Appendix 1

Matlab code for Cobb-Douglas function

```
function f = cobbfun(x,y)
[n1  n2] = size(x);
alpha = 0.3;
beta = 0.6
for a = 1 : n1
for b = 1 : n2
f(a,b) = (x(a,b)^alpha * y(a,b)^beta);
end
end
```

Matlab code for generating movie

```
clear;
close all;
data = load('SNIP_hugedata.txt');
x = data(:,1);
y = data(:,2).* 0.01;
N=80;
for i = 1:N
X = x((i-1)*100+i:i*100);
Y = y((i-1)*100+i:i*100);
[X1  Y1] = meshgrid(X,Y);
Z = cobbfun(X1,Y1);
disp(Z);
surf(X1,Y1,Z)
title('Cobb-Douglas Production function output');
ylabel('beta')
xlabel('alpha')
zlabel('Y')
if(mod(i,16) == 1)
gth = 1;
end

    M(i) = getframe(gcf);
    end
movie2avi(M,'WaveMoviePart2.avi');
```

Source code for computation of Self-citations and Journal Name Extraction :

https://github.com/MaQuest/computeSelfcites
https://github.com/sujithvm/internationality-journals
https://github.com/sujithvm/red-alert
The videos can be viewed at location:
https://drive.google.com/open?id=0B66EN2brcY_ScHYtVXdRRFZfQzg

Appendix 2

Journal Influence Score(JIS) The notion of internationality" proposed in the model embodying the work is based on the quantitative features of a journal. Journal Influence Score (JIS) serves as the most important tool for the formation of a cluster of internationality, derived from the scientometric data. A

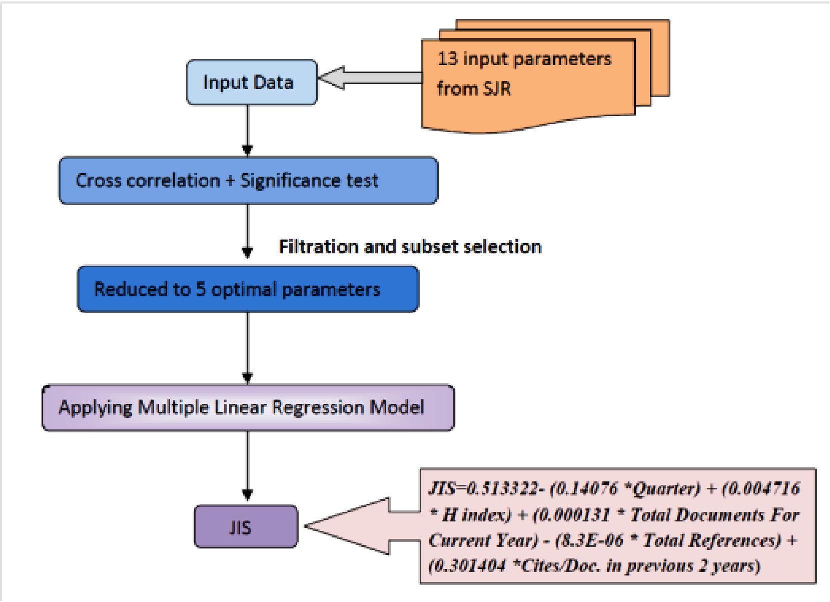
13 input parameters from SJR
Input Data
Cross correlation + Significance test
Filtration and subset selection
Reduced to 5 optimal parameters
Applying Multiple Linear Regression Model
JIS
JIS=0.513322- (0.14076 *Quarter) + (0.004716 * H index) + (0.000131 * Total Documents For Current Year) - (8.3E-06 * Total References) + (0.301404 *Cites/Doc. in previous 2 years)

Figure 1: Computation model of Journal Influence Score (JIS).

relatively new journal is then evaluated for internationality by measuring the proximity or inclusion to the known cluster, albeit loosely. The authors believe that the metric serves as a strong indicator of internationality. Such a score could help formulate a publication appraisal policy of institutions across the country. JIS could serve as a useful guideline for funding

As shown in Fig 1, we use a multiple linear regression (MLR) models where the JIS is the response variable. Thus, the response variable says y (JIS in our case), can be expressed as a function of k predictor variables $x_1, x_2, \ldots, x_k$ using a linear model of the form

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k + e$$

where $b_0, b_1, \ldots, b_k$ are fixed parameters that signify the weight of factors and e is the error.

Sample selection: For training and validating our model our source data for this study, we used data from the SCImago Journal and Country Rank (SJR) portal which contained journals in Elseviers Scopus. The portal includes the journals and country scientific indicators developed from the information contained in the Scopus database. The SCImago Journal and Country Rank [8] is a portal that includes the journals and country scientific indicators developed from the information contained in Scopus database. This data source provides the statistics for features mentioned below:

- SJR (SCImago Journal Rank) indicator: It expresses the average number of weighted citations received in the selected year by the documents published in the selected journal in the three previous years
- H Index: The h index expresses the journal's number of articles (h) that have received at least h citations.
- Total Docs./Total Documents: Output of the selected period. All types of documents are considered, including citable and noncitable documents.
- Total Docs. (3years): Published documents in the three previous years (selected year documents are excluded).
- Total References: It includes all the bibliographical references in a journal in the selected period.

- Total Cites (3years): Number of citations received in the selected year by a journal to the documents published in the three previous years.
- Citable Documents: Number of citable documents published by a journal in the three previous years (selected year documents are excluded). Exclusive articles, reviews, and conference papers are considered.
- Cites per Documents (2 years): Average citations per document in a 2 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the two previous years.
- Cites per Doc (3 years): Average citations per document in a 3 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the three previous years.
- Cites per Doc (4 years): Average citations per document in a 4 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the four previous years.
- Ref. / Doc.: Average number of references per document in the selected year.
- Self Cites: Number of journal's self-citations in the selected year to its own documents published in the three previous years.
- Non-citable documents (Available in the graphics) : Noncitable documents ratio in the period is considered.
- Cited Documents (Cited Doc.): Number of documents cited at least once in the three previous years.
- Uncited Documents (Uncited Doc.): Number of uncited documents in the three previous years.
- % International Collaboration: Document ratio whose affiliation includes more % than one country address.

Data acquisition: We used a set of 12 parameters available from the SCImago portal. Additionally we used the Quarter,Qi = i/4 where i was the quarter in which the journal was published. The input parameters (predictor variables) thus include the Quarter, H-Index, Total Docs 2012, Total Docs 3yrs, Total Cites 3yrs, Citable Docs 3yrs, Ref/Doc, Cites/Doc 3yrs and Total Ref. The quarter is considered as one of the input variables. Intuitively, any journal to be evaluated in the first Quarter of the year has more probability of having greater influence, considering the number of publications is mostly limited. Hence, the quarter of publication should be statistically significant. The results validate the use of quarter (in which the journal issue was published) in our model.

Statistical procedure: Starting with the initial set of input parameters, a two-phase approach was employed to obtain a more compact set of transformed variables. In the first step, the number of variables was reduced using correlation and MLR, and a down selected set of input variables was obtained. In the second step, pair wise correlation was applied on this reduced set and the few parameters that explained $>$ 90% of the variability were retained. The final model was an MLR model on the parameters retained after the second phase. These steps are described below.

Step 1 - Down selection using correlation with response variable and Multiple Linear Regression : In this phase, all the initially selected input parameters are used to analyze the correlation and regression statistics. The correlation of each individual parameter with the response variable was computed. Parameters which had both a low correlation ($<$0.4) as well as high p-value (> 0.05) were removed. As shown in table 1, the input variable Ref. / Doc can be removed. The regression was repeated multiple times until no parameters could be discarded based on above criteria.

Step 2 - Down selection based on pair wise correlation of the set of input variables obtained in Step : The down selected set of variables computed in Step 1 above for multiple journals was used to compute the overall variance from the co variance matrix. We computed pairwise correlations and identified a smaller set of variables such that the correlation between any two variables in this set was small. They can then be used to compute the percentage of variability accounted for individually as shown in table 1. This reduced the number further to only five input variables. The R2 value was very similar to when 9 input variables were considered. We did not do a Principal Component Analysis (PCA) since we were interested in down-selection of features. While in PCA the principal components are orthogonal to each other by design and it provides an elegant way of dimensionality reduction based on percentage variability explained, one problem is interpretation of the transformed variables with respect to the original input variables.