Imperial College
Business School

Imperial means
Intelligent Business

# Business Analytics Report

## Modelling of Bibliometric approach to the publications written by the professors at Imperial College Business School

MSc Business Analytics
2017 - 2018
Imperial College London

Name: Seongmin LEE
CID: 01247436

Word Count: 4,629

# Table of Contents

# Abstract

This report investigates the citation analysis of the 1,072 publications written by the professors at the business school of Imperial College London. In terms of calibrating articles' performance, the number of citation is the target variable for each journal. The primary purpose of the citation analysis is to measure the feature importance by using the statistical modelling and machine learning methods so as to build an effective way to stimulate the cited time for articles. Furthermore, by visualising a variety of the publications' factors, we can observe and grasp the journals' historical trend during the period between 2000 and 2018 throughout the explanatory analysis. The result shows that a journal impact factor, which is measured by the number of cited paper to the number of paper in each journal organization, has the most critical effect on the paper outcome, which suggests that the more influential journal association a publication is submitted in, the higher chances the business publications will get cited.

# 1. Introduction

Bibliometric analysis is based on measuring the performance of a publication. To be specific, a regression model on the bibliometric dataset indicates which predictors are statistically and economically significant to the citation increase. This method could be used for other industries such as the smartphone sales and marketing in that a product marketer can evaluate the influential factors to a sell-out quantity of a device. However, as a matter of the fact that publications do not have price tags, the bibliometric analysis cannot be implemented to the product marketing in the same way. Nevertheless, in terms of product, price, place and promotion, which is the universal concept of marketing strategy, both are analogous to each other on the condition of the same price range in the smartphone market. For instance, in the product, both are tangible, and cited time and sell-out quantity can be regarded as an outcome. When it comes to place, smartphones are distributed by selective channel distributors, which is similar to a journal organisation. Both distributors and journal publishers have an impact power to help to reach the higher target. Regarding promotion, both have an effect from word of mouth and search words such as keywords of publications and unique selling points of devices. Although ATL and BTL advertising has directly impacted the sales quantity in a consumer device market, there is a tendency that the purchase decision is made through the user experience at the offline stores in terms of customer purchase journey for smartphones. In other words, given the fact that both users have made up their minds after mostly having contact with the objects, both domains have an acquisition process in common.

It is impossible to analyse a dataset without a dataset, which means I had to scrape a dataset on the Internet or collect the related information manually. To obtain the information related to publications, I started to look into the official Imperial College website. And throughout the search engine, I could reach the college directory which is a web page for querying a list of the faculties by departments. By applying the filter to the business school department, the list of the professors' homepages in the department of the business school was secured. Fortunately, the professors' websites provide a list of publications, but with limited information only such as titles, co-authors, journal organisations, and related links. More importantly, some of the documents are linked with a website called the "Web of Science", where bibliometric figures for each journal is allowed to check. The information collected in the Web of Science is correctly suited for data analysis in view of the data integrity as well as the conformity. Not only with the suitable dataset, but the web service offers more extensive details such as abstracts, co-authors addresses, research fields, and references, compared to the professors' school account website. Also, I could reach calculated metrics such as impact factors by journal organisation or research field and h-index by authors, which are broadly used for bibliometrics. Therefore, I had no choice but to narrow the target publications down to the ones in the imperial professor accounts linked with the Web of Science although the chances to look through all of the dissertations have been lost.

The contributions of this study are divided into two parts: understanding of bibliometrics and recommendation for getting cited more for publications by Imperial College London. As mentioned in the beginning, grasping the bibliometric analysis will help to apply the method to a different domain such as a smartphone industry. As there is a big dataset provided by billions of users in the mobile device market, it is necessary to analyse the dataset from various points of views. For example, a smartphone maker wants a channel distributor with highest impact factors to handle newly released devices, instead of the merely largest distributor. In other words, a channel distributor with higher sell-out quantity per product is highly likely to cooperate with the maker. Secondly, the recommendation for professors at Imperial to get their publications cited more effectively will be provided, which is not only crucial for professors' reputations but also significant for the university rankings.

# 2. Literature Review

According to Sarli and Holmes (2011) as cited by Ale Ebrahim et al. (2013), the more keywords provided by authors are duplicated in the abstract, the higher the citation would get, which indicates that the similarity between the abstract and the keywords should be measured for the citations. Vanclay (2013) investigated how much Thomson Reuters impact factor(TRIF), h-index, and other author-related indicators have a causal effect on the citation. As a result of the research, the TRIF is the most reliable indicator of the increase of citation. Another research shows that the number of citation is positive to the authors' number (Aksnes, 2003). From the diversity point of view, there is an argument that the multi-country or multi-academic organisation outperform the homogenous authors' groups (Krause, 2009). According to an article published on the website of Nature, the number of the references in a publication is highly related to the cited time (Corbyn, 2010). Corbyn (2010) specifically demonstrated that a writer citing a peer expert output is expected to be quoted from the cited author. Menon and Phillips (2011) indicated that the odd-sized groups perform better than the even-sized groups because the odd-sized have stronger combination due to not being equally balanced, which mostly draws the better outputs.

# 3. Data Acquisition

Scraping from the Web of Science are divided into the five section: the citation, authors, research category, journal organisation, and references. Due to not only the limitation of computing power resource but the robust security of the Web of Science such as generating the random error messages, log in/out, IP block, and so on, the end of the analysis is more feasible if the target documents are specified. Thus, the target publications are satisfied with the conditions as follows:

1. The authors belong to the Imperial College business school on June 31st, 2018.
2. The publications should be shown on the official faculty account of Imperial College London, starting with http://www.imperial.ac.uk/people/+**faculty ID**
3. The target journals on the official faculty accounts are linked with the Web of Science

As a result, the 1,421 papers including NA values were obtained through several repeated trials due to the random website system error responses.

For the web crawling, the Python packages called the "Beautifulsoup" and the "Selenium" were used. The "Beautifulsoup" is famous for the web parsing in a format of HTML, and the "Selenium" is prevalent for the web-browser automation. Overall, the target web pages were firstly attained by the "Selenium", and then the "Beautifulsoup" library was implemented to acquire selected HTML strings.

## 3-1. Basic publications information [File Name: WOS_scraped_df.pkl]

Above all, it is essential to start by collecting the basic information of each publication such as a title, co-authors' names, cited times, a list of the references, research fields in order to extend the scraping target based on the basic information. As depicted in the appendix-1, these data can be approached by clicking the hyperlinks of the documents in the official Imperial account homepages of the professors which were directly connected to the corresponding publications' descriptions in the Web of Science. To avoid duplicate scraping, each publication URL linked with the Web of Science was also scraped as well to play a role as a primary key in the data table. Finally, the descriptions of the collected 15 vectors in the file are as follows:

1. WOS_URL: a publication URL linked with the Web of Science (Primary Key)
2. Title: publication name
3. Co-authors1: authors' IDs on the Web of Science
4. Co-authors2: authors' names
5. Journal Type: the journal organisation that a document was published
6. Research Area: 22 research fields defined by the Web of Science
7. WOS_category: 227 research categories defined by the Web of Science
8. Address: each author's organisation address

9. Organisation: the official name of the institution which each writer belongs to

10. Date: the published year and month

11. **Cited Time: the number of citation [Dependent Variable]**

12. Number of references: the number of references

13. Abstract: the whole sentences in the abstract

14. Keywords: the keywords provided by the authors

15. Publisher: the address of the journal organisation

The result from the first section demonstrates the basic information for each publication is enough for a predictive modelling. Thus, based on the obtained features in the first section, the final preprocessed data will be generated by being merged with other additional information from the section two to five.

## 3-2. Authors performance [File name: Author.csv]

It is mandatory for every author to create an ID and fill in a profile for creating an account on the Web of Science, which helps to measure the authors' performances such as total publications, total citation, and h-index associated with their publication histories. As shown in the appendix-2, the performance indexes can be approached by clicking on the authors' IDs in the web browser of the publication information. The main features of the writers' performance are as below:

1. ttl_publication: the number of total publication

2. h-index: the number of the cited journals divided by the total journal

3. sum_of_times_cited: total cited times for the whole publications

## 3-3. Research Categories impact factors [File Name: Category.csv]

Research categories help to measure how many research fields a journal covers. Although target documents have been produced by the business school, the number of the research fields that the total articles get involved in is 90 out of 227, which accounts for about 40% of the total categories. Furthermore, some of the areas are not even close to the business fields, such as 'Geography', 'History Of Social Sciences', 'Endocrinology & Metabolism', and 'Anesthesiology.' In considering that each field has different chances to get cited, it is better to acquire the impact factors for each research category. The Web of Science provides the analytical database call "Journal Citation Reports", which helps to search for each category's yearly impact factors from 1997 to 2016. As demonstrated in the appendix-3, the panel dataset for each research field was scraped. The main feature of this dataset is a median impact factor which indicates that how many times articles in the past two years get cited in the present one year. (INCITES INDICATORS HANDBOOK, 2014, pp7)

### 3-4. Journal impact factors [File Name: Journal Type.csv]

In view of politics, it is a common sense that an impact of an article published in the New York Times is most likely to be more influential than the one printed out by a local newspaper. Moreover, when it comes to the economy, the Wall Street Journal has dominant clout with people. In other words, every journal publisher has a different influence on each academy field. Furthermore, given the fact that each article should pass an examination process executed by experts in the specific areas, authors prefer to publish their dissertations in the more prominent journal organisation, which is not only representative to their career paths but also gives more chances to get the articles cited. As shown in the appendix-4, the journal impact factor in each publication page can be reached by clicking the "Journal Citation Report" button. By doing so, the panel dataset of the journal impact factor metric from 1997 to 2016 can be scraped.

### 3-5. Reference [File Name: Reference.csv]

In the web page displaying the basic publication information, the number of references can be only obtained. Given the fact that the cited times for each reference might be referred as a metric of publications' credibility, each reference information such as year of publishment and quoted time is also scraped in the references collection page through clicking the hyperlink labelled as "View All in Cited References page" as shown in the appendix-5.

## 4. Feature Engineering and Explanatory Analysis

In the previous section, the information relevant to the publications' cited times has been collected from the Web of Science. The total number of the dissertations is 1,421, which are enough to be representative to measure the publications performance written by the professors at the business school of Imperial College London. Based on the preprocessing of those datasets, features to predict the cited items will be generated in this section. The file name of the base dataset is the WOS_scraped_df.pkl, which includes the necessary information on each publication, such as title, authors, cited time, and published year. Moreover, the other datasets such as authors, references, journals, and research categories will be merged into the basic publication information table.

### 4-1. Year since published

As depicted below, the longer the year of a publication is passed, the higher chances a paper gets to be cited on average before the year of 2,000. To control this factor, the dissertations written after a year of 2,000 has been selected due to the small sample size for the journal published before a year of 2,000. Moreover, the "yrs_since_published" variable is calculated by subtracting 2019 to published years.

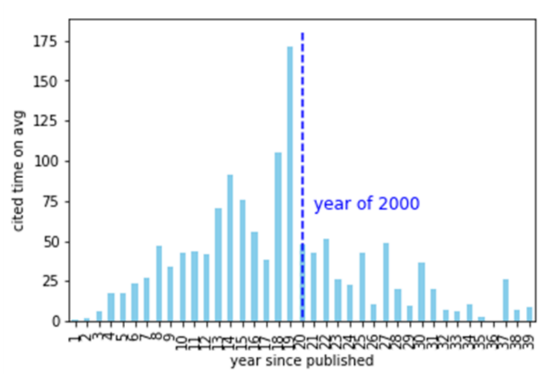


Figure1. Cited times on average by year

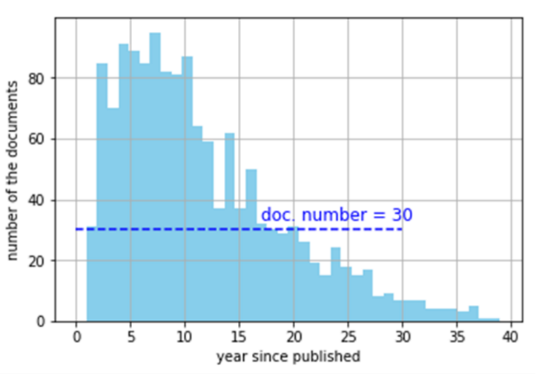


Figure2. Number of documents by year

## 4-2. Organization Country (Preprocessing Only)

To measure diversity, the country lists were extracted from the address column which includes the names of universities, address in details, and countries. The diversity features on organisations and nations will be generated in the latter section.

## 4-3. Author information

### 4-3-1. The number of the co-authors for each publication

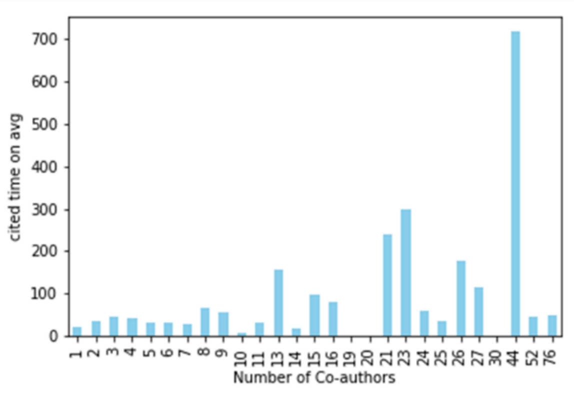


Figure3. Cited times by co-author numbers

As mentioned in the literature review, the citation is proportional to the number of authors. Authors-related indicators are primary factors to measure the publications performance in that a publication written by authors with a dominant reputation will be more likely to be read or cited by other authors. Above all, the number of the co-authors could be the simplest, but the most useful variable for the predictive modelling. As shown on the bar chart, there is a tendency that the larger the number of co-authors is, the higher cited time the publications mostly acquire.

### 4-3-2. Odd number of the authors (1 = odd-size group)

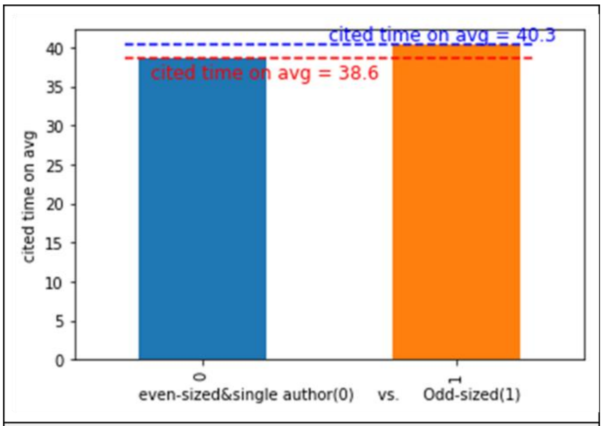


Figure4. Even vs. Odd-sized Groups

As pointing out in the article review, there is a result indicating that odd-sized groups outperform even-sized groups. Therefore, the dummy variable labelled as 1 is created if either the number of the group member is odd, otherwise 0. As depicted, the odd-sized group mostly gets the higher number of citation than the even sized group and a group of single authors.

### 4-3-3. The average h index of the authors

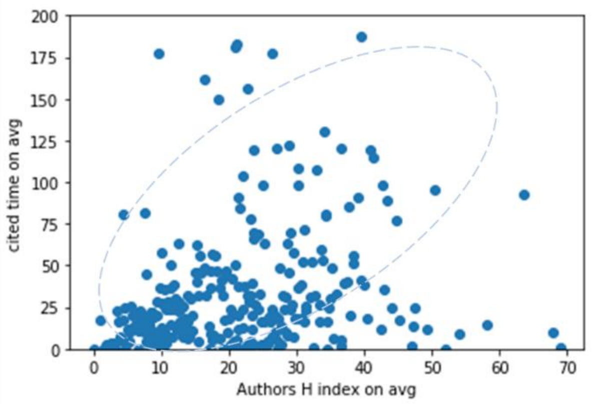


Figure5. Cite times by H-index

The author.csv file displays each author's performance in terms of the publication history. Among the achievement indicators, h-index is typical of measurement to authors' productivity. Therefore, the h-index on average in the group of the authors for a publication is calculated as a metric of the publications performance. As shown in the figure 5, the documents with the higher h index on average occupy on the upper right, which indicates the h index on average could be a proper metric to evaluate publications quality.

## 4-4. Reference information

When it comes to the references information, Corbyn demonstrated that a writer citing a peer expert output is expected to be quoted from the cited author (2010). Not only that, but quoting is a logical way for an article to strengthen argument on controvertible issues. So, the longer an article has a list of the references, the more likely the article is to be irrefutable. The dataset already includes the number of the references in each publication. In the Reference.csv dataset, we have reference title and year of being published and cited times for each paper in the base dataset. Calculated the cited time on average for the whole references in each publication, a new feature is created to describe the article credibility, which means that a journal supported by a higher cited references are more likely to be irrefutable. As depicted on the charts, the number of the references and the average cited times of a group of references is positively correlated to the citation.

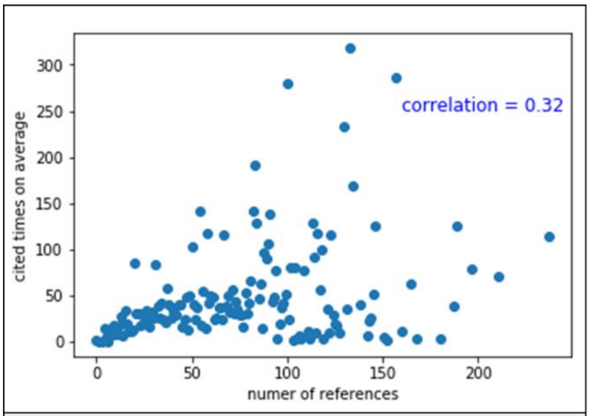


Figure6. Cited times by the ref. number

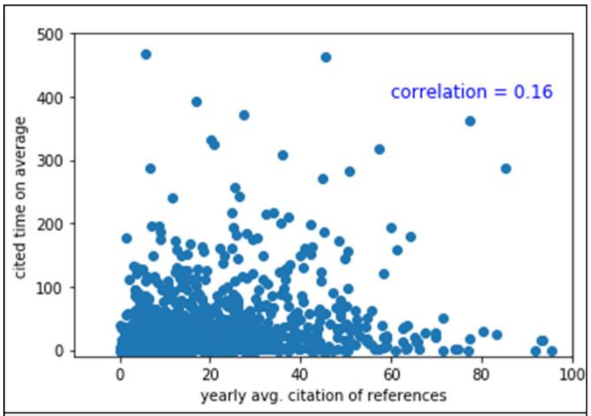


Figure7. Cited times by the ref. citations

## 4-5. Organization

Most of the co-authors belonged to certain official institutions when they published the dissertations. The quality of the publications has an effect on the university rankings. Reversely, the institution's level might help to form a group of competent authors or more likely to secure the funding. The dataset in the organizationbyResearchArea folder contains category normalised impact factors and ranking by each research field and universities as shown below. According to the indicator handbook published by the Web of Science, For the same journal type and year of publishment, and the research category, the category normalised citation impact is measured from a formal that the cited time is divided by the expected cited time (InCites INDICATORS HANDBOOK, 2014, p.11.). Therefore, the organisation ranking and category normalised citation impact on average in each research field have been aggregated for each publication. As shown in the below charts, both variables seem relevant to the cited time.

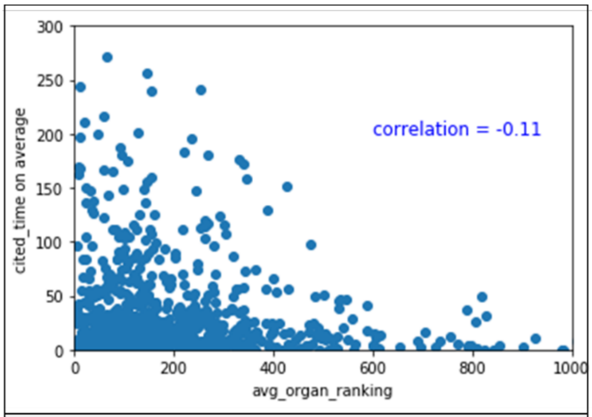


Figure8. Cited times by Org. Ranking

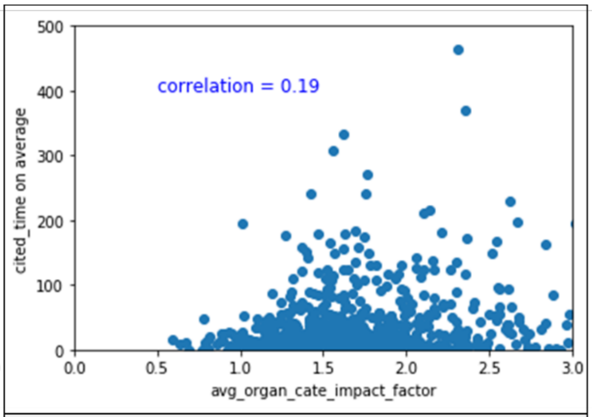


Figure9. Cited times by the CNCI

## 4-6. Research Field impact factor

Although one of the co-authors for each journal is the business school professor at Imperial College London, the target publications cover the 90 research categories which indicate that there must be interdisciplinary convergence research between business and other fields. In other words, there might be significantly different impact factors on which research fields a publication covers. Also, the dataset about the research categories is in the panel format, which helps to measure the impact factor during a period since the journal was published. Based on matching the research field and published year, the data in the column of the "aggregate impact factor", which is the cited time in two years divided by the number of the items in two years, has been calculated on average for a group of the research category belonging to each publication.

## 4-7. Journal impact factor

An article published by a renowned journal organisation will get cited more. In other, the influence or clout of a dissertation might be the vital indicator to predict the cited time of a paper. In the same way as the research field impact factor variable above, the journal impacts factor on average from the published year to the latest year has been created. According to the InCites Indicator Handbook published in 2014 by the Web of Science, journal impact factor is the number of the cited times divided by the items released over the past two years in each journal (InCites INDICATORS HANDBOOK, 2014, p.7).

## 4-8. Diversity

### 4-8-1. Ethnicity

When it comes to diversity, there are three features to be generated in the dataset: authors' ethnicities, organisations' countries, and educational institutions. First of all, based upon the last names of the target authors, the ethnicities can be estimated by using the python package called "ethnicolr." The basic idea of ethnicity is that authors with different growth backgrounds might look at a topic and solve it differently. In other words, the ethnicity diversity might help to create an environment of thesis-antithesis-synthesis paradigm. So, the dummy variable is created for 1 if multi-ethnicities, otherwise 0.

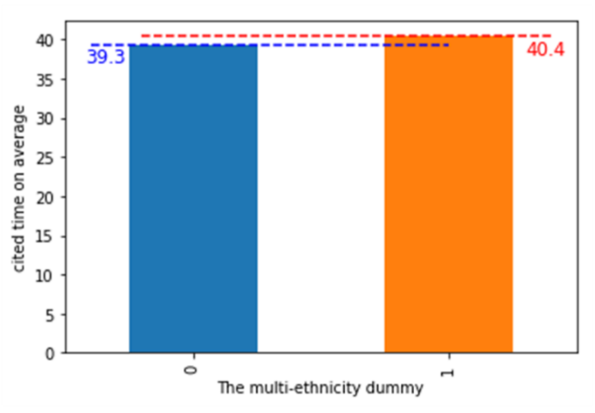


Figure10. Cited Times by ethnicity

#### 4-8-2. Organization Country

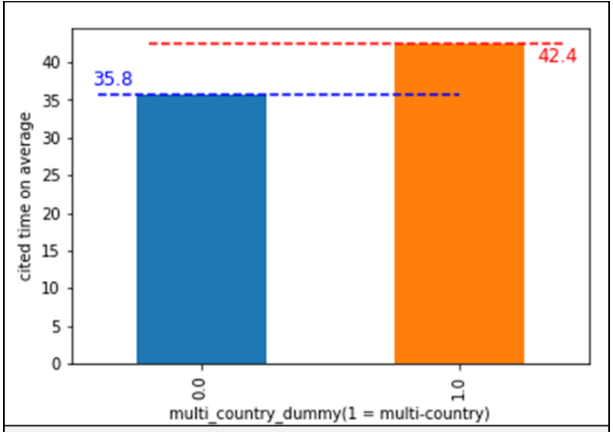


Figure11. Cited Times by multi-countries

The number of the countries that the academic institutions of the authors are located mostly ranges from one or two, which indicates that it is arduous for professors to work on research on an international scale. However, it is important to have higher diversity rate in that the diversity of organisations is related to verifying a generalised theory. For example, it might be harder for a group of authors at Imperial College London to generalise a method to be able to be applied to not only Europe but also other continents due to exceptional regional cases. However, a publication written by a group of authors based in different countries would be proven more broadly, which might lead to a more generalised dissertation. So, the dummy variable is created for 1 if multi-countries of the institutions, otherwise 0.

#### 4-8-3. Educational Institution

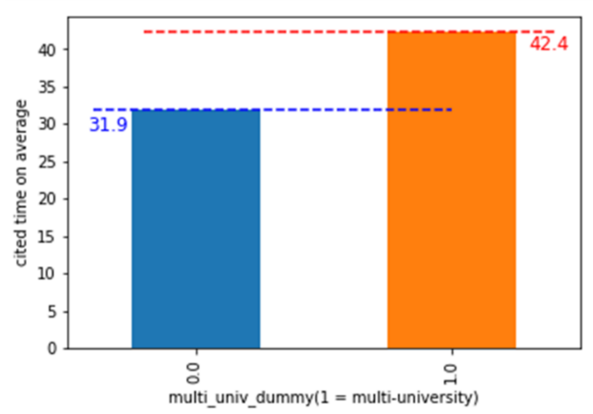


Figure12. Cited Times by Multi-universities

Regarding institution diversity, one school has a stronger academic field than another. Thus, a publication with a university diversity might have higher impact enhanced by fusion research. So, the dummy variable is created for 1 if multi-institutions, otherwise 0. The relationship between the cited times and multi-university dummy variables shows that the outputs made by multi-institutions are better than the one by a single institution.

### 4-9. Title+Keywords similar to Abstract

From the information search point of view, it is crucial to set the right title and keywords for a publication because the publication is more likely to be shown on the top search result. Furthermore, if the title and keywords are highly similar to the abstract, readers would continue to look into and finally quote the publication. Thus, the formula for measuring a similarity is defined as the intersect of two lists divided by the union of the two lists. So, the similarity measurement indicates how many times the words in a title and the keywords are found in the abstract for a document.

Throughout the preprocessing and feature engineering, 15 independent variables have been created for the 1,072 data lines after removing the null values in the abstract vector. The descriptions of each predictor are as following:

**Features Description**

| No. | Feature Name | Description |
|---|---|---|
| 1 | yrs_since_published | passed year since published.<br>ex) published year: 2018→yrs_since_published: 1 |
| 2 | co_author_num | the number of the co-authors |
| 3 | avg_authors_h_index | authors' h-index on average |
| 4 | sim | the similarity between title+keywords and abstract |
| 5 | num_of_ref | the number of the references for a target publication |
| 6 | ref_yearly_cited_avg | the cited time of the references on average for a target publication |
| 7 | avg_organ_ranking | a mean of the ranking of the organisations in the relevant research categories |
| 8 | avg_organ_cate_impact_factor | a mean of the category impact factor in the relevant research categories |
| 9 | avg_research_field_impact_factor | a mean of impact factor in the relevant research categories |
| 10 | avg_journal_impact_factor | a mean of impact factor in the relevant journal |
| 11 | multi_eth_dummy | 1 if the number of authors' ethnicities is above 1. Otherwise, 0. |
| 12 | organ_multi_country_dummy | 1 if the number of organisations' countries is above 1. Otherwise, 0. |
| 13 | multi_univ_dummy | 1 if the number of organisations is over 1. Otherwise, 0. |
| 14 | author_odd_num_dummy | 1 if the number of co-authors is odd and greater than 1. Otherwise, 0. |
| 15 | keyword_provided_dummy | 1 if keywords are provided. Otherwise, 0. |

# 5. Predictive Modelling and Interpretation

This section is divided into four parts: 1. Multicollinearity, 2. Feature Selection, 3. Model Evaluation, and 4. Model interpretation. As shown on the histogram, the figure seems like the exponential distribution, which describes that a majority of the sample is skewed to the left side. There are multiple modelling options such as negative binomial regression(NB regression), Poisson regression, or linear regression with taking the log of the target variable especially for the dependent variable like the number of event occurrence.

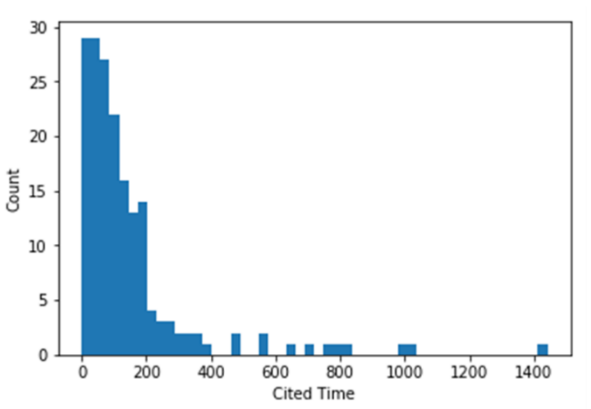
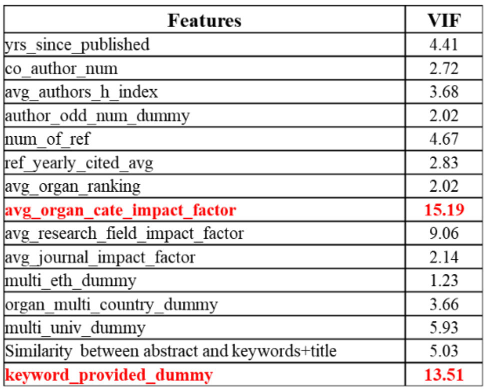


**Figure13. Histogram of Cited Times**

## 5-1. Multicollinearity

Multicollinearity can cause misinterpreting the effect of the independent variables in the models. The best way to solve the issue is removing the variables above the variance inflation factor(VIF) of 10 and highly correlated variables. By using the function "variance_inflation_factor" in the "statesmodels" package, the VIF was calculated as shown on the Figure 14. As a result, there are two variables above the VIF of 10, which are "avg_organ_cate_impact_factor", and "keyword_provided_dummy". Also, as depicted in the heatmap correlation chart, "organ_multi_country_dummy" is highly correlated to the "multi_univ_dummy," which indicates that one of them should be removed.



| Features | VIF |
|---|---|
| yrs_since_published | 4.41 |
| co_author_num | 2.72 |
| avg_authors_h_index | 3.68 |
| author_odd_num_dummy | 2.02 |
| num_of_ref | 4.67 |
| ref_yearly_cited_avg | 2.83 |
| avg_organ_ranking | 2.02 |
| avg_organ_cate_impact_factor | 15.19 |
| avg_research_field_impact_factor | 9.06 |
| avg_journal_impact_factor | 2.14 |
| multi_eth_dummy | 1.23 |
| organ_multi_country_dummy | 3.66 |
| multi_univ_dummy | 5.93 |
| Similarity between abstract and keywords+title | 5.03 |
| keyword_provided_dummy | 13.51 |

Figure14. Variance Inflation Factor

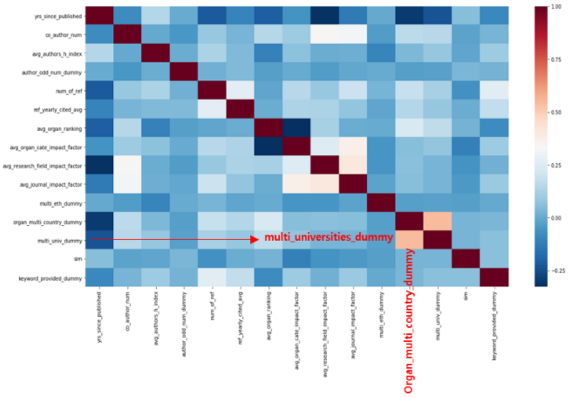


Figure15. Heatmap Correlation

## 5-2. Feature Selection

The Random Forest(RF) will be used to select the predictors by measuring the feature importance. To find out the best hyper-parameters of the RF models, the grid search method has been applied with the cross-validation of k = 20. As a result, the optimal model has built on the max feature = 7 and the number of estimators = 550. The results of Random Forest Regression with log(cited_time) as target variable shows that "avg_journal_impact_factor", "yrs_since_published", "num of ref", and "avg_authors_h_index" are the top 4 important variables as below. Based on the previous VIF, correlation heatmap, and the Random Forest feature importance, the three feature groups have been formulated as below.

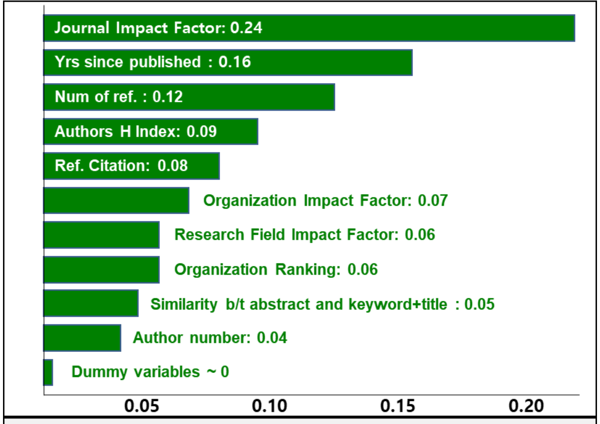


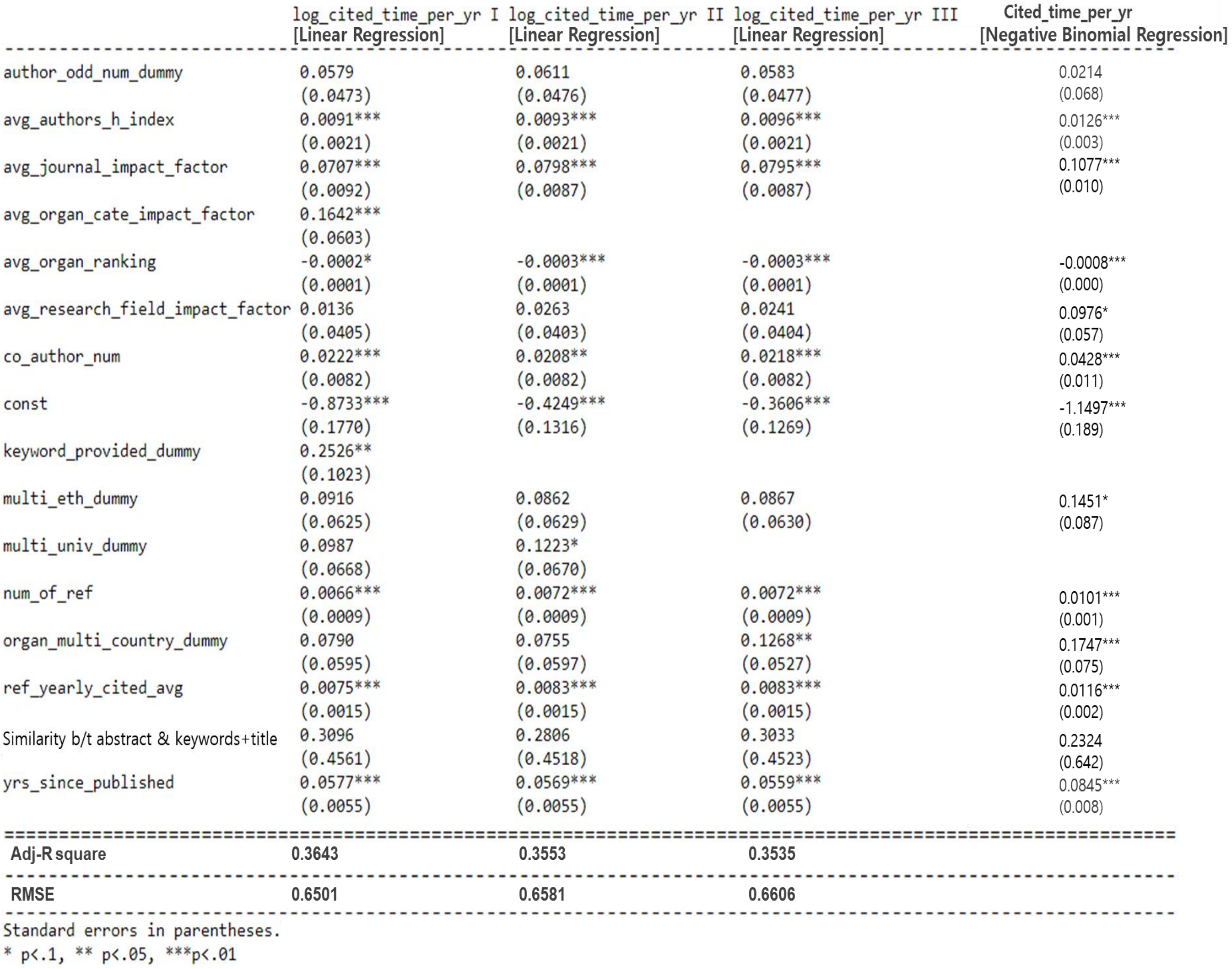
Figure16. Feature Importance by RF

1. all variables group: 15 variables
2. truncated group 1: 13 variables (dropped: "avg_organ_IF", "keyword _dummy")
3. truncated group 2: 12 variables (dropped: " avg_organ_IF", "keyword_dummy", " univ_dummy")

## 5-3. Model Evaluation (Linear Regression, Negative Binomial Regression)

To compare the three models' performance, the dataset is split into 80% of the trainset and 20% of the testset. For the regression model trained with all of the variables, the adjusted R square is 0.36, and the Root Mean Square Error(RMSE) is 0.65, which are the best among the three models. The results of the other two regressors have similar to each other, which is the adjusted R square of 0.35 and the RMSE of 0.66. Regarding interpretation, it is essential to select an unbiased model which satisfies the Gauss-Markov theorem. Therefore, the last model with the 12 features has been chosen to interpret the casual effects.

**Fig17. Regression Models Comparison**



|  | log_cited_time_per_yr I<br>[Linear Regression] | log_cited_time_per_yr II<br>[Linear Regression] | log_cited_time_per_yr III<br>[Linear Regression] | Cited_time_per_yr<br>[Negative Binomial Regression] |
|---|---|---|---|---|
| author_odd_num_dummy | 0.0579 | 0.0611 | 0.0583 | 0.0214 |
|  | (0.0473) | (0.0476) | (0.0477) | (0.068) |
| avg_authors_h_index | 0.0091*** | 0.0093*** | 0.0096*** | 0.0126*** |
|  | (0.0021) | (0.0021) | (0.0021) | (0.003) |
| avg_journal_impact_factor | 0.0707*** | 0.0798*** | 0.0795*** | 0.1077*** |
|  | (0.0092) | (0.0087) | (0.0087) | (0.010) |
| avg_organ_cate_impact_factor | 0.1642*** |  |  |  |
|  | (0.0603) |  |  |  |
| avg_organ_ranking | -0.0002* | -0.0003*** | -0.0003*** | -0.0008*** |
|  | (0.0001) | (0.0001) | (0.0001) | (0.000) |
| avg_research_field_impact_factor | 0.0136 | 0.0263 | 0.0241 | 0.0976* |
|  | (0.0405) | (0.0403) | (0.0404) | (0.057) |
| co_author_num | 0.0222*** | 0.0208** | 0.0218*** | 0.0428*** |
|  | (0.0082) | (0.0082) | (0.0082) | (0.011) |
| const | -0.8733*** | -0.4249*** | -0.3606*** | -1.1497*** |
|  | (0.1770) | (0.1316) | (0.1269) | (0.189) |
| keyword_provided_dummy | 0.2526** |  |  |  |
|  | (0.1023) |  |  |  |
| multi_eth_dummy | 0.0916 | 0.0862 | 0.0867 | 0.1451* |
|  | (0.0625) | (0.0629) | (0.0630) | (0.087) |
| multi_univ_dummy | 0.0987 | 0.1223* |  |  |
|  | (0.0668) | (0.0670) |  |  |
| num_of_ref | 0.0066*** | 0.0072*** | 0.0072*** | 0.0101*** |
|  | (0.0009) | (0.0009) | (0.0009) | (0.001) |
| organ_multi_country_dummy | 0.0790 | 0.0755 | 0.1268** | 0.1747*** |
|  | (0.0595) | (0.0597) | (0.0527) | (0.075) |
| ref_yearly_cited_avg | 0.0075*** | 0.0083*** | 0.0083*** | 0.0116*** |
|  | (0.0015) | (0.0015) | (0.0015) | (0.002) |
| Similarity b/t abstract & keywords+title | 0.3096 | 0.2806 | 0.3033 | 0.2324 |
|  | (0.4561) | (0.4518) | (0.4523) | (0.642) |
| yrs_since_published | 0.0577*** | 0.0569*** | 0.0559*** | 0.0845*** |
|  | (0.0055) | (0.0055) | (0.0055) | (0.008) |
| **Adj-R square** | **0.3643** | **0.3553** | **0.3535** |  |
| **RMSE** | **0.6501** | **0.6581** | **0.6606** |  |

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Due to the difference of the dependent variables between the linear regression and negative binomial regression, the two models' performance cannot be compared. However, given the dependent variable is a count of event occurrence, the negative binomial regression is a more proper model to infer in this case.

### 5-4. Model Interpretation

Above all, the journal impact factor is statistically and economically significant at the 95% confident interval, which indicates that the higher a journal has an impact factor, the more chances a publication gets to be cited. Moreover, the period of being published has a positive effect on the cited time, which demonstrates that every ten years passed after the publishment will give a chance for the 8% increase of the citation. When it comes to the author's performance, the number of co-authors is more critical than the average h-index of the co-authors in that both are statistically significant at the 95% confident interval. For the reference, citing publications with high citation is more critical than citing more publications, which refers to the significance of the quality-oriented citing. The other variables are insignificant or not economical in the regression model.

## 6. Discussion

The dataset only consists of the publication's information at the whole business school of Imperial College London. It would be much reliable if the specific department names as a dummy variable are included because Finance, Innovation and Entrepreneurship, and Management could have different impact factors regarding journal or research field. Secondly, the reference can be analysed deeper in a way to measure the similarity between each publication title and reference titles, which is a hypothesis that the less similar a title to the reference titles, the more unique the publication could be. However, that could not be executed because some of the reference titles are not provided in the Web of Science. On top of that, to generalise the output of the regression models, the dissertations published by other business schools also need modelling and comparing. By doing so, the unique factors affecting the cited times of the Imperial publications can be found.

## 7. Conclusion

In this report, we saw how the journal factors affect the citation. The regression result demonstrates that authors at the Imperial business school should consider which journal organisation the dissertations will be published above all with respect to increasing cited times. Furthermore, the more authors from diversified academic institutions cooperate, the higher citation the journal will get, which indicates that the external diversity is more significantly crucial than the internal diversity such as ethnicity and gender. Although the universities rankings where the authors belong are statistically significant, it does not affect the cited time economically, which might point out that the professors at Imperial are already in a position to receive attention in the world top 10 schools. Moreover, how many citations the references get are more important than the number of the referencing for a publication, which demonstrates that the quality for the references is more significant than the quantity of the referenced papers.

# 8. Reference

Aksnes, D. (2003). Characteristics of highly cited papers. Research Evaluation, 12(3), pp.159-170.

Ale Ebrahim, N., Salehi, H., Amin Embi, M., Habibi Tanha, F., Gholizadeh, H., Motahar, S. and Ordi, A. (2013). Effective Strategies for Increasing Citation Frequency. International Education Studies, 6(11).

Corbyn, Z. (2010). An easy way to boost a paper's citations. [online] Nature. Available at: https://www.nature.com/news/2010/100813/full/news.2010.406.html[Accessed 21 Jun. 2018].

InCites INDICATORS HANDBOOK. (2014). [ebook] Thomson Reuters, p.7&11. Available at: http://ipscience-help.thomsonreuters.com/inCites2Live/8980-TRS/version/default/part/AttachmentData/data/InCites-Indicators-Handbook-6%2019.pdf [Accessed 20 Jun. 2018].

Krause K. (2009). Increasing your Article's Citation Rates. [online] Bepress. Available at: https://works.bepress.com/kate_krause/12/ [Accessed 23 Jul. 2018].

Menon, T. and Phillips, K. (2011). Getting Even or Being at Odds? Cohesion in Even- and Odd-Sized Small Groups. Organization Science, 22(3), pp.738-753.

Sarli, C., & Holmes, K. (2011). Strategies for Enhancing the Impact of Research, Retrieved May 9, 2013, from https://beeker.wustl.edu/impact-assessment/strategies

Vanclay, J. (2013). Factors affecting citation rates in environmental science. Journal of Informetrics, 7(2), pp.265-271.

# 9. Appendix

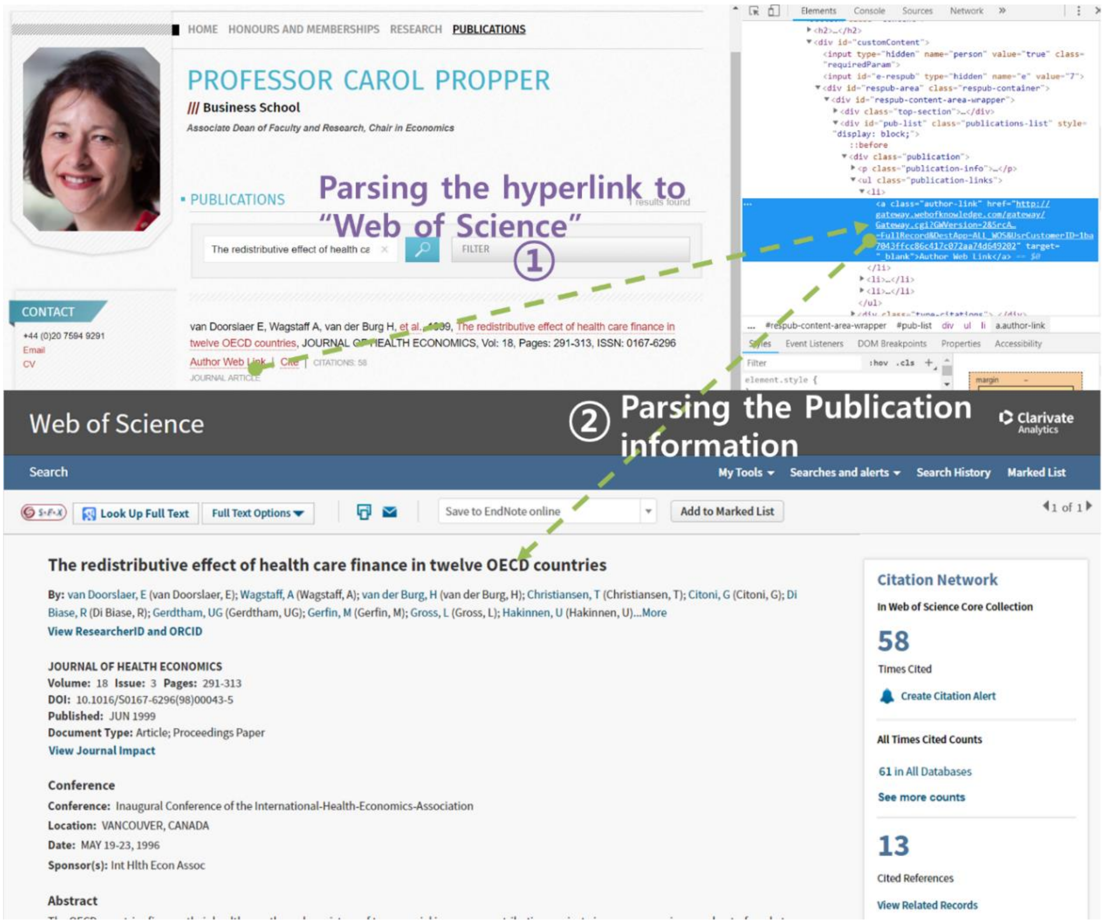
## Bibliography

Dhawan, S. and Gupta, B. (2005). Evaluation of Indian Physics Research on Journal Impact Factor and Citations Count: A Comparative Study. DESIDOC Bulletin of Information Technology, 25(3), pp.3-8.

Evans, T., Hopkins, N. and Kaube, B. (2012). Universality of performance indicators based on citation and reference counts. Scientometrics, 93(2), pp.473-495.

Yang, K. and Lee, J. (2013). Bibliometric Approach to Research Assessment: Publication Count, Citation Count, & Author Rank. Journal of Information Science Theory and Practice, 1(1), pp.27-41.

## Appendix-1 Imperial Account Scraping Process



① Parsing the hyperlink to "Web of Science"

van Doorslaer E, Wagstaff A, van der Burg H, et al., 1999, The redistributive effect of health care finance in twelve OECD countries, JOURNAL OF HEALTH ECONOMICS, Vol: 18, Pages: 291-313, ISSN: 0167-6296

Author Web Link | Cite | CITATIONS: 58

JOURNAL ARTICLE

```
▸<h2>…</h2>
▾<div id="customContent">
   <input type="hidden" name="person" value="true" class=
   "requiredParam">
   <input id="e-respub" type="hidden" name="e" value="7">
  ▾<div id="respub-area" class="respub-container">
    ▾<div id="respub-content-area-wrapper">
      ▸<div class="top-section">…</div>
      ▾<div id="pub-list" class="publications-list" style=
       "display: block;">
          ::before
        ▾<div class="publication">
          ▸<p class="publication-info">…</p>
          ▾<ul class="publication-links">
            ▾<li>
                <a class="author-link" href="http://
                gateway.webofknowledge.com/gateway/
                Gateway.cgi?GWVersion=2&SrcA…
                =FullRecord&DestApp=ALL_WOS&UsrCustomerID=1ba
                7043ffcc86c417c072aa74d649202" target=
                "_blank">Author Web Link</a> == $0
              </li>
            ▸<li>…</li>
            ▸<li>…</li>
            </ul>
```

… #respub-content-area-wrapper #pub-list div ul li a.author-link

element.style {

② Parsing the Publication information

**The redistributive effect of health care finance in twelve OECD countries**

By: van Doorslaer, E (van Doorslaer, E); Wagstaff, A (Wagstaff, A); van der Burg, H (van der Burg, H); Christiansen, T (Christiansen, T); Citoni, G (Citoni, G); Di Biase, R (Di Biase, R); Gerdtham, UG (Gerdtham, UG); Gerfin, M (Gerfin, M); Gross, L (Gross, L); Hakinnen, U (Hakinnen, U)...More

View ResearcherID and ORCID

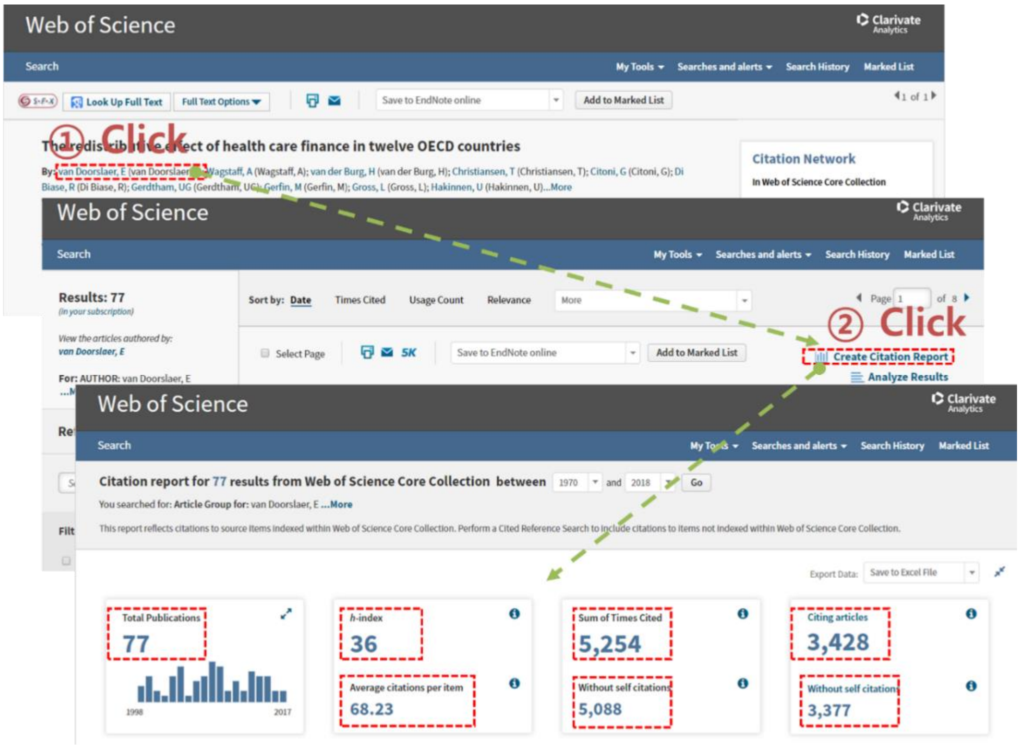
JOURNAL OF HEALTH ECONOMICS
Volume: 18 Issue: 3 Pages: 291-313
DOI: 10.1016/S0167-6296(98)00043-5
Published: JUN 1999
Document Type: Article; Proceedings Paper
View Journal Impact

Conference
Conference: Inaugural Conference of the International-Health-Economics-Association
Location: VANCOUVER, CANADA
Date: MAY 19-23, 1996
Sponsor(s): Int Hlth Econ Assoc

Abstract

Citation Network
In Web of Science Core Collection
58
Times Cited
Create Citation Alert
All Times Cited Counts
61 in All Databases
See more counts
13
Cited References
View Related Records

# Appendix-2 Author Performance Scraping Process



The redistributive effect of health care finance in twelve OECD countries

By: van Doorslaer, E (van Doorslaer, E); Wagstaff, A (Wagstaff, A); van der Burg, H (van der Burg, H); Christiansen, T (Christiansen, T); Citoni, G (Citoni, G); Di Biase, R (Di Biase, R); Gerdtham, UG (Gerdtham, UG); Gerfin, M (Gerfin, M); Gross, L (Gross, L); Hakinnen, U (Hakinnen, U)...More

Results: 77
*(in your subscription)*

*View the articles authored by:*
*van Doorslaer, E*

For: AUTHOR: van Doorslaer, E

Citation report for 77 results from Web of Science Core Collection between 1970 and 2018

You searched for: Article Group for: van Doorslaer, E ...More

This report reflects citations to source items indexed within Web of Science Core Collection. Perform a Cited Reference Search to include citations to items not indexed within Web of Science Core Collection.

Total Publications: 77

*h*-index: 36

Average citations per item: 68.23

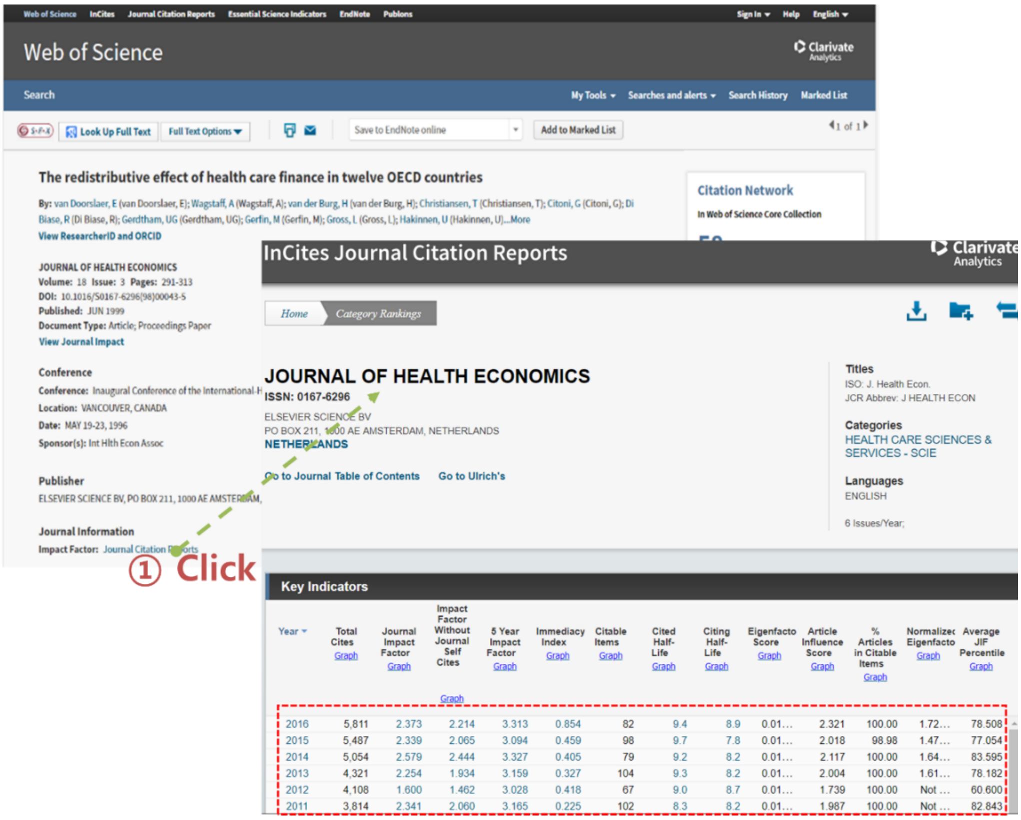
Sum of Times Cited: 5,254

Without self citations: 5,088

Citing articles: 3,428

Without self citations: 3,377

# Appendix-3 Research Category Scraping Process



**The redistributive effect of health care finance in twelve OECD countries**

By: van Doorslaer, E (van Doorslaer, E); Wagstaff, A (Wagstaff, A); van der Burg, H (van der Burg, H); Christiansen, T (Christiansen, T); Citoni, G (Citoni, G); Di Biase, R (Di Biase, R); Gerdtham, UG (Gerdtham, UG); Gerfin, M (Gerfin, M); Gross, L (Gross, L); Hakinnen, U (Hakinnen, U)...More

View ResearcherID and ORCID

JOURNAL OF HEALTH ECONOMICS
Volume: 18 Issue: 3 Pages: 291-313
DOI: 10.1016/S0167-6296(98)00043-5
Published: JUN 1999
Document Type: Article; Proceedings Paper
View Journal Impact

Conference
Conference: Inaugural Conference of the International-H
Location: VANCOUVER, CANADA
Date: MAY 19-23, 1996
Sponsor(s): Int Hlth Econ Assoc

Publisher
ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM,

Journal Information
Impact Factor: Journal Citation Reports

JOURNAL OF HEALTH ECONOMICS
ISSN: 0167-6296
ELSEVIER SCIENCE BV
PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS
NETHERLANDS

Go to Journal Table of Contents    Go to Ulrich's

Titles
ISO: J. Health Econ.
JCR Abbrev: J HEALTH ECON

Categories
HEALTH CARE SCIENCES & SERVICES - SCIE

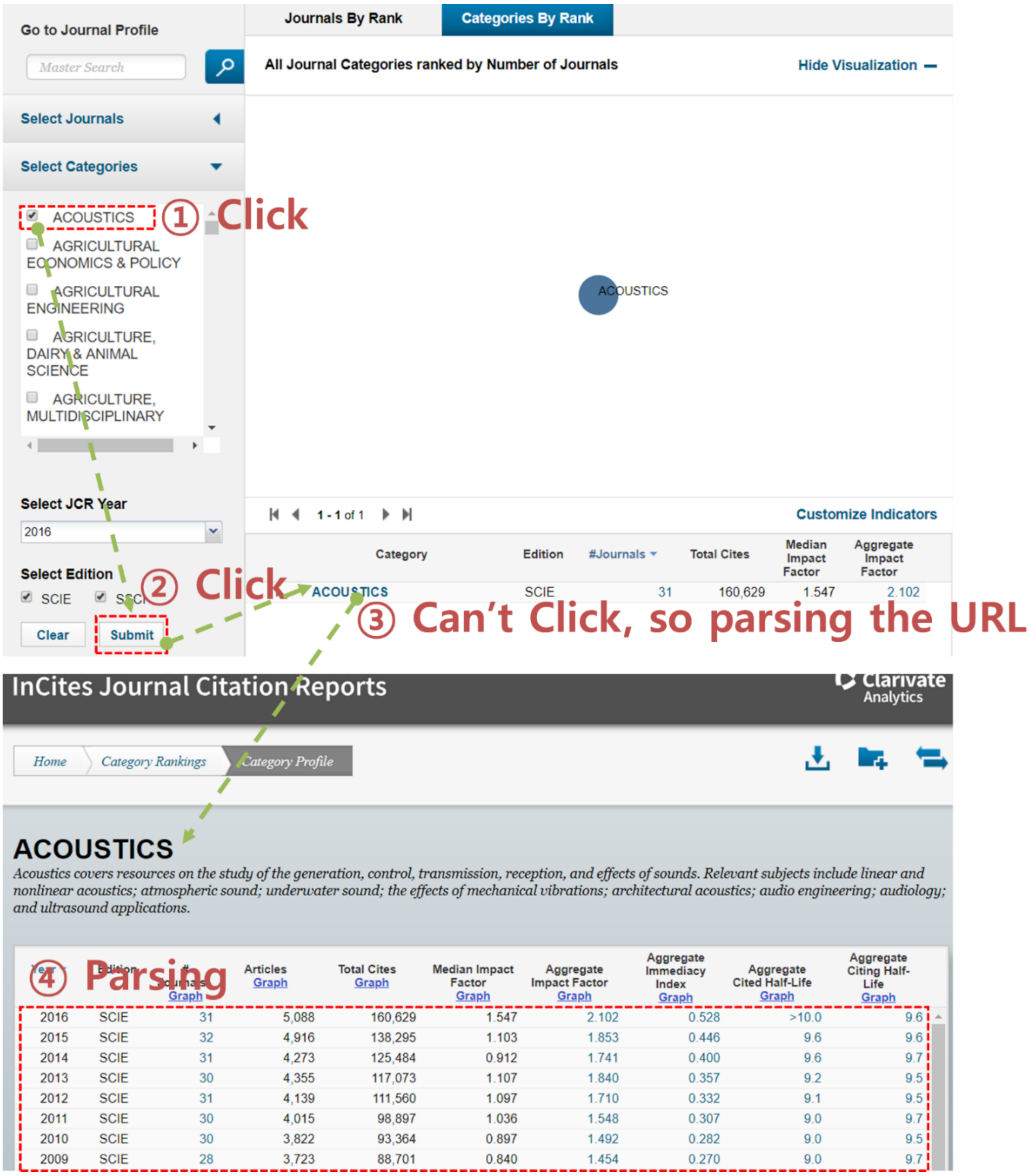
Languages
ENGLISH

6 Issues/Year;

Key Indicators

| Year | Total Cites Graph | Journal Impact Factor Graph | Impact Factor Without Journal Self Cites Graph | 5 Year Impact Factor Graph | Immediacy Index Graph | Citable Items Graph | Cited Half-Life Graph | Citing Half-Life Graph | Eigenfacto Score Graph | Article Influence Score Graph | % Articles in Citable Items Graph | Normalizec Eigenfacto Graph | Average JIF Percentile Graph |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2016 | 5,811 | 2.373 | 2.214 | 3.313 | 0.854 | 82 | 9.4 | 8.9 | 0.01… | 2.321 | 100.00 | 1.72… | 78.508 |
| 2015 | 5,487 | 2.339 | 2.065 | 3.094 | 0.459 | 98 | 9.7 | 7.8 | 0.01… | 2.018 | 98.98 | 1.47… | 77.054 |
| 2014 | 5,054 | 2.579 | 2.444 | 3.327 | 0.405 | 79 | 9.2 | 8.2 | 0.01… | 2.117 | 100.00 | 1.64… | 83.595 |
| 2013 | 4,321 | 2.254 | 1.934 | 3.159 | 0.327 | 104 | 9.3 | 8.2 | 0.01… | 2.004 | 100.00 | 1.61… | 78.182 |
| 2012 | 4,108 | 1.600 | 1.462 | 3.028 | 0.418 | 67 | 9.0 | 8.7 | 0.01… | 1.739 | 100.00 | Not … | 60.600 |
| 2011 | 3,814 | 2.341 | 2.060 | 3.165 | 0.225 | 102 | 8.3 | 8.2 | 0.01… | 1.987 | 100.00 | Not … | 82.843 |

# Appendix-4 Journal Impact Factor Scraping Process



① Click

② Click

③ Can’t Click, so parsing the URL

④ Parsing

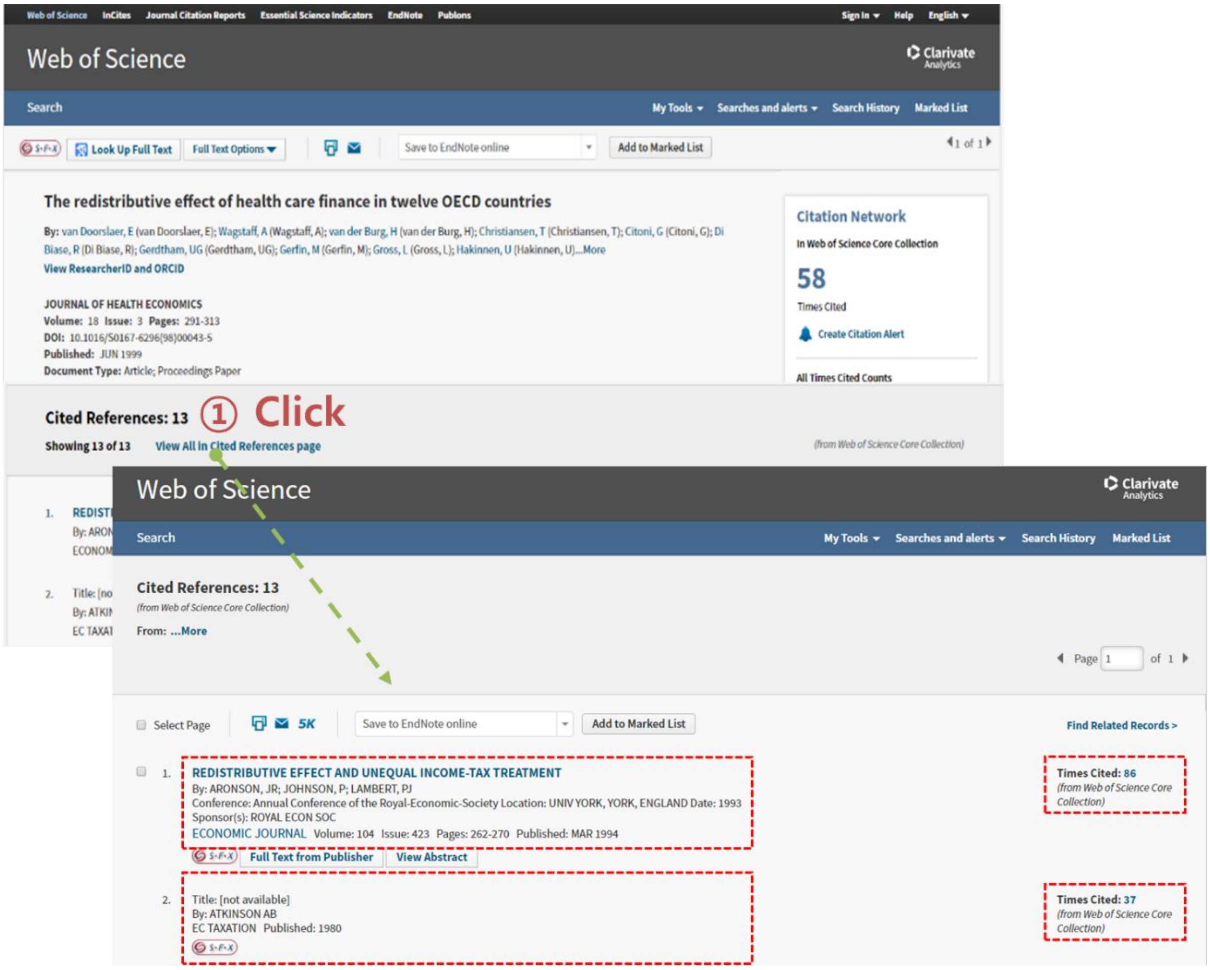
| Category | Edition | #Journals | Total Cites | Median Impact Factor | Aggregate Impact Factor |
|---|---|---|---|---|---|
| ACOUSTICS | SCIE | 31 | 160,629 | 1.547 | 2.102 |

## ACOUSTICS

*Acoustics covers resources on the study of the generation, control, transmission, reception, and effects of sounds. Relevant subjects include linear and nonlinear acoustics; atmospheric sound; underwater sound; the effects of mechanical vibrations; architectural acoustics; audio engineering; audiology; and ultrasound applications.*

| Year | Edition | # Journals Graph | Articles Graph | Total Cites Graph | Median Impact Factor Graph | Aggregate Impact Factor Graph | Aggregate Immediacy Index Graph | Aggregate Cited Half-Life Graph | Aggregate Citing Half-Life Graph |
|---|---|---|---|---|---|---|---|---|---|
| 2016 | SCIE | 31 | 5,088 | 160,629 | 1.547 | 2.102 | 0.528 | >10.0 | 9.6 |
| 2015 | SCIE | 32 | 4,916 | 138,295 | 1.103 | 1.853 | 0.446 | 9.6 | 9.6 |
| 2014 | SCIE | 31 | 4,273 | 125,484 | 0.912 | 1.741 | 0.400 | 9.6 | 9.7 |
| 2013 | SCIE | 30 | 4,355 | 117,073 | 1.107 | 1.840 | 0.357 | 9.2 | 9.5 |
| 2012 | SCIE | 31 | 4,139 | 111,560 | 1.097 | 1.710 | 0.332 | 9.1 | 9.5 |
| 2011 | SCIE | 30 | 4,015 | 98,897 | 1.036 | 1.548 | 0.307 | 9.0 | 9.7 |
| 2010 | SCIE | 30 | 3,822 | 93,364 | 0.897 | 1.492 | 0.282 | 9.0 | 9.5 |
| 2009 | SCIE | 28 | 3,723 | 88,701 | 0.840 | 1.454 | 0.270 | 9.0 | 9.7 |

# Appendix-5 References Scraping Process



**The redistributive effect of health care finance in twelve OECD countries**

By: van Doorslaer, E (van Doorslaer, E); Wagstaff, A (Wagstaff, A); van der Burg, H (van der Burg, H); Christiansen, T (Christiansen, T); Citoni, G (Citoni, G); Di Biase, R (Di Biase, R); Gerdtham, UG (Gerdtham, UG); Gerfin, M (Gerfin, M); Gross, L (Gross, L); Hakinnen, U (Hakinnen, U)...More

View ResearcherID and ORCID

JOURNAL OF HEALTH ECONOMICS
Volume: 18 Issue: 3 Pages: 291-313
DOI: 10.1016/S0167-6296(98)00043-5
Published: JUN 1999
Document Type: Article; Proceedings Paper

Citation Network
In Web of Science Core Collection
58
Times Cited
Create Citation Alert
All Times Cited Counts

Cited References: 13 ① Click
Showing 13 of 13 View All in Cited References page
(from Web of Science Core Collection)

Cited References: 13
(from Web of Science Core Collection)
From: ...More

□ 1. REDISTRIBUTIVE EFFECT AND UNEQUAL INCOME-TAX TREATMENT
By: ARONSON, JR; JOHNSON, P; LAMBERT, PJ
Conference: Annual Conference of the Royal-Economic-Society Location: UNIV YORK, YORK, ENGLAND Date: 1993
Sponsor(s): ROYAL ECON SOC
ECONOMIC JOURNAL Volume: 104 Issue: 423 Pages: 262-270 Published: MAR 1994

Times Cited: 86
(from Web of Science Core Collection)

2. Title: [not available]
By: ATKINSON AB
EC TAXATION Published: 1980

Times Cited: 37
(from Web of Science Core Collection)