998 project 2: What matters publication number and citation number?

Yifan Mei

March 22, 2018

Summary

This paper is discussing how can we predict "output", as measured by publications and citations, based on specific conditioning factors including:Project size as proxied by total expenditures, Relative portion of funding allocated to faculty, staff, and students, Funding mechanism (competitive, formula, other), Basic vs. applied science proportions, Time since completion, Science area (keywords) based on the data awarded by USDA to researchers at the University of Wisconsin–Madison during the period 2008-2016.

The main challenge for the project are: How to organize data from 4 different excel tables to put them together; How to deal with the missing data;How to deal with the data values we get from the data base and transform them in. Two multiple linear regression models are applied to either publication number or maximium citation number of the projects' publications (we choose max citation number of publication instead of sum of citation numbers because of the high similarity of publications within the same project. It is not reasonable to count the similar papers citation for different times.) over the factors we interested plus two impportant factors that affect publication number and citation number: publication journals' impact factors and the cost of time for each project. The result is as following:

For the publication numbers, we have strong evidence to say that the increase of 1,000,000 us dollar total expenditure is related to the increase of 2 publications, and we have weak evidence to say the more total days spent on the project also leads to increase of publication with 10000 days leader to increase of 1.8 publications. Besides, we have weak evidence to say the interaction of funding mechanisum: NIFA NON FORMULA and other study fields contributes to the increase of publication number. Once these two interacted, it leades to 4.61 publication increase.

For the maximal citation number, we have strong evidence to say that the funding mechanisum: NIFA NON FORMULA and the time since the project ends contributes to the increase of maximium citation number of the project's publiations. For a Nifa non formula mechanism, the max citation is 2.5. For the increase of 100 day the project ends, the max citation increase 1. Meanwhile, we have moderate evidence on the interaction of mechanism NIFA Non furmula and the study field: agriculture and medical and multisubjects that they leads to the decrease of maximium citation number on 25.32, 33.97, 32.06, which means that these field are not attract readers a lot. Besides, we weak evidence to say that 100 more total spend date, 1.131 more the maximium citation.

Introduction

Background

U.S. federal science funding agencies collectively spent between $130 and $180 billion per year during the past decade to support university and private foundation research. Agencies allocate much of this funding to researchers through competitive-grant programs that invite individual researchers and research teams to propose research projects with specific aims, hypotheses, and research methods.

For this specific project, I seeks associate patterns of research output, as measured by publications and citations, with characteristics of researchers, research areas, and funding mechanisms by examining data awarded by USDA to researchers at the University of Wisconsin–Madison during the period 2008-2016. These data are derived exclusively from administrative sources and is part of a larger project involving nearly 50 major research universities and covering all major U.S. funding agencies. Additional background on this (UMETRICS) project can be found at the website for the Institute for Research on Innovation and Science (IRIS) at the University of Michigan, and at the related Innovation Measurement Initiative website of the U.S. Census Bureau.

We derive our data from four sources as outlined in the following tables. The USDA maintains a publicly accessible database of all their funded research projects (past and current), and the University of Wisconsin—Madison assembles data on a quarterly basis that it submits to IRIS for the UMETRICS project. Additionally, we searched the references databases Scopus and the Web of Science (available to campus students, faculty, and staff) for each publication that references a project in the set of USDA funded project at the UW. Information about each publication are contained in the relevant tables.

Main Questions

The primary question is, How can we predict “output”, as measured by publications and citations, based on specific conditioning factors including: Project size as proxied by total expenditures Relative portion of funding allocated to faculty, staff, and students Funding mechanism (competitive, formula, other) Basic vs. applied science proportions Time since completion Science area (keywords)

Main Challenges

The main challenge for the project are: How to organize data from 4 different excel tables to put them together How to deal with the missing data How to deal with the data values we get from the data base and transform them in

Data Process

Firstly, by using same doi and assession number for each project or publication, I merge the 4 excel files provided by the client together.

Then, I did a time transformation: I merged the 2 types of dates: projects' start dates and projects' end dates to projects' costs of dates. Considering the time effect to numbers of citition, I also calculated the differences between the projects' end dates and our data collected date as a factor named "time for citing". By calculating the difference between end date and data collected date, we also do not need the variable: "project status", because it is the same as "time for citing".

Now, I get a first version of full data. however, there are a lot of missing data here. One type of missing data needed to be think at first: missing values for title for publications. If the titles for publication are missing, it is impossible for anyone to figure out what exactly they are and whether they really existed. In this way, the lines where the title of publication is missing are deleted. Besides, considering the same meaning of 2 factors: number of authors and authors, I merged the authors for publications with number of authors together into number of authors and delete authors of publications.

Besides, I delete all lines that the expenditures are negative because they are not reasonable.

Then, after observing the data, I find out that for the publishments under the same project, the contents are quite similar, thus counting each published journal or book's citation number for the same porject is not reasonable. Instead, I use the publishment with max citation for each project, and count the number of publishment for each project. Besides, I used the 2017 journal ranks listed on http://www.scimagojr.com/journalrank.php?year=2017 as the impact factors for the max citation paper in each project.

Now, it is time to classify the subjects. I use a very powerful package named text2vec in R. It automatically splits all subjects into single words and build a vocabulary database based onthese single words except useless words: "AND", "&", "SCIENCE", "SCIENCES", "TECHNOLOGY", "TECHNOLOGIES", "ALL", "(", ")", ",", ";", "APPLIED", "STUDIES", "SMALL. Then, I got the top 30 words that appears most times (See Appendix Figure1). Then, I merge the top 50 words into 6 types: biology, agriculture, medical, animal, ecology, multisubjects, other.

Now, before I can do analysis about the data, I deal with the missing values in the dataset. One type of missing data should be dealed with at first: the missing data in factors "pct_basic_research" (basic research rate), "pct_applied_research"(applied research rate), and "pct_develop_research"(developed research), because the sum of these three vectors are 100%. Thus, I use random number on the range of 0% to 100% to generate the rate of basic research for the missing values, and then I use the random number on the range of 0% to the difference of 100% and the repaired missing values in basic research rate to generate the applied research rate. Then, the developed research rate is 100% - repaired missing values in basic research rate - repaired missing values in applied research rate. Then, a method named Multiple Imputation is used for all the rest of missing value. The strength for this method is

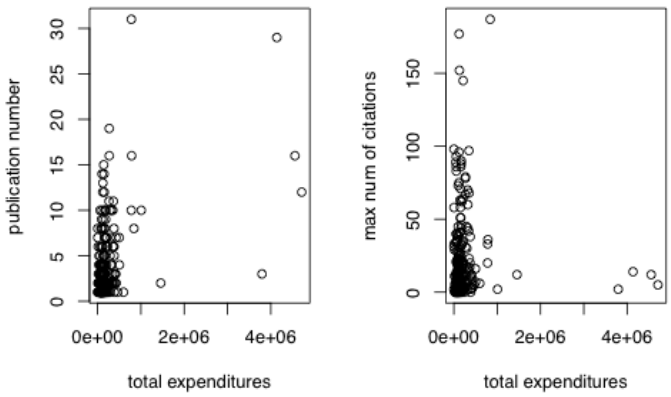

Figure 1: The scatterplots of total expenditures over publication number and citation number. Most expenditures are less than 1,000,000. The expenditure over 2,000,000 does not contribute to high publication directly to both publication number and citation number

that, comparing to traditional simple methods of dealing with missing data: omitting them, use mode or median or mean of the total factor as the value of missing data, it imputes the missing entries of the incomplete data sets, not once, but m (number of each vector) times. In this way, imputed values are drawn for a distribution. In this way, the missing values can be "repaired" differently. for each missing entry)

Basic Analysis about the factors we are interested in

After a long way, we finally get a clear version of data without any missing data, and we can take a brief look about the simple relationship of factors we interested toward the publication number and the maximium citation number. The analytic plots are listed as below:

Final Model

After I see the simple relationships between the outputs, publication numbers and citation numbers to the 6 interested variables, I think it is reasonable to use multiple linear regression model to detect the relationships among them in general because of some linear relationships for the simple models. Besides, considering the potential impact on the reputation of projects' published journals and the dates the project spend, I think it is reasonable to add these two terms together into the regression model. Besides, according to the result of interaction plot, there does exist interaction effect between the study effect and the funding mechanism, thus, the interaction term is also added to the regression.

The results are listed as following:

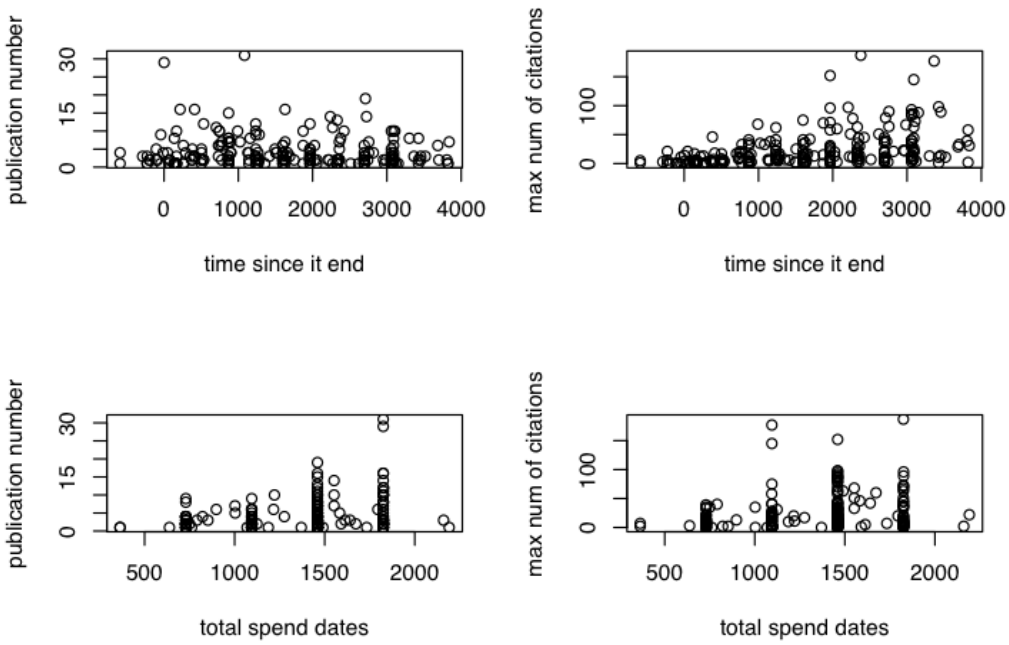

Figure 2: The scatterplos of time between end data of project and the data collected date toward the publication number and citations. The time between end data of project and the data collected date does not contribute obviously to the publication number, but a little bit contribution to the increase of max citation number. The total time a project spend does has upper trend to the publication number and the citation number

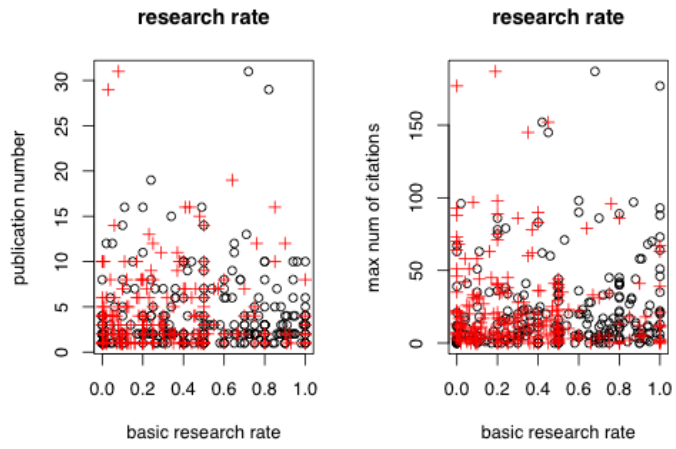

Figure 3: The scatterplots of research rate (basic, applied) towards publication number and maximium citation. There is no obvious pattern for eith basic or applied research rate toward either publication number or maximium citation.

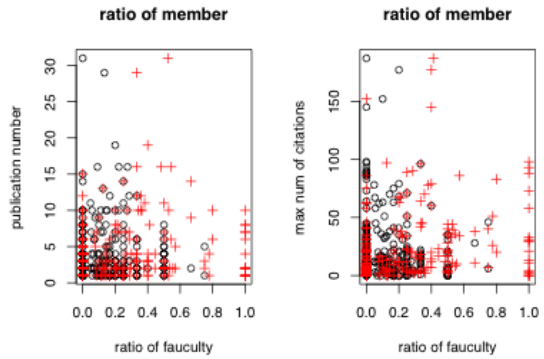

Figure 4: The scatterplots of the the ratio of faculty and the the ratio of staff over publication number and max citation number

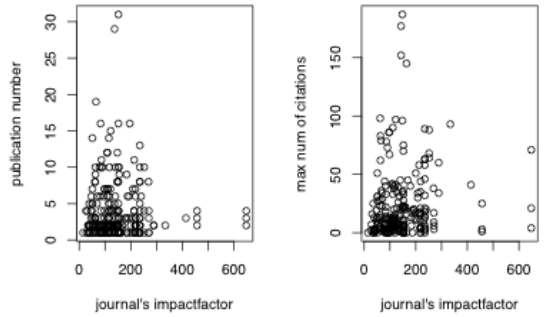

Figure 5: journal's impact factor over publication number and citation number. The impact factor gathers below 300. Around 200 impact number has largest citation.

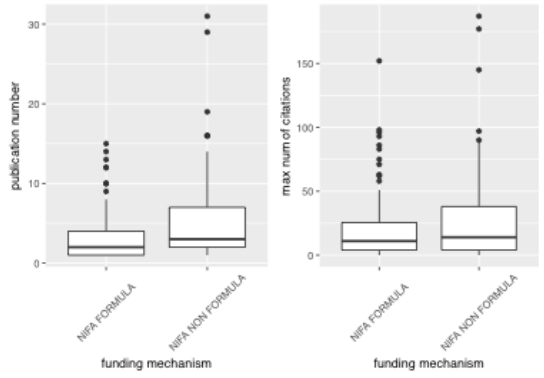

Figure 6: Boxplots for funding mechanisum over publication number and maximium citation number. According to the boxplot for funding mechanisum over publication number and maximium citation number, we can see nearly the same median but subtle difference of shape between two mechanisums over both publication number and maximium citation number.

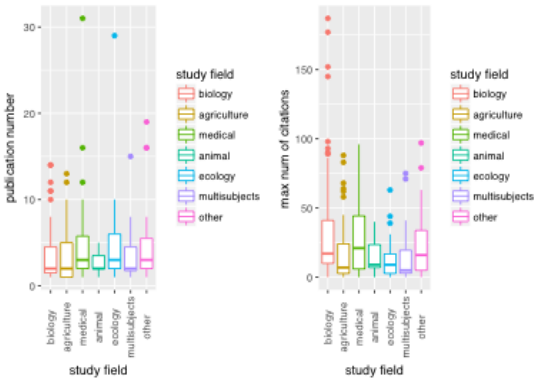

Figure 7: Boxplots for study fields over publication number and maximium citation number. Comparing the boxplot for study field over publication number and maximium citation number, we can see obvious differences of study fields over the maximium citation number, and slightly diferrences of study fields over publication number

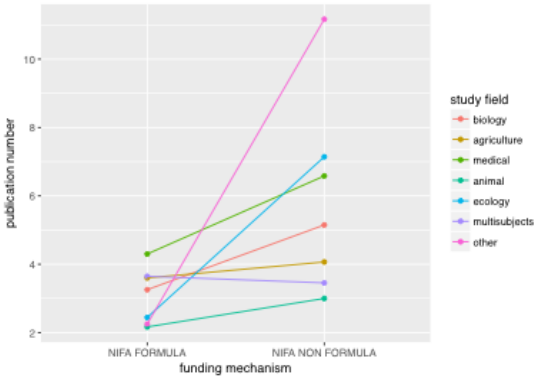

Figure 8: Interaction plot of funding mechanisum and study fields. Based on the interaction plot, we can see interaction effect for nearly all study fields to the funding mechanisums except the animal field

Table 1: Multiple linear regression model based for publication number over journal's impactfactor, study field, total expenditures, time since it end, total spend dates, funding mechanism, ratio of basic research, ratio of applied research, ratio of fauculty, ratio of staff. Based on the model, we have strong evidence to say that the increase of 1,000,000 us dollar total expenditure is related to the increase of 2 publications, and we have weak evidence to say the more total days spent on the project also leads to increase of publication with 10000 days leader to increase of 1.8 publications. Besides, we have weak evidence to say the interaction of funding mechanisum: NIFA NON FORMULA and other study fields contributes to the increase of publication number. Once these two interacted, it leades to 4.61 publication increase.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2946567	1.7497325	0.7399169	0.4601729
‘journal's impactfactor‘	-0.0022057	0.0028891	-0.7634705	0.4460357
‘study field‘agriculture	0.3625068	0.9052137	0.4004654	0.6892189
‘study field‘medical	0.8866562	1.3284017	0.6674609	0.5052071
‘study field‘animal	-1.4848515	1.6630077	-0.8928711	0.3729434
‘study field‘ecology	-0.4337644	1.1055385	-0.3923558	0.6951914
‘study field‘multisubjects	0.4689360	1.0885280	0.4307983	0.6670548
‘study field‘other	-0.5743522	1.2374814	-0.4641299	0.6430331
‘total expenditures‘	0.0000021	0.0000005	4.2007227	0.0000393
‘time since it end‘	-0.0001728	0.0002800	-0.6173032	0.5377003
‘total spend dates‘	0.0018957	0.0007675	2.4698969	0.0143086
‘funding mechanism‘NIFA NON FORMULA	0.9975722	1.0730313	0.9296767	0.3536009
‘ratio of basic research‘	-0.0086233	0.0132176	-0.6524061	0.5148494
‘ratio of applied research‘	-0.0050536	0.0148080	-0.3412761	0.7332354
‘ratio of fauculty‘	1.5898524	1.8807180	0.8453433	0.3988767
‘ratio of staff‘	1.7314464	1.1073952	1.5635306	0.1194264
‘study field‘agriculture:‘funding mechanism‘NIFA NON FORMULA	-1.8512105	1.4195274	-1.3041034	0.1936192
‘study field‘medical:‘funding mechanism‘NIFA NON FORMULA	0.2388932	1.9328727	0.1235949	0.9017537
‘study field‘animal:‘funding mechanism‘NIFA NON FORMULA	-0.2615466	2.6765337	-0.0977184	0.9222487
‘study field‘ecology:‘funding mechanism‘NIFA NON FORMULA	1.3657174	1.7502461	0.7803002	0.4360886
‘study field‘multisubjects:‘funding mechanism‘NIFA NON FORMULA	-2.0285683	1.7955811	-1.1297559	0.2598618
‘study field‘other:‘funding mechanism‘NIFA NON FORMULA	4.3869845	2.1857915	2.0070461	0.0460207

Table 2: Multiple linear regression model based for max num of citations over journal's impactfactor, study field, total expenditures, time since it end, total spend dates, funding mechanism, ratio of basic research, ratio of applied research, ratio of fauculty, ratio of staff. According to the result, we have strong evidence to say that the funding mechanisum: NIFA NON FORMULA and the time since the project ends contributes to the increase of maximium citation number of the project's publiations. For a Nifa non formula mechanism, the max citation is 2.5. For the increase of 100 day the project ends, the max citation increase 1. Meanwhile, we have moderate evidence on the interaction of mechanism NIFA Non furmula and the study field: agriculture and medical and multisubjects that they leads to the decrease of maximium citation number on 25.32, 33.97, 32.06, which means that these field are not attract readers a lot. Besides, we weak evidence to say that 100 more total spend date, 1.131 more the maximium citation.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-20.6808122	11.7288585	-1.7632417	0.0793072
‘journal's impactfactor‘	0.0234622	0.0193661	1.2115123	0.2270542
‘study field‘agriculture	-7.9248174	6.0678552	-1.3060327	0.1929634
‘study field‘medical	7.2818699	8.9045815	0.8177667	0.4144129
‘study field‘animal	-13.7332052	11.1475219	-1.2319514	0.2193384
‘study field‘ecology	-13.5986430	7.4106779	-1.8350066	0.0679125
‘study field‘multisubjects	-2.7423323	7.2966525	-0.3758343	0.7074177
‘study field‘other	-4.4526599	8.2951216	-0.5367805	0.5919847
‘total expenditures‘	-0.0000024	0.0000034	-0.6974321	0.4863002
‘time since it end‘	0.0116962	0.0018768	6.2318749	0.0000000
‘total spend dates‘	0.0126246	0.0051449	2.4537967	0.0149463
‘funding mechanism‘NIFA NON FORMULA	23.1277419	7.1927748	3.2154130	0.0015072
‘ratio of basic research‘	0.1016494	0.0886008	1.1472742	0.2525675
‘ratio of applied research‘	-0.0176763	0.0992617	-0.1780773	0.8588331
‘ratio of fauculty‘	14.6294050	12.6068839	1.1604299	0.2471851
‘ratio of staff‘	0.1737517	7.4231242	0.0234068	0.9813479
‘study field‘agriculture:‘funding mechanism‘NIFA NON FORMULA	-23.2767884	9.5154176	-2.4462183	0.0152551
‘study field‘medical:‘funding mechanism‘NIFA NON FORMULA	-31.8727917	12.9564888	-2.4599868	0.0146982
‘study field‘animal:‘funding mechanism‘NIFA NON FORMULA	-22.8642297	17.9414193	-1.2743824	0.2039296
‘study field‘ecology:‘funding mechanism‘NIFA NON FORMULA	-10.0430220	11.7323012	-0.8560147	0.3929607
‘study field‘multisubjects:‘funding mechanism‘NIFA NON FORMULA	-30.3006004	12.0361919	-2.5174574	0.0125637
‘study field‘other:‘funding mechanism‘NIFA NON FORMULA	1.5596555	14.6518619	0.1064476	0.9153284

Conclusion

Based on the regression results, we find out that the factors that matters the publication numbers are not the same as the factors that matters the maximium citation number.

For the publication numbers, we have strong evidence to say that the increase of 1,000,000 us dollar total expenditure is related to the increase of 2 publications, and we have weak evidence to say the more total days spent on the project also leads to increase of publication with 10000 days leader to increase of 1.8 publications. Besides, we have weak evidence to say the interaction of funding mechanisum: NIFA NON FORMULA and other study fields contributes to the increase of publication number. Once these two interacted, it leades to 4.61 publication increase.

For the maximal citation number, we have strong evidence to say that the funding mechanisum: NIFA NON FORMULA and the time since the project ends contributes to the increase of maximium citation number of the project's publiations. For a Nifa non formula mechanism, the max citation is 2.5. For the increase of 100 day the project ends, the max citation increase 1. Meanwhile, we have moderate evidence on the interaction of mechanism NIFA Non furmula and the study field: agriculture and medical and multisubjects that they leads to the decrease of maximium citation number on 25.32, 33.97, 32.06, which means that these field are not attract readers a lot. Besides, we weak evidence to say that 100 more total spend date, 1.131 more the maximium citation.

Reference

Distributed Representations of Words and Phrases and their Compositionality; Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean; comhttps://arxiv.org/pdf/1310.4546.pdf

Journal Ranking level http://www.scimagojr.com/journalrank.php?year=2017

Appendix

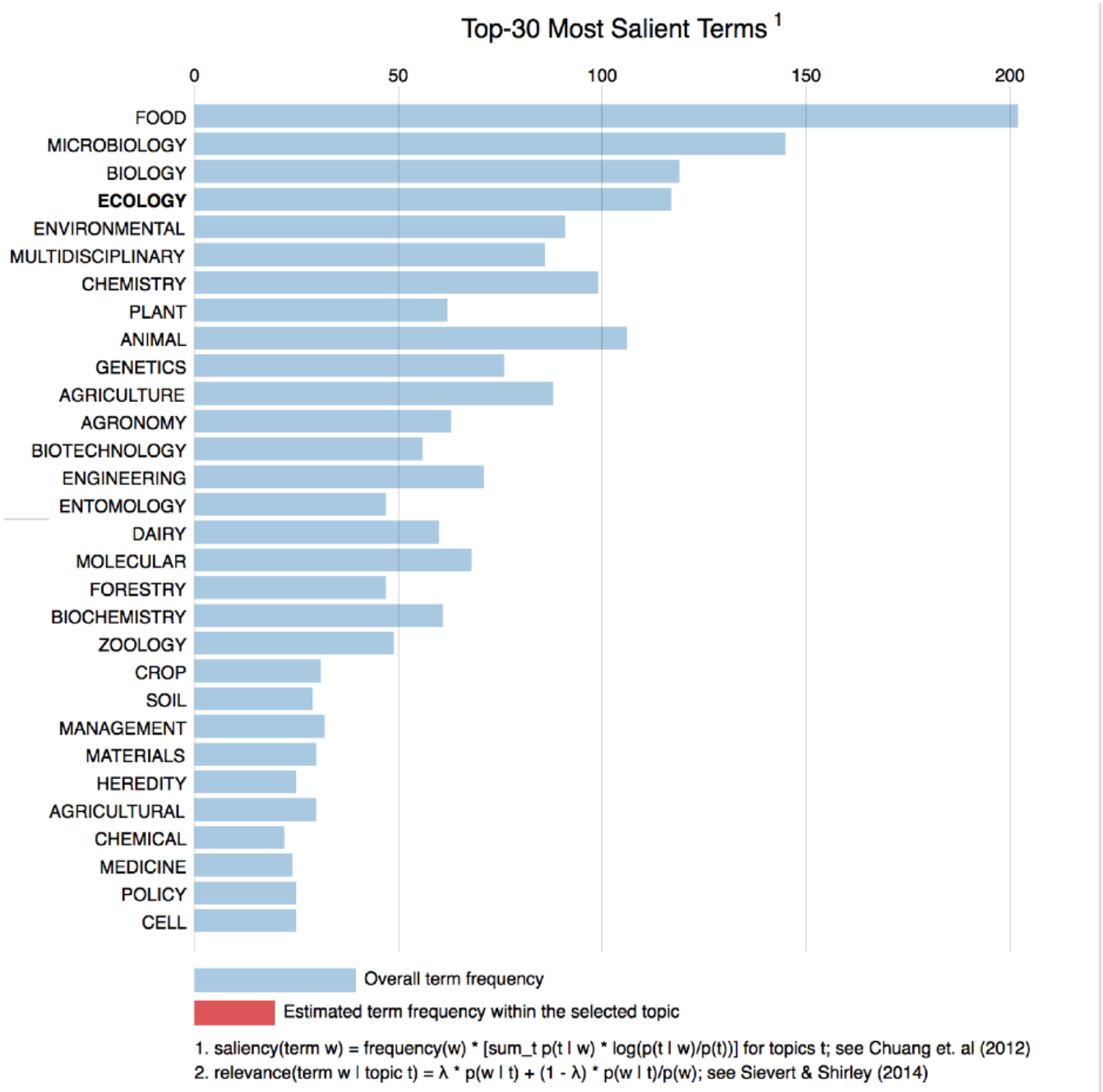

Figure 9: Top 30 key words among the subjects of the projects.