Field Effects in Predicting Exceptional Growth in Research Communities

Richard Klavans[*], Kevin W. Boyack[**] and Caleb Smith[***]

[*]*rklavans@mapofscience.com*
0000-0002-9118-8955
SciTech Strategies, Inc., USA

[**] *kboyack@mapofscience.com*
0000-0001-7814-8951
SciTech Strategies, Inc., USA

[***] *caleb.smith@gmail.com*
0000-0002-9734-9705
University of Michigan Medical School, USA

Using a model of the literature indexed in Scopus, we have increased the accuracy of our ability to predict which of 20,747 research communities would achieve exceptional growth from 32.2 to 39.6 using double exponential smoothing of inertial indicators and by doing predictions in each of 26 fields rather than across the entire model. Each field nominated two (out of a possible 123) indicators as 'best predictors' following the procedure described in our previous studies. Significant diversity was found in which indicators performed best in each field, suggesting that field effects should be accounted for in predictive analytics. Nevertheless, there were groupings of contiguous fields with a surprising level of homogeneity in predictive indicators. Possible reasons for the similarities and differences are discussed.

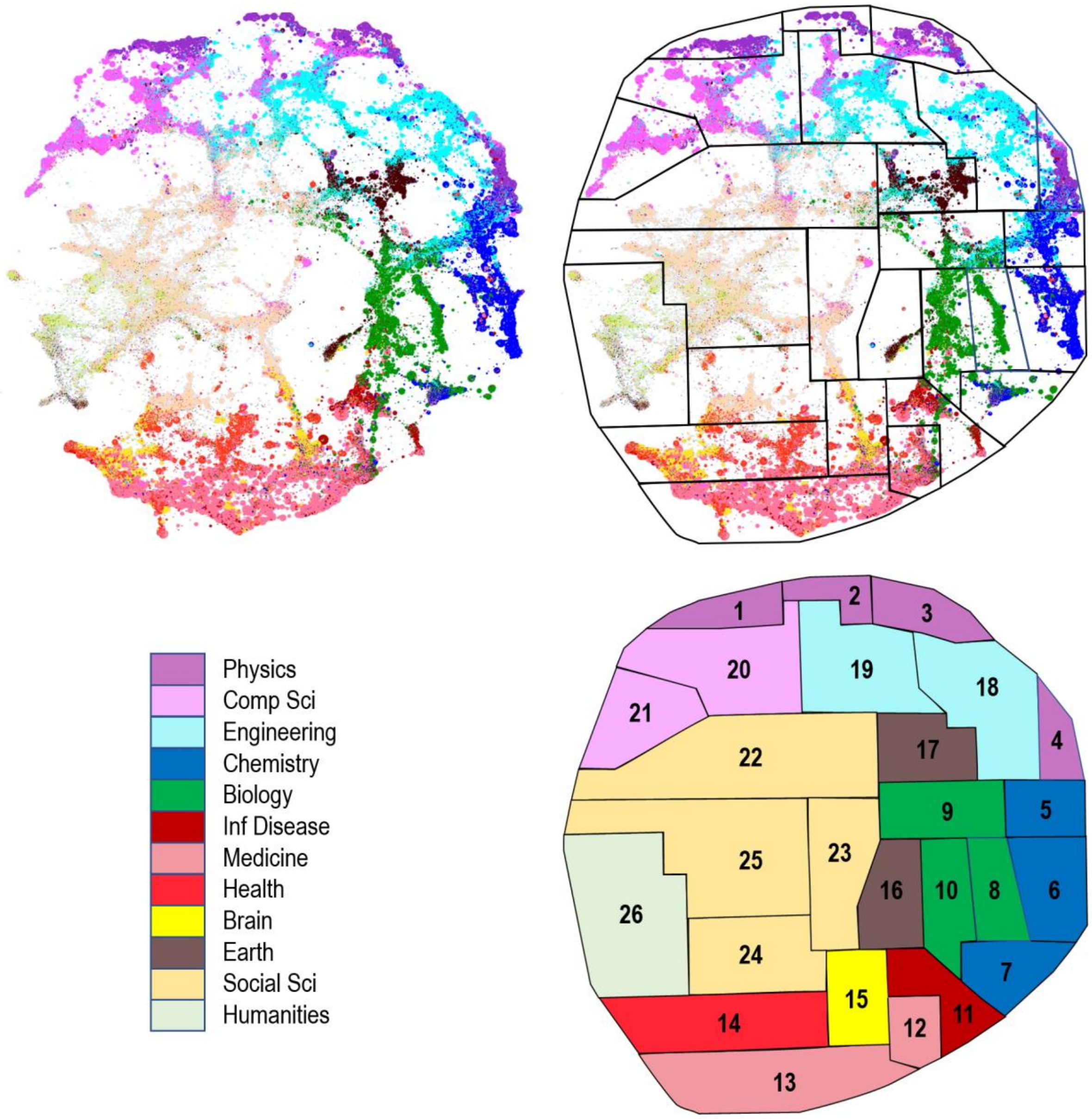
1. Introduction

Predictive analytics is a frontier area in scientometrics, one which we expect will attract increasing attention in coming years. Our recent work in this area has been to predict which research communities in the global structure of science will achieve exceptional growth three or four years forward (Klavans, Boyack, & Murdick, 2020). Most recently we used a model of the entire Scopus database in which 36.6M indexed documents from 1996-2016 were assigned to nearly 100k research communities (RCs) using a new iterative clustering methodology based on direct citation (Boyack & Klavans, 2022) and employing the Leiden algorithm (Traag, Waltman, & Van Eck, 2019). Documents from 2017-2019 were then added to the RCs using their citation patterns, and over 80 indicators from the 2016 baseline were used to predict RC growth as of 2019 in the 20,747 RCs that were large enough to consider. Two indicators were found to predict those RCs that would achieve an 8% annual growth rate over three years with a threat score (true positives divided by the sum of true positives, false positives and false negatives) of 32.2. The first indicator, an inertial composite based on numbers of papers, references, citations and authors, achieved a 29.8 threat score on its own. The second indicator was the number of a group of authors we call visiting foxes – those that publish in many RCs but who are 'visiting' (or publishing less in) the RC of interest – added over 2 points to the score. We have continued to seek improvements in our ability to predict exceptional growth. This paper reports on recent progress made using two improvements: exponential smoothing of inertial indicators and incorporation of field effects.

2. Background

We build upon Kuhn's assumption that researchers are members of RCs that work on research problems. These communities tend to be much smaller than fields; Kuhn mentioned that an RC might be populated by 100 researchers. We take this to mean 100 core researchers – those that publish regularly in the community. He also argued that these communities could be detected by looking at citation links as a reflection of the communication patterns between researchers (Kuhn, 1974). To operationalize this theory, we cluster the entire Scopus database using direct citation analysis in a way that yields roughly 100k clusters that contain, on average, about 100 core researchers. Communities identified in this way tend to be robust in terms of citation density, with nearly half of the (> 1 billion) citation links ending up between papers in the same RC. As one might expect, larger RCs tend to have a higher citation density while smaller RCs tend to have a lower citation density. We then use textual similarity between RCs with the DrL graph layout routine to create a map that shows the relationship between these communities. Figure 1 (upper left) shows the resulting map of RCs where each is colored using its dominant high-level discipline based on journals (Börner et al., 2012).

Figure 1: Identifying 26 fields (sectors) of research using a map of the Scopus database.



To investigate field effects, RCs must be grouped according to some logic that approximates fields. One way to do this would be to simply use our discipline (color) groupings. However, this ignores the multidisciplinary nature of several sections of the map, and we wanted a larger number of groups. Thus, we have identified sectors (hereafter, fields) geographically

(Figure 1, upper right) in a manner similar to how one might divide a continent, composed of cites with diverse languages, into nations. Each of the 12 general disciplines is represented. As examples, physics (dark purple) is broken down into 4 highly concentrated (and separately located) fields. The humanities (light green) are represented by only one field that is sparsely populated by RCs. Some fields are nearly mono-disciplinary such as #21 (mostly computer science). Others contain substantial components from multiple disciplines such as #14 which contains large numbers of medicine and brain science RCs in addition to the health sciences RCs that predominate.

3. Results

3.1. Exponential smoothing

Our previous work showed that inertial (historical) indicators were far better at predicting exceptional growth than annual point values. These were calculated using vitality (Boyack & Klavans, 2022), a measure that discounts annual indicator values by age. We wanted to explore the possibility that different ways of dealing with time series indicators might improve our results. After experimenting with different methods, we found that an indicator based on Holt-Winters double exponential smoothing (Holt, 1957; Winters, 1960) of the time series of counts was a better predictor than our previous best, raising the threat score from 29.8 to 33.1 for the single indicator. This indicator, which we call XDErp, is a composite of three individual double exponentially smoothed indicators that approximate growth in papers, references and exports. Coupling this with our visiting foxes (named A6) indicator in a two variable predictor raised the threat score to 34.5 across the 20,747 RCs mentioned previously.

3.2. Field effects

In this study we used the same stepwise regression method and same dependent variable (binary [0,1] indicating that the RC achieved 8% annualized growth or not) that were used in our previous study (Boyack & Klavans, 2022). However, in this study we not only investigated many more variables (123 rather than 81), but we also did the regressions for each of the fields identified in Figure 1 rather than for the entire model. The additional variables included the double exponentially smoothed indicator as well as variations of existing vitality indicators using shorter time series lengths (3 and 7 years rather than 10) and a new composite indicator (C1) created using machine learning over the set of indicators.

Each field was allowed to identify two (of the 123) indicators as potential predictors of exceptional growth. This selection process was done in a hierarchical fashion. The indicator that generated the best threat score was selected first and then the indicator that could improve the threat score the most was selected second. For 12 fields, the best overall indicator was the same as the indicator that was the best for the entire model (XDErp) – the composite indicator based on double exponentially smoothed counts. However, none of these 12 fields nominated the best second indicator (A6) from the entire model.

Table 1 lists the indicators and corresponding threat scores for the 12 fields for which XDErp was the best single predictive indicator. Numbers of RCs, RCs with exceptional growth (XG), and the associated discipline are listed for each field. The top two indicators for each field are then listed along with the threat scores corresponding to the single best variable (TS1) and the combination of the two best variables (TS1+2). Also listed is the gain (ΔTSorig) over what the threat score would have been using the two best variables (XDErp/A6) for the entire model. For example, for field #20, the threat score using the two original variables, XDErp and A6, is 45.1 (not shown), which is 2.5 points less than that obtained using XDErp and N8.

Table 1. Threat scores by field for the 12 fields where XDErp is the best predictor.

Field	**#RC**	**XG**	**Discipline**	**lndicators**	**TS1**	**TS1+2**	**ΔTSorig**
20	1400	208	Comp Sci	XDErp/N8	44.4	47.6	2.5
12	954	66	Medical	XDErp/Nd	39.8	45.6	6.1
19	1628	214	Engineering	XDErp/P0	41.5	43.3	0.0
21	815	103	Comp Sci	XDErp/P0	38.7	41.5	3.0
1	746	29	Physics	XDErp/vFv3	35.2	39.6	5.1
22	744	110	Social Sci	XDErp/C1	37.0	38.0	0.7
11	765	37	Inf Disease	XDErp/A5	29.2	36.2	9.2
18	1361	108	Engineering	XDErp/G1	33.0	36.2	2.7
17	1177	89	Earth	XDErp/Ac	33.5	35.8	1.7
8	347	17	Biology	XDErp/N0	22.9	33.3	11.1
14	1719	99	Health	XDErp/I6	25.2	27.0	3.1
25	554	25	Social Sci	XDErp/N0	16.7	18.5	4.2
	12210	1105	Total		36.9	39.7	2.6

Six of the second indicators are associated with the characteristics of authors. This includes the characterization of authors with different publication strategies (A5; Ac; vFv3); the number of papers with an industry author in 2016 (P0) and the inertial growth rate in male authors (G1). Four of the second indicators were based on RC-level network characteristics, including the percentage of links within RCs (N0), number of triangles (N8) and degree centrality (Nd). Descriptions of most indicators appearing in Table 1 and Table 2 are provided in the Appendix of (Boyack & Klavans, 2022).

Table 2 lists the fields where the first variable was not XDErp. As for the 12 fields in Table 1, the best field-level two-indicator threat score (TS1+2) was then compared to the threat score obtained using the two original indicators, XDErp and A6. For example, field 4 had a threat score of 51.9 using its top two indicators, while it would have had a threat score of only 32.3 (not shown) if it had been forced to use XDErp and A6 as indicators. This is the most extreme example, but nonetheless illustrates the improvement in threat score when each field is considered separately. There were four fields with less than 10 cases of exceptional growth that were not considered due to insufficient sample size to provide meaningful results. The average gain in threat score from using field-level indicators is greater for those fields that chose a different first indicator (6.3) than for those that chose the overall best indicator (XDErp) as their first indicator (2.6).

Table 2. Threat scores by field for fields where XDErp is not the best predictor.

Field	**#RC**	**XG**	**Discipline**	**lndicators**	**TS1**	**TS1+2**	**ΔTSorig**
4	409	16	Physics	G1/A2	48.1	51.9	19.6
5	887	110	Chemistry	vA7/A6	44.0	46.0	3.7
6	837	42	Chemistry	G1/N9	36.4	41.3	7.1
15	854	35	Brain	vA7/N3	33.3	39.1	12.0
7	639	55	Chemistry	vA7/A1	31.4	36.3	8.8
10	543	19	Biology	vXs7/I6	33.3	36.1	0.0
3	665	27	Physics	L4/A6	33.3	35.3	2.6
24	445	40	Social Sci	vXs7/C1	33.3	34.7	5.2

13	2131	39	Medical	L3/L4	32.4	33.8	8.5
23	425	37	Social Sci	A4v/C1	25.7	30.6	3.6
2	255	9	Physics	insufficient sample size			
9	206	3	Biology	insufficient sample size			
26	185	2	Humanities	insufficient sample size			
16	56	2	Earth	insufficient sample size			
	8537	436	Total		36.0	39.2	6.3

As was the case for the 12 fields in Table 1, the most predictive indicator for the 10 fields with sufficient sample size in Table 2 was an inertial (time series) indicator. However, for six of these cases it was based on counts related to authors rather than counts related to papers. The best predictive indicator type (paper or author) is shown for each field in Figure 2. It is interesting that four of the fields where an author-based indicator performed best are contiguous – fields #4-7 at the right-hand side of the map. These include the three fields focused on chemistry and one that is a physics/chemistry mix centered on materials science. We don't know what it is about chemistry as a disciplinary culture that leads to authors being so strongly correlated with growth, but simply note it here as a curiosity and leave details for future research.

Figure 2: Best and second-best indicator types for each field; paper (P), author (A).

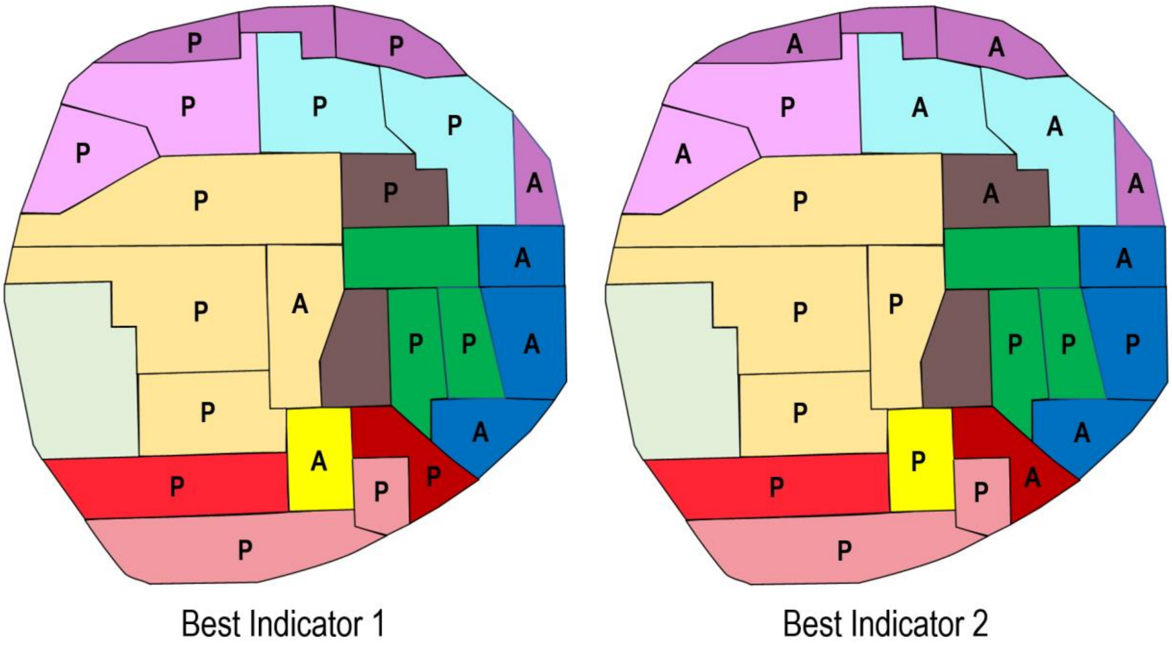
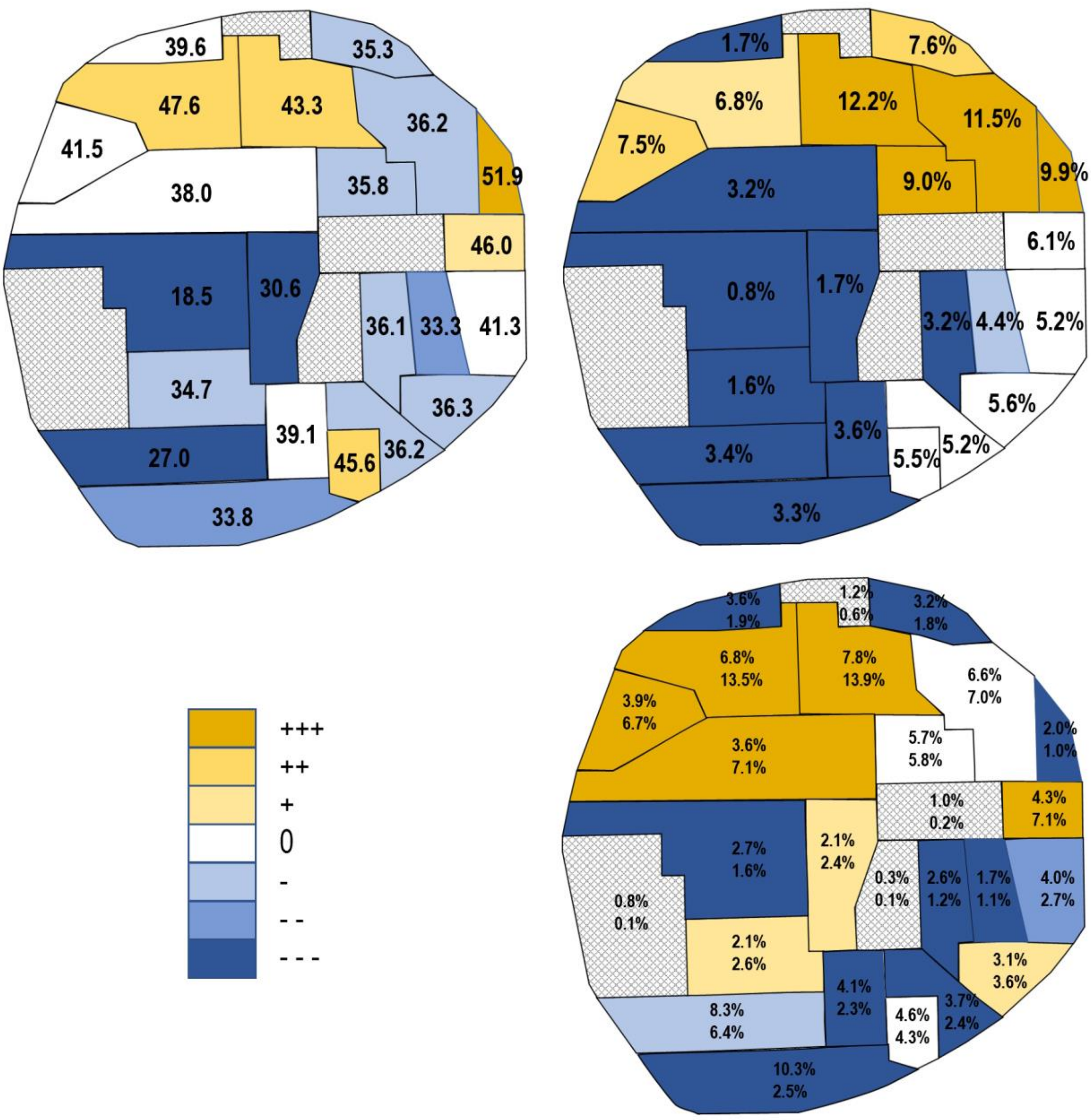


Figure 2 also shows whether the second-best indicator for each field is paper-based or author-based. Three of the four fields (#4,5,7) where an author-based indicator was the best predictor also have an author-based indicator as the second-best predictor. However, many other fields (in engineering, physics, computer science and earth science) also have an author-based indicator as their second choice. This is consistent with an author-based indicator being the second indicator for the overall model.

The overall gain in threat score is as follows. If the overall best two indicators (XDErp and A6) are used for the entire sample of 20,747 research communities, the threat score is 34.5. If the same two indicators are applied on a field-by-field basis (with four fields dropped from the analysis due to sample size) the threat score increases to 36.1. If each field chooses their own two indicators (and the same four fields are excluded) the threat score increases to 39.6.

Figure 3: Threat scores (upper left), industry authorship percentages (upper right) and RC distributions (lower right) by field. For RCs, the upper number is the percentage of RCs in the field while the lower number is the percentage of extreme growth RCs in the field. Colors indicate how far above or below average the values are by field for each panel.



4. Discussion and Implications

There are three fields (#19-21, engineering and computer science) that seem to be similar in that they are contiguous to each other on the map, have higher than average levels of industry concentration (see Figure 3) and have high threat scores that are less affected by field effects. The percentage of papers with an industrial author in these three fields is 9.1% (the global percentage is 5.6%). If we constrain each field to only use XDErp and A6, the weighted average threat score is 43.1. The weighted average threat score when these four fields use separately chosen indicators is 44.7. It is also interesting that these four fields have higher than expected numbers of extreme growth (XG) communities (Figure 3, lower right). For instance, field #19 only contains 7.8% of the RCs across the entire map but contains 13.9% of the RCs that experienced extreme growth. Thus, these four fields contain over 34% of all cases of extreme growth while they only contain 18.6% of the RCs. Extreme growth may be generally more pronounced in fields with higher industry participation.

However, there are counter examples. For instance, field #22 has nearly twice as many extreme growth RCs as expected while only having industry participation of 3.2%, well below average. While this field is dominated by social sciences, it does contain quite a few RCs from computer science and engineering, and the social science represented are primarily from economics and business. Thus, the mathematical and statistical underpinning of this field may be why it has so many extreme growth RCs despite not having high industry participation. Medicine-related fields (#11-15) have lower than average industry involvement, and also have lower than expected numbers of extreme growth RCs.

The four fields with low numbers of extreme growth events (those with the cross-hatch shading in Figure 3) only comprise 3.3% of the RCs in the entire model. Thus, we are not overly concerned about the lack of ability to predict which RCs will achieve extreme growth in these fields separately. They can be modelled as part of the whole.

In summary, we have increased the accuracy of our ability to predict which of 20,747 research communities would achieve exceptional growth from 32.2 to 39.6 using double exponential smoothing of inertial indicators and by doing predictions in each of 26 fields rather than across the entire model. The fact that none of the 26 fields nominated the same two predictive indicators that worked best over the overall model, and the relatively large gain in accuracy, together suggest that field effects should be considered in predictive analytics.

Open science practices

Our model and this analysis are created using Scopus data. Given that we license these data from Elsevier, paper-level data cannot be shared. In principle, however, we are strong supporters of open science practices. For instance, we have created a similar model using PubMed data that is openly available (Boyack, Smith, & Klavans, 2020). While we have not yet replicated the prediction analysis from this study on our PubMed model, we expect that this could be done with similar results. We have plans to do this and will make those results openly available at that time.

Author contributions

Richard Klavans: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing. Kevin W. Boyack: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing. Caleb Smith: Formal analysis, Writing.

Competing interests

The authors have no competing interests.

Funding information

This work was funded by a contract from the Center for Security and Emerging Technologies (CSET) at Georgetown University.

References

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., . . . Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464. doi:10.1371/journal.pone.0039464

Boyack, K. W., & Klavans, R. (2022). An improved practical approach to forecasting exceptional growth in research. *Quantitative Science Studies*. doi:10.1162/qss_a_00202

Boyack, K. W., Smith, C., & Klavans, R. (2020). A detailed open access model of the PubMed literature. *Scientific Data, 7,* 408. doi:10.1038/s41597-020-00749-y

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Memorandum, Vol. 52, Carnegie Institute of Technology, Pittsburgh*.

Klavans, R., Boyack, K. W., & Murdick, D. A. (2020). A novel approach to predicting exceptional growth in research. *PLoS ONE, 15*(9), e0239177. doi:10.1371/journal.pone.0239177

Kuhn, T. S. (1974). Second thoughts on paradigms. In F. Suppe (Ed.), *The Structure of Scientific Theories* (pp. 459-482). Urbana: University of Illinois Press.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*, 5233. doi:10.1038/s41598-019-41695-z

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science, 6*(3), 324-342. doi:10.1287/mnsc.6.3.324