). Let's make science metrics more scientific. Nature, 464(7288), 488–489.

Leon, L., & Kress, N. (2012). Looking at resource sharing costs. Interlending & Document Supply, 40(2), 81–87.

Medeiros, N. (2007). Usage statistics of e-serials.

Metz, P. (1992). Thirteen steps to avoiding bad luck in a serials cancellation project. The Journal of Academic Librarianship, 18(2), 76–82.

Metz, P., & Cosgriff, J. (2000). Building a comprehensive serials decision database at Virginia Tech. Library Collections, Acquisitions & Technical Services, 61(4), 324–334.

Miller-Francisco, E. (2003). Managing electronic resources in a time of shrinking budgets. Library Collections, Acquisitions & Technical Services, 27(4), 507–512.

Minasny, B., Hartemink, A.E., & McBratney, A. (2007). Soil science and the h-index. Scientometrics, 73(3), 257–264.

Oosthuizen, J.C., & Fenton, J.E. (2014). Alternatives to the impact factor. The Surgeon, 12(5), 239–243.

Ortiz-Cordova, A., & Jansen, B.J. (2012). Classifying web search queries to dentify high revenue generating customers. Journal of the American Society for Information Science and Technology, 63(7), 1426–1441.

Pesch, O. (2012). Usage Factor for journals: A new measure for scholarly impact. The Serials Librarian, 63(3–4), 261–268.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. Research Policy, 41(7), 1262–1282.

Shepherd, P.T. (2012). The COUNTER Code of Practice for e-Resources: Counter Online Metrics (pp. 1–29). Chicago, IL: The Inaugural Coase Lecture.

Sykes, A.O. (1993). An introduction to regression analysis. In An introduction to regression analysis Coase lecture (20th ed., pp. 1–33). Chicago, IL: University of Chicago Law School.

DOI: 10.1002/asi