Journal of MultiDisciplinary Evaluation
Volume 12, Issue 27, 2016

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Predicting Research Productivity in International Evaluation Journals Across Countries

Christoph E. Mueller
German Research Institute for Public Administration

Hansjoerg Gaus
Saarland University, Center for Evaluation

Ingo Konradt
GESIS Leibniz Institute for the Social Sciences

Background: Progress in evaluation research depends on the continuous generation of scholarly knowledge and its dissemination in the community. One way of disseminating findings is to publish in scientific journals and researchers, institutions, and even whole countries are assessed by their output in these journals. Particularly with regard to countries, there is an uneven distribution of research productivity in evaluation journals. A viable model for predicting countries' research output in international evaluation journals, however, has not yet been developed.

Purpose: The purpose of this study was to develop and test a model for the prediction of countries' research output in international evaluation journals by predictors from the research, economic, and social/political system.

Setting: NA

Intervention: NA

Research Design: A cross-sectional design was used for predicting research output in evaluation journals across countries.

Data Collection and Analysis: Our sample consists of 65 countries that made contributions to ten international peer-reviewed evaluation journals. We collected data for the period from 2009 to 2013 and predicted the number of authorships across countries by using boosted regression trees, a machine learning procedure.

Findings: Our model provided accurate predictions of countries' research output. Research productivity in the social sciences had the strongest effect, followed by economic prosperity, control of corruption, and age of evaluation society. The model was generalizable to another period of time with only marginal loss in predictive accuracy.

Keywords: *research productivity; journal publications; evaluation research; evaluation journals; boosted regression trees.*

Introduction

Progress in the field of evaluation research depends on the continuous generation of scholarly knowledge and its dissemination in the community of evaluation researchers. Communicating research findings in the field reflects the *permanence* of activities involved in evaluation systems (Leeuw & Furubo, 2008) and contributes to a discourse about evaluation, which is considered to be a dimension of *evaluation culture* (Jacob, Speer, & Furubo, 2015).

One way of disseminating research findings is to publish in scientific journals. This has become a very important strategy of communicating scholarly knowledge in modern science (e.g., Canagarajah, 2002; Origgi & Ramello, 2015) and is a critical function of research communities (Vinluan, 2012). This is also true for evaluation research, where the "international dialogue is typically associated with publications, particularly journals" (Diaz-Puente, Cazorla, & Dorrego, 2007, p. 400). In order for articles to contribute to an international dialogue, they should be published in international journals. Basically, journals can be considered international if they are published in English since this is the predominant language in modern science (Flowerdew, 1999; Short et al., 2001). Moreover, international journals "should bring together authors who cross national, continental, and language boundaries to develop contributions to these publications" (Diaz-Puente et al., 2007, p. 400) and should have policies "of publishing high-level, international-refereed scientific articles from authors from all the countries of the world" (Gutiérrez & López-Nieva, 2001, p. 54).

The importance of journal publications in the field of evaluation research is highlighted by the fact that various aspects of researchers' publishing behavior in evaluation-focused journals have already been investigated (e.g., Christie & Fleischer, 2010; Coryn et al., 2016; Diaz-Puente et al., 2007; Heberger Christie, & Alkin, 2010; Nielsen & Winther, 2014). However, there are still research questions that have not been answered so far, one of these being why there is an uneven distribution of research productivity in international evaluation journals across countries. As Diaz-Puente et al. (2007) have demonstrated, North America–particularly because of the United States (US)–has by far the highest research output with more than two thirds of the contributions in international evaluation journals from 2000 to 2005. By contrast, European countries were responsible for less than a quarter of the contributions. Despite some hints pointing towards the factors which may be responsible for this uneven distribution (e.g., Diaz-Puente et al., 2007), a viable model for predicting countries' research output in international evaluation journals has not yet been developed. Thus, in this study we develop and test a model to find evidence that helps understand why the output in evaluation journals varies between countries. Before we present the data, methods, and results of our analyses, we describe the theoretical framework of our study.

Theoretical Framework

In conceptualizing research productivity, we follow a widely applied definition (e.g., Basu, 2010; Ramsden, 1994; Vinluan, 2012) and use the term synonymously with the output of publications in scientific journals within a given period of time. In this study, we focus on investigating potential predictors of research productivity in international evaluation journals at country level. More specifically, we are interested in finding out to what degree macro-level variables that represent aspects of countries' research, economic, and social/political systems are suited to predicting their research output in international evaluation journals.

As regards the research system, we assume that countries' output in evaluation journals is associated with the level of their *research productivity in the social sciences*. Since evaluation research is considered to be a "social science activity" (Rossi, Lipsey, & Freeman, 2004), we assume that it is strongly represented in countries where the social sciences in general have a high standing. Thus, we expect countries with strong social sciences–that is, countries which are very productive in this area–to be stronger in evaluation research, and to have a higher output in evaluation journals.

Furthermore, we believe that characteristics of a country's academic sector play an important role. Although the academic sector may consist of more than just universities–for example, publicly funded research institutes–it is usually the universities that are dominant in the production of peer-reviewed literature in evaluation journals (Nielsen & Winther, 2014). Hence, we assume that countries' research productivity is positively correlated with the *research performance of their universities*. The better a country's universities generally perform in research activities, the better they perform in publishing in international evaluation journals. Moreover, we believe that the

size of the academic sector has an effect on a country's research productivity too. We expect the number of publications in international evaluation journals to increase when the number of universities rises (e.g., Meo et al., 2013), simply because there are more researchers who could publish articles.

A discipline-related predictor for countries' output in evaluation journals is their *evaluation culture/tradition*. Nielsen and Winther (2014, p. 327) found that the dominance of a country with regard to its output in evaluation journals may be explained by the fact that its evaluation tradition evolved earlier and is more mature. Unfortunately, besides the index of evaluation culture developed by Furubo and Sandahl (2002) and updated by Jacob et al. (2015)–available for only 19 countries–we did not find any data on this predictor. We did however obtain sufficient data for the *age of evaluation societies/associations*, which we treat as a surrogate for evaluation culture. The existence of evaluation societies is an expression of evaluation culture (Furubo & Sandahl, 2002) and they "bring together evaluators from multiple disciplines to share knowledge and experiences, bridge disciplinary divides, debate issues of fundamental importance, set standards and ethical guidelines, build skills, and chart the future as a group with a strong and shared identity" (Love & Russon, 2000, p. 450). These processes can encourage cooperation and stimulate innovations in the field, which may eventually lead to publications in evaluation journals. We expect these processes to become more intense over time and thus assume an increase in countries' output in evaluation journals with increasing age of their evaluation societies.

The last research-related indicator is the *size of the continental evaluation journal market*. Following the idea of Diaz-Puente et al. (2007), we assume that evaluation researchers predominantly publish in journals located on their home continent. Thus, we expect countries' research output in evaluation journals to be positively correlated with the size of the evaluation journal market on their home continent. In other words: the more international evaluation journals are edited on the home continent of a country, the higher its output in international evaluation journals.

Furthermore, we assume a positive association between countries' output in evaluation journals and their *economic prosperity*. There is evidence that economic prosperity is positively correlated with countries' expenditure on R&D (e.g., Lane, 2011), which in turn increases the performance of the research system and research productivity in terms of scientific publications (e.g., Meo et al., 2013). Moreover, we presume that economically strong countries are capable of spending more money on publicly funded social interventions (e.g., Tanzi & Schuknecht, 2000), which may contribute to a higher intensity of evaluative activities. As a consequence, we expect an increase in the number of evaluation researchers, which we believe increases the likelihood of publishing in evaluation journals.

We also expect two constructs from the social/political sphere to be important predictors. One of these is *corruption*. Corruption is known to be negatively correlated with economic prosperity (e.g., Husted, 1999), which in turn is positively related to the general performance of research systems. Moreover, because evaluation is an instrument capable of uncovering corruptive practices, we expect evaluative activities (including evaluation research) to be weaker in countries where corruption is a problem. We thus assume a negative correlation between the degree of corruption and countries' output in evaluation journals.

The second social/political predictor is called *civil liberties*. We consider the existence of civil liberties as an important antecedent for the professional independence of evaluators, which is required for conducting evaluation research (e.g., Markiewicz, 2008). The independence of evaluators and their research depends on civil liberties such as the rule of law, organizational rights, freedom of expression and opinion, personal autonomy, and contractual security. We thus assume that the intensity of evaluation research is positively correlated with the scope of civil liberties in a country. Put differently, we expect countries with more civil liberties to have higher outputs in evaluation journals.

Finally, following Diaz-Puente et al. (2007), we expect that there are *linguistic boundaries* with regard to publishing in international evaluation journals. More precisely, a "lack of familiarity with the English language is probably hindering the number of international submissions and causing the poor quality of some of these submissions" (p. 412). Consequently, we expect countries in which evaluation researchers are more familiar with the English language to have higher outputs in international evaluation journals.

Method

Measures

Research Productivity in International Evaluation Journals. We collected data for the dependent variable on the basis of ten journals focusing on aspects of evaluation research, practice, concepts, and methods. In doing so, we focused exclusively on *designated evaluation journals* and left aside articles related to the field of evaluation that were published in domain-specific, specific disciplinary, or generic social science methodology journals (Nielsen & Winther, 2014, p. 313). Moreover, we excluded journals without *peer review* and only considered journals where all the articles are published in *English*. Thus, journals such as the partially French-language *Canadian Journal of Program Evaluation* or the German-language *Zeitschrift für Evaluation* were excluded from data collection. Finally, we only considered journals which had the term *evaluation* in their title.

In the next step, we had to decide on the period of time for which data were to be collected. Because capturing all the articles published in the last few decades would have been very time-consuming, we decided to consider only articles published from 2009 to 2013. In this period, we scanned 1,260 articles and collected information on the institutions to which their authors were affiliated. We then determined the nationality of these institutions. We employed the nationality of the institutions instead of that of the authors because many authors work abroad, which is why the scientific value creation takes place in the home countries of the institutions. In the case of freelancers, we assigned the country in which these authors were working.

Because some authors made more than one contribution in the period of time considered and because articles were frequently published by several authors affiliated with institutions from different countries, we employed the number of *authorships* instead of the number of articles as the unit in our analysis. An authorship is an author's single contribution to one published article. Consequently, the number of authorships in our dataset is substantially larger than the number of articles. We did not distinguish between first and other authorships because previous research has shown that analyzing total authorships provides similar results to analyzing first authorships only (Diaz-Puente et al., 2007).

Finally, there are some particularities that have to be noted. Firstly, we only considered original research articles and excluded book reviews, comments, discussions, and other formats that were not original research articles. Secondly, we did not consider articles that were published online first, but only those that had already appeared in edited issues. Thirdly, some authors were affiliated to more than one institution belonging to different countries or to institutions that could not be assigned to a single country (e.g., World Bank, UN organizations). In these cases, authorships were subsumed under the category *international*. Fourthly, at the time of data collection, there was only one issue published by the *Evaluation Review* for 2013. Table 1 presents the number of articles and authorships in our sample from 2009 to 2013.

Table 1
Articles and Authorships in Selected Journals (2009 — 2013)

	AR(n)	*AR(%)*	*AS(n)*	*AS(%)*
American Journal of Evaluation	103	8.2	264	7.5
Educational Evaluation and Policy Analysis	117	9.3	305	8.7
Educational Research and Evaluation	164	13.0	401	11.4
Evaluation	115	9.1	248	7.1
Evaluation and Program Planning	288	22.9	898	25.5
Evaluation and The Health Professions	135	10.7	515	14.6
Evaluation Review	93	7.4	318	9.0
Journal of MultiDisciplinary Evaluation	51	4.0	124	3.5
Practical Assessment, Research & Evaluation	85	6.7	176	5.0
Studies in Educational Evaluation	109	8.7	268	7.6
Total	1,260	100.0	3,517	100.0

Note. AR(n) = absolute number of articles*; AR(%)* = proportion of articles; *AS(n)* = absolute number of authorships; and *AS*(%) = proportion of authorships.

Research Productivity in the Social Sciences. One way to think about a country's research productivity in the social sciences is to consider its output of citable documents in a certain period of time (Ramsden, 1994). The more citable documents a country produces in the social sciences within a defined period, the higher its research productivity in the period and area concerned. We used the number of citable documents produced by countries in the social sciences from 2009 to 2013. The data were retrieved from the SCImago Journal and Country Rank (http://www.scimagojr.com). The mean value of the variable in our prediction sample[1] was 58,680 (*SD* = 61,290).

Research Performance of the Academic Sector. We intended to use the average number of universities placed in the top 500 of the 'Academic Ranking of World Universities' per country (http://www.shanghairanking.com) from 2009 to 2013, which identifies the world's best universities by using indicators such as highly cited researchers, papers indexed in major citation indices, or the per capita academic performance of the institution. Yet in preliminary analyses we found that the indicator was strongly correlated with research productivity in the social sciences ($r > .95$). We thus excluded it from further analyses because it did not add any new information to our model.

Size of the Academic Sector. We measured the size of the academic sector by using the number of universities located in a country. We retrieved the data from the 'Webometrics Ranking of World Universities', 2015 edition (http://www.webometrics.info). The mean value was 329.93 (*SD* = 625.97).

Age of Evaluation Society/Association. We gathered data on the age of evaluation societies from the website of the 'International Organisation for Cooperation in Evaluation' (http://www.ioce.net). As there were some countries for which no information was available on that website, we searched the Internet and duly found data in some cases. When assigning numeric values to the variable, we had to deal with the fact that some societies were founded within the period from 2009 to 2013. In order to compute a variable that reflects this circumstance, we chose 2014 as the starting point for our calculations and subtracted the founding year of each evaluation society from the value 2014. The mean age was 9.33 years (*SD* = 3.52).

Size of Continental Evaluation Journal Market. We divided the ten evaluation journals into two journal markets, namely North America (7 journals) and Europe (3 journals).[2] Subsequently, we assigned the value seven to the two North American countries USA and Canada[3] because they represent the North American journal market. Secondly, all European countries (including Israel) received the value three because three of the journals were assigned to the European journal market. Finally, we assigned the value zero to all the remaining countries because no journal included in our sample is edited on their continents. In total, 34 countries received the value 0, 22 countries received the value 3, and 2 countries received the value 7.

Economic Prosperity. We operationalized economic prosperity by countries' average per capita GDP from 2009 to 2013. We used per capita GDP instead of absolute GDP[4] because it considers the size of countries in terms of their overall population. Thus, the per capita GDP shows the relative economic performance of countries. We obtained data for per capita GDP from the World Bank (http://data.worldbank.org). The mean value was 23,785 US$ (*SD* = 25,378).

Control of Corruption. In order to operationalize the degree of corruption in a country, we used the percentile ranking of the *Control of Corruption*

[1] Due to missing values, we did not use all cases for prediction. Descriptive statistics were only calculated for countries which were included in the predictive analyses.

[2] Nielsen and Winther (2014) subsumed the journal "Studies in Educational Evaluation" under the category "other" and Diaz-Puente et al. (2007) characterized it as being Asian because Israel is the home country of the institution that sponsors the journal. We assigned the journal to the European journal market because of the special relationship between the Israeli and European research systems.

[3] Due to the cultural, political, and linguistic boundaries between the US/Canada and Mexico, we did not assign Mexico to the North American journal market.

[4] Originally, we intended to include the absolute GDP in our model too. Yet this variable was highly correlated with other predictors, particularly the number of universities ($r = .90$). In order to prevent issues related to multicollinearity, we did not include absolute GDP in our model.

index developed by the World Bank[5]. This index reflects "perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as 'capture' of the state by elites and private interests" (Kaufmann, Kraay, & Mastruzzi, 2011, p. 223). The value of a country on the index indicates its percentile rank among all countries. It ranges from 0 (lowest rank) to 100 (highest rank), with lower values indicating less control of corruption. We employed the average rank from 2009 to 2013. The mean value was 60.47 (*SD* = 30.53).

Civil Liberties. To measure civil liberties, we used the average of the Freedom House civil liberties index from 2010 to 2014, reflecting events from 2009 to 2013. This index is part of the *Freedom in the World* survey (https://freedomhouse.org) and covers aspects such as freedom of expression and belief, associational and organizational rights, rule of law, and personal autonomy without interference from the state. The index can take values from 1 to 7, a rating of 1 standing for the highest degree of freedom in respect of civil liberties. The mean value was 2.58 (*SD* = 1.69). None of the countries in the sample received the value 7.

Linguistic Boundaries. We used the variable 'English as an Official Language' for capturing linguistic boundaries. We assume that researchers working in countries where English is an official language have a better knowledge of English than those working in countries where English is not an official language. We created a dummy-coded variable where countries which do have English as an official language received the value one and all other countries the value zero.

Data Analysis

Before estimating our model, we log-transformed the dependent variable in order to reduce the degree of skewness [6] and treated the natural logarithm of the number of authorships (hereinafter referred to as ln(AS)) as a continuous dependent variable. To predict ln(AS), we used a machine learning technique known as *boosted regression trees* (BRT; Friedman, 2001; Hastie, Tibshirani, & Friedman, 2009). This technique has several desirable characteristics; for example, it does not force us to make assumptions about the functional form of the relations between predictors and the dependent variable, it does not impose restrictions as regards the number of predictors, it automatically identifies and models interactions between the predictors, and it considerably enhanced the predictive power of our model.[7]

BRT is a non-parametric approach that predicts the values of a response variable by combining two different algorithms, namely regression trees and boosting. Regression trees are a recursive partitioning method that predicts a response variable by using a series of rules in order to identify regions that have the most homogeneous responses to predictors and fit a constant to each of those regions (Elith, Leathwick, & Hastie, 2008). The result of this process can then be visualized in the form of a decision tree. When fitting a regression tree, the order of the selected predictors and the split points are chosen in such a way that the prediction errors are minimized (Elith et al., 2008). In the case of a continuous dependent variable, prediction errors are usually measured by the squared difference between the observed and fitted values.

Frequently, single regression trees are not sufficiently accurate in predicting the response. In order to improve model accuracy, regression trees are thus combined with the boosting algorithm. Schonlau (2005) provides a concise explanation of how boosted regression trees work in the case of continuous dependent variables:

> The average *y*-value is used as a first guess for predicting all observations. This is analogous to fitting a linear regression model that consists of the intercept only. The residuals from the model are computed. A regression tree is fitted to the residuals. For each terminal node, the average *y*-value of the residuals that the node contains is computed. The regression tree is used to predict the residuals. (In the first step, this means that a regression tree is fitted to the difference between the observation and the average *y*-value. The tree then predicts those differences.) The boosting regression model–consisting of the sum of all previous regression trees–is updated to reflect the current regression tree. The residuals are updated to

[5] http://info.worldbank.org/governance/wgi/index.aspx#home

[6] We also estimated our model with the untransformed number of authorships and fitted a boosted Poisson regression tree model for predicting the response. Neither the overall fit of our model nor the influences of the single predictors differed substantially. Thus, we believe that using ln(AS) was an appropriate choice.

[7] BRT clearly outperformed OLS regression in terms of predictive accuracy. When we compared the predictive performance of BRT and OLS, the BRT model explained 64.8% of the variance in ln(AS) in a test dataset whereas the OLS model was only capable of explaining 40.6%.

> reflect the changes in the boosting regression model; a tree is fitted to the new residuals, and so forth. (p. 336)

The final model is a linear combination of all fitted regression trees and can be thought of as a regression model where each term is a tree (Elith et al., 2008). The predicted values in the final model are calculated by multiplying the sum of all trees by the *learning rate*, which is introduced below.

Besides specifying an appropriate loss criterion (which in our case is the squared error loss), researchers have to make several other decisions. One important choice concerns the number of splits (J) that are used for fitting each tree. Allowing J splits per tree is equivalent to a model with up to J-way interactions. According to Hastie et al. (2009), a number from four to eight splits generally works well. In our study, we allowed four splits per tree.

Another decision concerns the number of trees to be fitted. Because of the functional flexibility of the BRT method, fitting more and more trees can lead to formidable overfitting (which means that the model will fit the data on which it was trained well, but that it is not generalizable to other observations from the same population). Thus, the number of iterations has to be restricted so that the model is still generalizable. A way to find the optimal number of trees is to split the full dataset into a training and a test dataset and fit the training data with exactly the number of trees that maximizes the log likelihood on the test data (Schonlau, 2005). We randomly assigned 60% of our observations to the training and 40% to the test dataset. The automatic identification of the optimal tree number is implemented in the Stata plugin *boost* (Schonlau, 2005), which we used for all predictions. Since randomly splitting the data leads to different compositions of training and test data each time this process is repeated, we conducted the whole BRT analysis 100 times and averaged the predicted values over the 100 runs in order to obtain our final predictions.

Furthermore, there are two commonly used variations of the boosting algorithm which we employed. One of these, the learning rate, reduces the impact of each additional tree in order to avoid overfitting. Usually, small learning rates are chosen because it is more effective to improve a model by taking many small steps than by taking a few large ones (e.g., Schonlau, 2005). We used a common learning rate of 0.01. The other strategy often followed is called *bagging*, which improves the approximation accuracy of boosting. At each iteration, a subsample is randomly drawn without replacement from the training dataset. This subset is then used for fitting the regression tree and computing the model update for the current iteration (Friedman, 2002). We employed a bagging fraction of 80%.

The results provided by BRT analyses are different from those one obtains when using traditional regression models. Although there is an R^2 value computed for the test dataset, BRT does not provide regression coefficients. Instead, it works with the concept of the *relative influence* of predictors (Friedman, 2001). This measure is computed as the improvement in squared error as a result of using a variable to form splits, averaged across all regression trees (Friedman & Meulman, 2003). The relative influence of each predictor is scaled, which means that the sum of all the relative influences of the predictors equals 100. Higher values indicate stronger influences on the dependent variable.

Relative influences do not tell us what the functional form of the relation between a predictor and the dependent variable is. A common way of learning about this functional form is to visualize conditional effects of the predictors with partial dependence plots. These plots represent visualizations of the marginal effects of the single predictors while holding all other predictors constant. In addition to the conditional effects estimated on the basis of the full dataset, we also plotted conditional effects that were estimated without considering the US. We did this in order to assess the sensitivity of our model to the greatest outlier in our dataset.

Finally, the utility of a predictive model depends on its external validity, which means that it should not only provide accurate predictions in the dataset on which it was developed but also in different settings. Therefore, in order to assess whether our model is generalizable, we collected additional data on almost all the predictors[8] for a different period of time (2005 and 2006). We then used the model fitted on the data from 2009 to 2013 and fed it with the new observations. Based on this new information, we were able to predict ln(AS) for the period from 2005 to 2006 and compare the predicted with the observed values of ln(AS) within that period. The new dataset contained 39 countries (with 1,064 authorships),

[8] Due to a lack of data, we used the same information on the number of universities as we did in the model for predicting ln(AS) for the years from 2009 to 2013. Moreover, for the civil liberties index we were only able to use data for the year 2006 because there was no information available for 2005.

of which 34 (with 1,030 authorships) were able to be used for prediction.

Results

Distribution of Authorships across Countries. Table 2 presents the distribution of the absolute and relative frequencies of countries' authorships. In total, we observed 3,517 authorships by authors affiliated with institutions from 65 countries. As expected, the US is far ahead of all the other countries with more than half the total number of authorships. It is followed by Canada, the Netherlands, and the United Kingdom, all of these having contributed at least 5% of the total number of authorships. Of the remaining countries, only Australia, Germany, and Belgium are responsible for more than 2% of the total number of authorships.

Table 2
Absolute Numbers and Proportions of Authorships by Country

Country	*n(AS)*	%	Country	*n(AS)*	%
United States	1,872	53.2	Poland	6	0.2
Canada	270	7.7	Slovenia	5	0.1
Netherlands	222	6.3	Cyprus	4	0.1
United Kingdom	177	5.0	Japan	4	0.1
Australia	114	3.2	Mexico	4	0.1
Germany	91	2.6	Peru	4	0.1
Belgium	89	2.5	Brazil	2	0.1
Spain	51	1.5	Dominican Rep.	2	0.1
Israel	47	1.3	Estonia	2	0.1
Taiwan	47	1.3	Jordan	2	0.1
Italy	44	1.3	Kenya	2	0.1
Sweden	43	1.2	Luxemburg	2	0.1
Finland	39	1.1	Romania	2	0.1
Austria	37	1.1	Sri Lanka	2	0.1
Greece	25	0.7	Afghanistan	1	<0.1
South Africa	24	0.7	Bahrain	1	<0.1
International	22	0.6	Benin	1	<0.1
New Zealand	22	0.6	Bhutan	1	<0.1
Chile	21	0.6	Bolivia	1	<0.1
France	21	0.6	Colombia	1	<0.1
Norway	21	0.6	Dem. Rep. Congo	1	<0.1
Turkey	20	0.6	Ecuador	1	<0.1
China	19	0.5	Ghana	1	<0.1
Switzerland	19	0.5	Haiti	1	<0.1
Denmark	17	0.5	Malaysia	1	<0.1
Ireland	16	0.5	Palestine	1	<0.1
Portugal	14	0.4	Philippines	1	<0.1
Singapore	11	0.3	Russia	1	<0.1
Burkina Faso	10	0.3	South Korea	1	<0.1
Hungary	9	0.3	Tanzania	1	<0.1
Thailand	8	0.2	Tunisia	1	<0.1
India	7	0.2	Uganda	1	<0.1
Iran	6	0.2	Zimbabwe	1	<0.1
Total *n(AS)* = 3,517					

Note. *n(AS)* = absolute number of authorships per country.

Predicting Authorships across Countries. Due to missing data, we only used 58 of the 65 countries for BRT analysis. These 58 countries, however, account for 97.8% of the total number of authorships observed. Table 3 presents the results of the BRT analysis. As regards overall model information, the R^2 value in the test dataset is especially important because it allows us to assess

whether the model estimated on the training dataset was overfitted. The model was capable of making accurate predictions in the test dataset (R^2 = 64.8%), suggesting that overfitting was not a problem.

Table 3
Relative Influence of Predictors

Area	Predictor	Relative Influence
Research system	Research productivity in social sciences	48.9%
	Age of evaluation society	8.3%
	Number of universities	7.4%
	Size of continental journal market	2.6%
Economic and social/political system	GDP per capita	16.8%
	Control of Corruption index	10.0%
	Civil liberties index	5.7%
	English as an official language	0.2%
Model information	R^2 in test dataset	64.8%
	Optimal number of trees	2,139

When it comes to the influence of the individual predictors, Table 3 shows that research productivity in the social sciences is by far the most important predictor with a relative influence of almost 50%. The second most important predictor with a relative influence of 16.8% is the per capita GDP, followed by the control of corruption index with 10.0% and age of evaluation society with 8.3%. The number of universities and the civil liberties index also have an influence greater than 5%. The variables size of continental journal market and English as an official language possess low predictive power with values below 3% and 1% respectively. When the influences are added together, results suggest that the research-system-related predictors have twice as much influence on predicting countries' output in evaluation journals as the economic and social/political indicators combined.

As regards the nature of the relations between the predictors and ln(AS) (Figure 1), we find that the strongest increase in ln(AS) is associated with research productivity in the social sciences. The functional form of this relation is non-linear and suggests a ceiling effect because ln(AS) strongly increases with changes in research productivity up to 30,000 documents and almost stops increasing at values above that. Apart from some slightly different developments at about 140,000 documents, the conditional effects estimated with and without the US are very similar. With regard to the age of evaluation societies, we find that ln(AS) remains stable from zero up to 15 years. From 15 to 24 years, there is an increase in ln(AS), but after having reached 24 years it stops mounting again. Although the functional form is similar between the models with and without the US up to the age of 18 years, we observe differences in the development above that, suggesting that the effect is somewhat sensitive to the strongest outlier in the dataset.

The conditional effect of the size of the academic sector is almost the same for the datasets with and without the US. In both datasets, the predicted values of ln(AS) decrease with an increase in the number of universities until this number reaches about 1,250. Above that number, the predicted values of ln(AS) remain relatively stable. Finally, we find that the conditional effect of the size of the journal market points in the expected direction. The predicted values of ln(AS) are smallest for countries that have no access to a continental journal market and slightly increase for countries which have access to the European journal market. They increase further when countries have access to the North American market. Yet the graph shows that the conditional effect of this predictor is small, which is true for the models estimated both with and without the US.

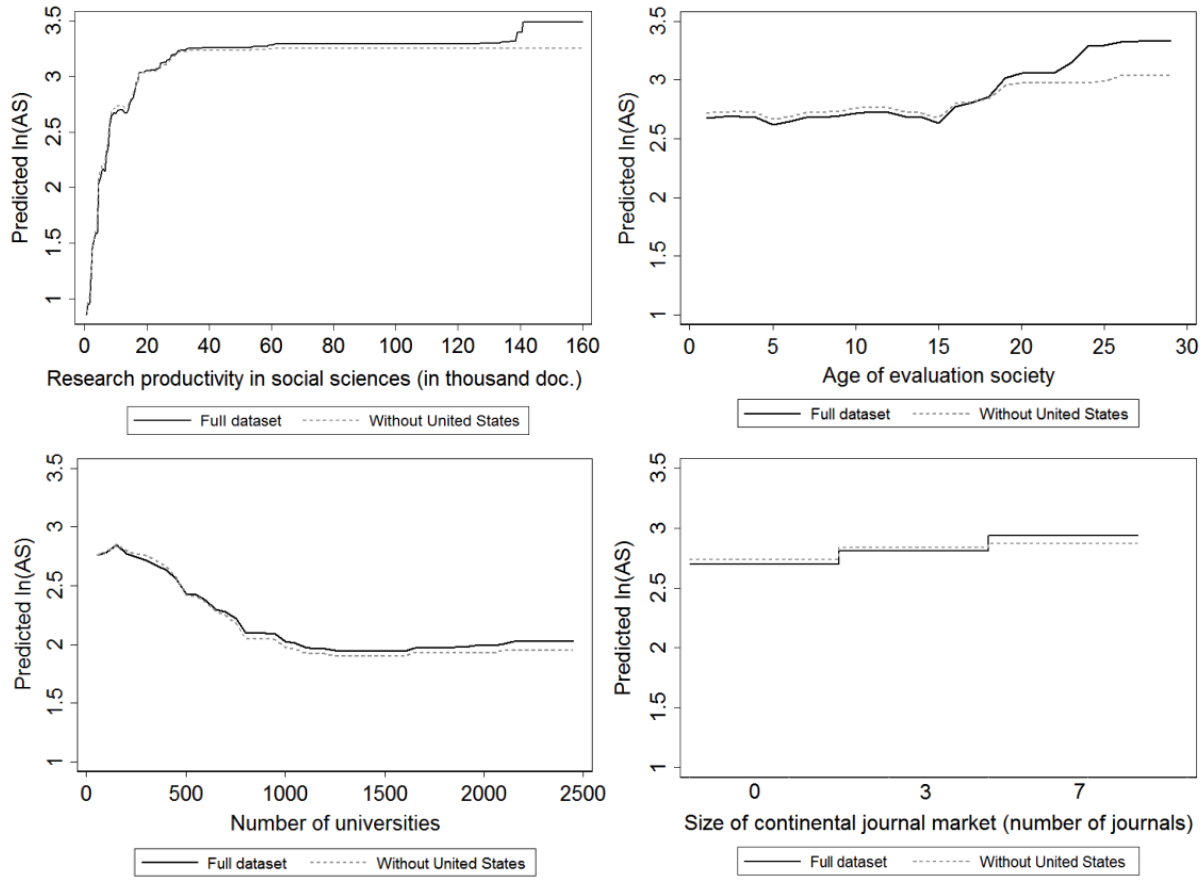

Figure 1. Partial dependence plots (research-system-related predictors)

With regard to the four remaining predictors, the plots in Figure 2 show that the per capita GDP and control of corruption show a positive relation with ln(AS). In both graphs, removing the US from the estimation did not alter the shape of the curves. As regards per capita GDP, we find an increase in the predicted values of ln(AS) from small values of the predictor up to mid-range values of about 42,000 US$. Above this threshold, after a slight decrease, the predicted values of ln(AS) remain stable. On the contrary, however, we find that the predicted ln(AS) is stable for low levels of control of corruption, indicating that there are no differences in ln(AS) between countries where there is only low control of corruption. Yet when the index exceeds the value of 58, we observe a slight but continuous increase in predicted ln(AS).

With regard to civil liberties, we find that countries with the highest value on the civil liberties index show the highest predicted values of ln(AS), whilst countries with the values 2 and 3 produce fewer authorships than countries with the value 1, but more than countries with values greater than 3. Interestingly, the predicted ln(AS) remains stable with values above 3 on the civil liberties index, suggesting a floor effect after having reached a threshold. Here too, we do not find any substantial differences between the models estimated with and without the US. Finally, Figure 2 shows that there are practically no differences in the predicted ln(AS) between countries where English is an official language and countries where it is not.

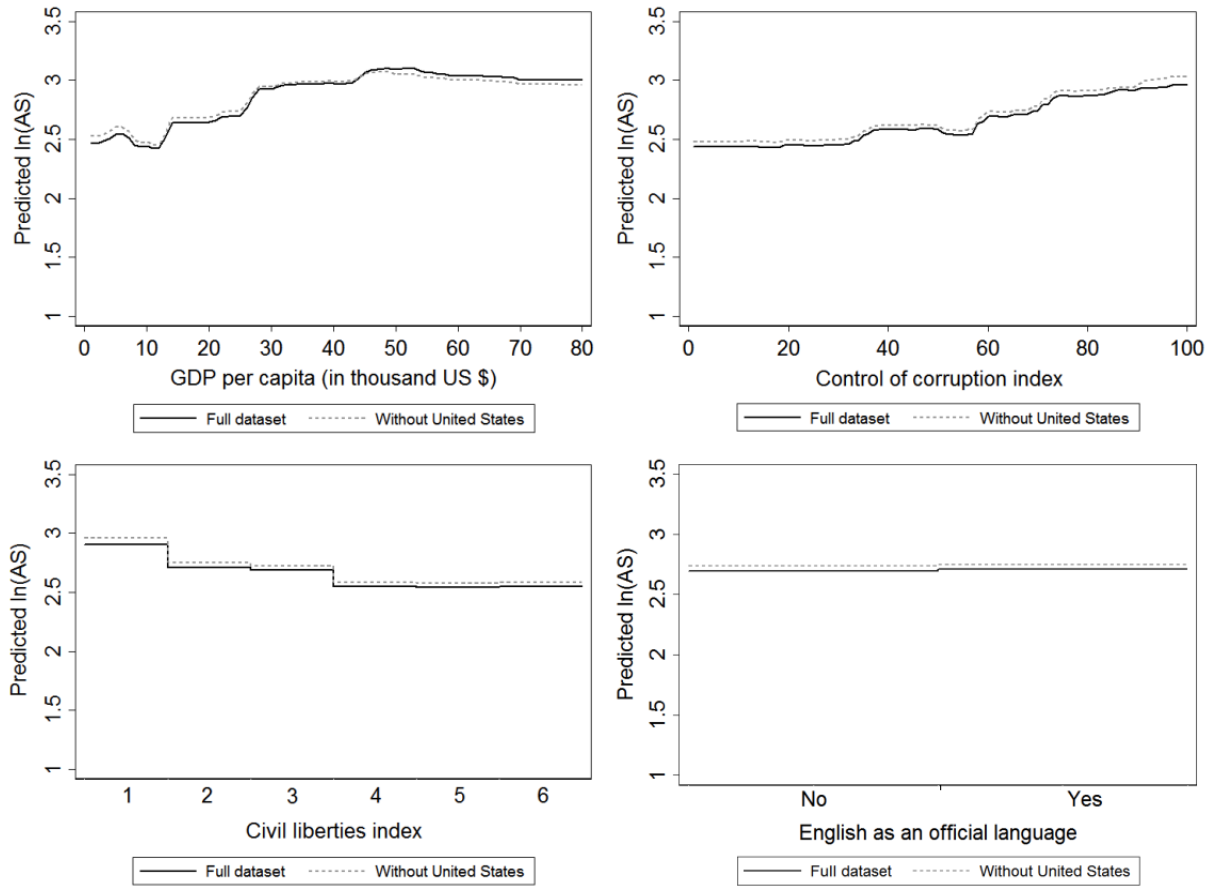

Figure 2. Partial dependence plots (economic/social system-related predictors)

External Validity. Applying our model to data from the years 2005 and 2006 suggests that our originally fitted model is fairly generalizable. We found that the original model–fitted to data from the years 2009 to 2013–was capable of explaining 60.9% of the variance of ln(AS) in the dataset from 2005 and 2006. This means that the model's predictive power only marginally decreased when compared to the proportion of variance explained in the dataset from 2009 to 2013.

Discussion

The results of the empirical analysis mostly correspond with our assumptions. As regards the research-related predictors, the strongest association was observed between *research productivity in the social sciences* and ln(AS). The findings show that the strong increase in the predicted ln(AS) is only present in lower regions of the predictor and that there are practically no further increases in the middle or higher ranges. This suggests that the existence of a certain level of productivity in social science research is a necessary condition for a high output in international evaluation journals, but that its continuous increase does not lead to a continuous increase in productivity in evaluation journals. Yet there were only six countries in our prediction sample which had produced more than 30,000 documents and only two with more than 40,000. Because BRT does not extrapolate on the basis of a pre-specified functional form (as traditional regression models do), the stable curve progression above 30,000 documents is also an expression of the low density of predictor data in this range.

When it comes to the *age of evaluation societies*, we observed a much smaller effect, though it did go in the expected direction. The functional form of the relation can be described in three stages. In the first stage (zero to 15 years), the predicted ln(AS) remains stable, probably because evaluation societies need time to develop the potentials for stimulating evaluation research activities. In the second stage (15 to 24 years), there is an increase in the predicted ln(AS), which

might be explained by the circumstance that after an orientation and development phase, evaluation societies may unfold their potential due to well established networking and cooperation processes. Finally, in the third stage (above 24 years), we observed stable productivity, which may be a hint that the potential for stimulating increases in ln(AS) is virtually exhausted as from a certain age.

The *number of universities* is negatively correlated with the output in evaluation journals until a threshold is reached. This functional form contradicts our assumptions and findings from previous studies (e.g., Meo et al., 2013). An explanation for the absence of a positive association may be that evaluation research is a discipline which is not part of the basic research program of universities. Thus, in contrast to more popular disciplines, the field of evaluation research may not necessarily increase with a rise in the number of universities. Yet this interpretation only explains why there is no positive relation, not why there is a negative one in lower ranges of the predictor. However, we do not have a comprehensive explanation for this finding. Thus, further research in this direction is needed.

In respect of the *size of the continental evaluation journal market* we observed only a small effect on countries' output in evaluation journals, though it did go in the expected direction. This finding suggests that the journals considered in our sample are indeed *international*, not only because they are English-language but also because research output in these journals only differs marginally between countries located in different continental journal markets.

Economic prosperity is the second most influential predictor in our model. The observed effect on ln(AS) is in line with our expectations; countries' output in evaluation journals increases when their per capita GDP rises. This positive association, however, only exists until a threshold of about 42,000 US$ is reached. Afterwards, ln(AS) remains stable despite further increases in per capita GDP. Thus, similarly to the predictor *research productivity in the social sciences*, an increase in the predictor only affects the output in international evaluation journals until a certain degree of economic prosperity is reached. A reason for the cessation of the increase in ln(AS) may be found in the fact that–except for Japan–all the countries in our sample with a per capita GDP greater than 42,000 US$ are part of the "western world" and have similar research systems and traditions.

As regards *control of corruption* and *civil liberties*, we found that our assumptions are supported. We observed that countries with more control of corruption produced more output in evaluation journals than countries with less control of corruption. Moreover, countries with more civil liberties have more contributions in evaluation journals than those with fewer civil liberties. However, for the output in evaluation journals it is irrelevant whether countries are extremely corrupt, or only very corrupt. Similarly, publishing in evaluation journals does not depend on whether countries have no civil liberties or only very few. Both of these findings suggest that certain levels of corruption control and civil liberty have to be attained if evaluation research activities are to become manifest in evaluation journals.

Finally, the results do not support our expectation regarding *linguistic boundaries*. It did not make any difference in terms of the productivity in evaluation journals whether English was an official language of a country or not. One reason for this finding may lie in the nature of the indicator employed, because it does not consider the individual language abilities of researchers. Another reason may be that linguistic boundaries simply do not exist and that the journals in our sample do indeed constitute an international journal market.

Limitations

Our study has two limitations that need to be discussed. First, the relations between the predictors and the dependent variable are correlative and must not be interpreted in a strictly causal manner. As is the case with many predictive models, relevant predictors may have been omitted in our model. Relations observed may to some extent thus represent spurious effects which are in fact caused by unobserved third variables. This circumstance does not lower the predictive validity of our model, but it does have implications for drawing conclusions about how to increase research productivity in international evaluation journals.

A second limitation concerns the generalizability of our model. Although the original model fitted on data from the years 2009 to 2013 predicted ln(AS) from 2005 and 2006 well with only marginal loss in predictive accuracy, it may perform less well with data collected from the more distant past or the future. Moreover, we only considered ten international peer-reviewed evaluation journals. Thus, we cannot exclude the possibility that our model would provide different estimates when more or other journals were considered, let alone evaluation-related articles

published in journals located outside the area of evaluation research.

Conclusion

This research was devoted to developing and testing a model for the prediction of countries' research output in international evaluation journals. We included eight predictors from the research, economic, and social/political system in our model and found that it provided accurate predictions. This was even true when the fitted model was tested with data for another period of time, suggesting that the model is externally valid to some degree.

Our main conclusion from the study is that the research productivity of countries in international evaluation journals can be predicted fairly well by using standard macro-level indicators capturing aspects of the research, economic, and social/political systems of countries. One reason for the good model performance is the application of BRT, which considerably increased predictive power when compared to traditional OLS regression. However, using BRT does not overcome the problem of omitted variable bias, which is why we could not draw any reliable causal inferences. In order to deal with this issue, future research could try to identify the causal mechanisms behind the correlational relations we found in our study. Moreover, with regard to external validity, it would be interesting to see whether or not our model also works for predicting countries' research output in areas other than evaluation research.

References

Basu, A. (2010). Does a country's scientific 'productivity' depend critically on the number of country journals indexed? *Scientometrics*, *82*, 507–516.

Canagarajah, A. S. (2002). A geopolitics of academic writing. Pittsburgh, PA: University of Pittsburgh Press.

Christie, C. A., & Fleischer, D. N. (2010). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation*, *31*, 326–346.

Coryn, C. L. S., Ozeki, S., Wilson, L. N., Greenman II, G. D., Schröter, D. C., Hobson, K. A., Azzam, T., & Vo, A. T. (2016). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. *American Journal of Evaluation*, *37*, 159-173.

Diaz-Puente, J. M., Cazorla, A., & Dorrego, A. (2007). Crossing national, continental, and linguistic boundaries: Toward a worldwide evaluation research community in journals of evaluation. *American Journal of Evaluation*, *28*, 399–415.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

Flowerdew, J. (1999). Writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing*, *8*, 123–145.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*, 367–378.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*, 1365–1381.

Furubo, J.-E., & Sandahl, R. (2002). A diffusion perspective on global developments in evaluation. In J.-E. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 1–23). New Brunswick, NJ: Transaction Publishers.

Gutiérrez, J., & López-Nieva, P. (2001). Are international journals of human geography really international? *Progress in Human Geography*, *25*, 53–69.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer New York.

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation*, *31*, 24–44.

Husted, B. W. (1999). Wealth, culture, and corruption. *Journal of International Business Studies*, *30*, 339–359.

Jacob, S., Speer, S., & Furubo, J.-E. (2015). The institutionalization of evaluation matters: Updating the International Atlas of Evaluation 10 years later. *Evaluation*, *21*, 6–31.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2011). The worldwide governance indicators: Methodology and analytical issues. *Hague Journal on the Rule of Law*, *3*, 220–246.

Lane, P. R. (2011). Innovation and financial globalization. In J. Y. Lin & B. Pleskovic (Eds.), *Annual World Bank conference on development economics – 2010: Lessons from East Asia and the global financial crisis* (pp. 309–332). Washington D.C.: The World Bank.

Leeuw, F. L., & Furubo, J.-E. (2008). Evaluation systems: What are they and why study them? *Evaluation*, *14*, 157–169.

Love, A., & Russon, C. (2000). Building a worldwide evaluation community: Past, present, and future. *Evaluation and Program Planning*, *23*, 449–459.

Markiewicz, A. (2008). The political context of evaluation: What does this mean for independence and objectivity? *Evaluation Journal of Australasia*, *8*, 35.

Meo, S. A., Al Masri, Abeer A., Usmani, A. M., Memon, A. N., Zaidi, S. Z., & Preis, T. (2013). Correction: Impact of GDP, spending on R&D, number of universities and scientific journals on research publications among Asian countries. *PLoS ONE*, 8.

Nielsen, S. B., & Winther, D. M. (2014). A Nordic evaluation tradition? A look at the peer-reviewed evaluation literature. *Evaluation*, *20*, 311–331.

Origgi, G., & Ramello, G. B. (2015). Current dynamics of scholarly publishing. *Evaluation Review*, *39*, 3–18.

Ramsden, P. (1994). Describing and explaining research productivity. *Higher Education*, *28*, 207–226.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a Stata plugin. *The Stata Journal*, *5*, 330–354.

Short, J. R., Boniche, A., Kim, Y., & Li, P. L. (2001). Cultural globalization, global English, and geography journals. *The Professional Geographer*, *53*, 1–11.

Tanzi, V., & Schuknecht, L. (2000). *Public spending in the 20th century: A global perspective*. Cambridge, UK: Cambridge University Press.

Vinluan, L. R. (2012). Research productivity in education and psychology in the Philippines and comparison with ASEAN countries. *Scientometrics*, *91*, 277–294.