Exploring predictors of scientific performance with decision tree analysis: The case of research excellence in early career mathematics

Jonas Lindahl*

*jonas.lindahl@umu.se
Inforsk (Department of sociology), Umeå University, Umeå, SE-901 87 (Sweden)

ABSTRACT
The purpose of this study was (1) to introduce the exploratory method of decision tree analysis as a complementary alternative to current confirmatory methods used in scientometric prediction studies of research performance; and (2) as an illustrative case, to explore predictors of future research excellence at the individual level among 493 early career mathematicians in the sub-field of number theory between 1999 and 2010. A conceptual introduction to decision tree analysis is provided including an overview of the main steps of the tree-building algorithm and the statistical method of cross-validation used to evaluate the performance of decision tree models. A decision tree analysis of 493 mathematicians was conducted to find useful predictors and important relationships between variables in the context of predicting research excellence. The results suggest that the number of prestige journal publications and a topically diverse output are important predictors of future research excellence. Researchers with no prestige journal publications are very unlikely to produce excellent research. Limitations of decision three analysis are discussed.

INTRODUCTION
Bibliometric indicators are increasingly used as decision support tools in academia (Abramo, Cicero, & D'Angelo, 2013). Indicators that are used as decision support tools should satisfy at least two basic assumptions: (1) indicators of past performance should be able to predict future scientific performance (e.g., Danell, 2011); and (2) indicators should be free from inherent biases (e.g., Moed, 2005). These assumptions have been tested in numerous scientometric prediction studies at different levels of aggregation (e.g., Jensen, Rouquier, & Croissant, 2009; Penner, Pan, Petersen, Kaski, & Fortunato, 2013; Dubois, Rochet, & Schlenker, 2014; Havemann & Larsen, 2015).

Most scientometric prediction studies are conducted in a confirmatory framework based on testing hypothesis. In this study I suggest the framework of exploratory data analysis (EDA), and specifically decision trees, as an underutilized source of methods that could complement the dominating confirmatory framework (Strobl, Malley, Tutz, & Maxwell, Scott, 2009; King, & Resick, 2014).

Decision trees has a number of desirable features in the context of predicting research performance from the micro to the macro level. Decision trees (King & Resick, 2014): (a) are non-parametric; (b) are flexible and can handle many different variable types; (c) can identify non-linear relationships; (d) can identify useful predictors; (e) show interactions between

predictors without the need to specify these in advance; and (f) are transparent, intuitive to interpret, and can be used as visual tools to inform decision making.

Publication praxis and citation behaviours differ between scientific fields (Moed, 2005). Decision trees provide an opportunity to explore research fields with few prior assumptions. Such data driven exploration can generate new hypothesis to test with confirmatory methods. Decision tree analysis may prove particularly useful for the study of peripheral and emerging areas of research where knowledge is scarce to begin with.

The purpose of this study is (1) to introduce the exploratory method of decision tree analysis as a complementary alternative to current methods used in scientometric prediction studies of research performance; and, as an illustrative case, (2) to identify important predictors, interactions between predictors, and the effect of combinations of publication track record characteristics to predict research excellence at the individual level among 493 early career mathematicians in the sub-field of number theory between 1999 and 2010.

METHOD

Data collection

The dataset consisted of article publication track records of 493 authors in number theory. The authors were selected on the basis of (1) at least one published article in class 11 (i.e., Number theory) in the Mathematics Subject Classification (MSC) scheme between 1999 and 2003; (2) an article publication career of ≥ eight years; and (3) that the share of articles belonging to the MSC class in the track record of an author was ≥ than the share of any other MSC class found in that authors track record.

Publications (*N*=4654) was retrieved from the MathSciNet (MSN) database and matched against publications indexed in the citation indices available through Web of Science (WoS) Core Collection to obtain citation data. The final dataset retrieved from WoS consisted of 2975 articles, reviews, notes and letters.

Design and variables

The design comprise two time periods: Period 1 (P1) and Period 2 (P2). P1 is the first four years in the publication career of an author. P2 is the fifth to the eighth year. The publication career of an author begin with the first MSN article publication.

The response variable consisted of a binary variable indicating if an author publish ≥ one article in P2 that can be considered excellent (i.e., an excellent researcher). An article is defined as excellent if it has a document type, publication year, and field normalized citation score (FNCS) adjusted for multiple Web of Science Categories ≥ the 90th percentile given a reference set (Lundberg, 2007). I used the article FNCSs of the publications of the 493 authors in P2 as a reference set to calculate the percentile.

Seven predictors were used in the analyses. Predictor:

(1) address publication rate in p1 (coding: *P*) and consist of the number of MSN articles in P1;
(2) address journal prestige and consist of the number of articles published in journals with a SNIP value ≥ the 75th percentile in the CWTS Journal Indicators list (CWTS,

2015; Waltman, Van Eck, Van Leeuwen, & Visser, 2012) in the publication year of the article in P1 (coding: *Top_Jour*);
(3) address collaboration and consist of the average number of authors per publication during P1 (coding: *Avg_Co_Au*);
(4) address topical diversity (coding: *Topic_Div*) and consist of the number of different MSC classes an author has published in during P1;
(5) address mobility in P1 and consist of the number of MSN publications at different universities (coding: *Mob_Univ*);
(6) address university prestige (coding: *Top_Univ*) and is a binary predictor indicating if an author has published ≥ 1 publication at a top university in P1;
(7) is a binary predictor and address whether an author has published ≥ 50% of her/his output at institutions located in English speaking countries during P1 (coding: *English*).

Decision tree analysis
A decision tree is built by an algorithm that successively split the initial dataset into smaller sub-groups based on splitting rules (King, & Resick, 2014). When the predictors and the response variable are chosen the decision tree is built in three main steps.

In step one all predictors are evaluated to find the "best" binary split. The decision tree algorithm starts with the total dataset. The best binary split is a cut-off threshold among the values in the chosen predictor. When the cut-off is determined the dataset is divided in two sub-groups. The goal of splitting is to assign authors with similar values in the response variable in the same group so that the two sub-groups are more homogenous than the previous group (King, & Resick, 2014).

In step two the splitting procedure in step one is performed on the total dataset and two sub-groups are created. Each sub-group is treated as a new dataset; a cut-off for the best binary split is determined and two sub-groups is created. Successively the decision tree algorithm creates smaller and more homogenous groups. All predictors are evaluated for each potential split (King, & Resick, 2014).

In the third step the splitting procedure is ended by some stopping criteria. The definition of the best split and the stopping criteria depends on the decision tree algorithm. In this study I used an implementation of the algorithm for conditional inference trees (Hothorn, Hornik, & Zeileis, 2006) in the Party package available through R (R Core Team, 2015). With this algorithm the best binary split is determined by testing the global null hypothesis of independence between each of the predictors and the response variable with permutation tests and further between the each possible binary sub-set of the chosen predictor and the response (Hothorn, Hornik & Zeileis, 2006). The predictor and cut-off threshold resulting in the strongest association with the response variable (i.e., lowest p value) are chosen for the split. The splitting procedure stop when the global null hypothesis of independence between all possible combinations of the predictors and the response are rejected at some pre-specified level of alpha (e.g., 0.05; Hothorn, Hornik & Zeileis, 2006). Predictors that has not been chosen for a split when the splitting procedure stops are not included in the tree model.

The performance of decision tree models are usually evaluated with the statistical method of cross-validation (Maimon & Rokach, 2008). The basic principle of cross-validation is to split

the initial dataset into a training set and a test set. The decision tree model is first trained on the training set and then validated on the test set. The purpose of cross-validation is to evaluate the generalizability of the decision tree model by fitting the trained model on new data (i.e., the test set) (Maimon & Rokach, 2008). By fitting the trained model on new data a more realistic performance measure can be obtained since models tend to have a better fit on the training set due to overfitting, compared to a new dataset that is sampled from the same population as the training set (Maimon & Rokach, 2008).

In this study I tested the model with the commonly used method of 10-fold cross-validation, where the dataset is split into 10 equally sized and non-overlapping folds (i.e., sub-groups) (Maimon & Rokach, 2008). The cross-validation consists of ten iterations. In each iteration one fold is used as the test set and the remaining nine folds are used as the training set. In each iteration some appropriate metric of model performance is calculated. As a result of the 10-fold cross-validation the ten values of the chosen metric are averaged to produce a single cross-validated performance measure (Maimon & Rokach, 2008).

RESULTS

Decision tree analysis: Predictors of research excellence among early career mathematicians

Figure 1 depicts a decision tree which consist of a single root node (oval) at the top, a number of internal nodes (ovals) and several terminal nodes (bar charts) at the bottom. The nodes in the tree is connected by branches. Each split is represented by a predictor label visible in the node denoting which predictor was used for the split (e.g., *Top_Jour*). The value at which the best split occurred in the predictor is placed along the branches between nodes. The splitting procedure stops at the terminal nodes. Each terminal node provide a bar chart indicating the proportion of authors in each class and the *n* of authors in that group.

32 authors had missing values on the affiliation based predictors and was excluded from the analysis. The analyses was performed with 461 authors. Of the 461 authors, 71 was defined as excellent (i.e., incidence=15.4%). The binary response variable and seven predictors was used as input for the decision tree depicted in Figure 1: *P*; *Top_Jour*; *Avg_Co_Au*; *Topic_Div*; *Mob_Univ*; *Top_Univ*; *English*. However, only three of these predictors, *Top_Jour*, *Topic_Div,* and *English* was actually used in the tree. This indicates that *P, Avg_Co_Au*, *Mob_Univ,* and *Top_Univ*, did not contribute to the model.

I used the *10-fold cross validated area under the receiver operating characteristic* (ROC) *curve* (AUC) to evaluate the prediction accuracy of the decision tree (King & Resick, 2014). The AUC is not sensitive to skewed class distribution which make it an appropriate metric in the context of bibliometric data (Maimon & Rokach, 2008). The decision tree model had a cross-validated AUC of 0.73, indicating acceptable discrimination between excellent authors and non-excellent authors according to the rule of thumb interpretation of AUC-values suggested by Hosmer and Lemeshow (2000).

Figure 1. Decision tree consisting of three predictors of research excellence among 461 mathematicians

Predicting research excellence as combinations of publication track record characteristics

Each author follow a path through the tree and end up in a terminal node (Figure 1). The path through the tree reveal the combinations of publication track record characteristics (as defined

by the predictors and the predictor values) that is required for an author to end up in that particular terminal node.

At the root node, all authors (N=461) are evaluated for a potential split (node 1). Authors with ≤ 3 in *Top_Jour* follow the left branch and authors with > 3 in *Top_Jour* follow the right branch. The group of authors with > 3 in *Top_Jour* is further split on the predictor *Topic_Div*. Authors with a *Topic_Div* of ≤ 3 end up in terminal node 10, a group consisting of 8 authors of which 12.5% is excellent (Table 1). The group of authors with a topical diversity > 3 end up in terminal node 11, a group consisting of 32 authors of which 62.5% is excellent (Table 1). The appearance of *Topic_Div* (node 9) in the branch to the right of *Top_Jour* (node 1) but not to the left represents an interaction. Thus, *Topic_Div* has an effect on future research excellence at high levels of *Top_Jour*. *Top_Jour* and *Topic_Div* seem to be the best predictors of future research excellence.

Table 1. Showing combinations of predictors and predictor values required to end up in a particular terminal node.

TN*	Combinations of publication track record characteristics	% of excellent authors	*n* of authors	% of total authors
11	Top_Jour > 3; Topic_Div > 3	62.5%	32	6.9%
10	Top_Jour > 3; Topic_Div ≤ 3	12.5%	8	1.7%
6	Top_Jour ≤ 3; Top_Jour > 0; Topic_Div ≤ 3; English = Yes	26.8%	71	15.4%
7	Top_Jour ≤ 3; Top_Jour > 0; Topic_Div ≤ 3; English = No	6.8%	103	22.3%
8	Top_Jour ≤ 3; Top_Jour > 0; Topic_Div > 3	26.2%	61	13.2%
3	Top_Jour ≤ 3; Top_Jour ≤ 0	4.3%	186	40.3%

* Terminal Node

Authors following the left branch with ≤ 3 but > 0 in *Top_Jour* is further split on the predictor *Topic_Div*. Authors with > 3 in *Topic_Div* end up in terminal node 8, a group consisting of 61 authors of which 26.2% is excellent (Table 1). Authors with a *Topic_Div* value ≤ 3 is further split on the binary predictor *English*. An author with < 50% of the publication output at an institution in an English speaking country end up in the group represented by terminal node 7, of which 12.5% is excellent (Table 1). Authors with ≥ 50% end up in terminal node 8 of which 26.7% is excellent.

The predictor *English* is important among authors with > 0 but ≤ 3 in *Top_Jour* and ≤ 3 in *Topic_Div* (Table 1). This interaction indicate that publishing in an English speaking environment early in the career increase the probability of producing excellent research in P2 at low levels of *Top_Jour* and *Topic_Div*.

The group of authors with ≤ 3 papers in top journals in the first split and ≤ 0 papers in top journals in the second split, end up in terminal node 3, a group consisting of 186 authors of which 4.3% is excellent. At this level of *Top_Jour*, the predictors *Topic_Div* and *English* has no effect on the outcome.

DISCUSSION AND CONCLUSION

Decision trees can identify useful predictors and important relationships between variables without the need to specify a model a priori. In the case of 493 mathematicians in number theory seven predictors of research excellence was used as input, but only three, *Top_Jour*, *Topic_Div*, and *English*, was included in the tree. These results suggest that early career publication strategies where prestige journal publications and a topically diverse output is important for the production of future excellent research (as defined in this study). Previous research has shown that topical diversity has a positive effect on productivity (Dubois, Rochet, & Schlenker, 2014).

A particularly useful feature of decision trees is their ability to reveal interactions between predictors without specifying these in advance. An interesting interaction between the predictors, *Top_Jour*, *Topic_Div*, and *English*, was identified. At lower levels of *Top_Jour* and *Topic_Div* the predictor *English* had an effect on the response variable indicating a compensatory effect of early career publication activity in an English speaking environments. This interaction affected 37.7% of the included authors. Since the response variable used as a proxy for research excellence in this study is based on WoS data, the effect of *English* on the outcome may be a consequence of the well-known English language bias inherent in WoS (Moed, 2005).

Another useful feature of decision trees is the ability to show how different combinations of publication track record characteristics affect the outcome. Results show that authors with at least four publications in top journals and publications in at least four different subject areas have the highest probability to produce excellent research in P2. High topical diversity also has an effect on lower levels of *Top_Jour*. Authors with no publications in top journals are very unlikely to produce excellent research in P2. The information provided can be used to explore which combinations of publication characteristics that would be prioritized given some research policy selection criteria (e.g., a citation based indicator).

One limitation with decision trees is a tendency towards instability (i.e., small changes in the data can cause significant changes in the tree structure) in some situations (e.g., small sample size; overfitted trees) (King & Resick, 2014). Given this limitation decision trees are suitable as a complementary method to identify relationships between predictors that can be further tested with confirmatory methods.

REFERENCES

Abramo, G, Cicero, T, & D'Angelo, A. (2013). Are the authors of highly cited articles also the most productive ones? *Journal of Informetrics, 8*(1), 89-97.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal Of The American Society For Information Science And Technology, 62*(1), 50-60.

Dubois, P., Rochet, J., & Schlenker, J. (2014). Productivity and mobility in academic research: Evidence from mathematicians. *Scientometrics, 98*(3), 1669-1701.

Havemann, F., & Larsen, B. (2015). Bibliometric indicators of young authors in astrophysics: Can later stars be predicted? *Scientometrics, 102*(2), 1413-1434.

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

Jensen, P., Rouquier, J. B., & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, *78*(3), 467-479.

King, M., & Resick, P. (2014). Data Mining in Psychological Treatment Research: A Primer on Classification and Regression Trees. *Journal of Consulting and Clinical Psychology, 82*(5), 895-905.

Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics, 1*(2), 145-154.

Maimon, O., & Rokach, L. (2008). *Data Mining With Decision Trees: Theory and Applications*. London: World Scientific.

Moed, H. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., & Fortunato, S. (2013). On the Predictability of Future Impact in Science. *Scientific Reports*, *3*, 3052.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Strobl, C., Malley, J., Tutz, G., & Maxwell, S. E. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods, 14*(4), 323-348.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674.