



**Determining cognitive distance between publication portfolios of  
evaluators and evaluatees in research evaluation: A case study of  
Biology department**

**TECHNICAL REPORT**

**A. I. M. Jakaria Rahman and Raf Guns**

[jakaria.rahman@uantwerpen.be](mailto:jakaria.rahman@uantwerpen.be), [raf.guns@uantwerpen.be](mailto:raf.guns@uantwerpen.be)  
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences,  
University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium

**Antwerp, 2017**

This technical report is prepared in the context of A. I. M. Jakaria Rahman's PhD project on *Determining cognitive distance between publication portfolios of evaluators and evaluatees in research evaluation: Exploration of informetric methods*. Similar technical reports on Biomedical Sciences, Chemistry Pharmaceutical Sciences, Physics and Veterinary Sciences department are also available at the institutional repository of the University of Antwerp (<https://repository.uantwerpen.be>).

# Table of Contents

<b>List of Tables .....</b>	<b>v</b>
<b>List of Figures.....</b>	<b>vi</b>
<b>1 Introduction.....</b>	<b>1</b>
<b>2 Cognitive distance based on Web of Science subject categories.....</b>	<b>3</b>
2.1 Data collection process .....	3
a) Research groups data retrieval .....	3
b) Panel members data retrieval.....	4
2.2 Correlation between publication profiles of research groups together and panel.....	5
a) Pearson's correlation coefficient and Spearman's rank-order correlation coefficient ..	5
b) Top-Down correlation coefficient .....	7
2.3 Web of Science subject categories similarity matrix.....	9
2.4 Web of Science subject categories overlay map creation .....	12
2.5 Bootstrapping and confidence intervals.....	13
2.6 Barycenter method .....	14
a) Barycenter calculation .....	14
b) Euclidean distance between barycenters .....	16
c) Barycenter overlay map.....	17
2.7 Similarity-adapted publication vector method.....	20
a) Similarity-adapted publication vector calculation .....	20
b) Euclidean distance between similarity-adapted publication vectors .....	21
c) Similarity-adapted publication vector overlay map.....	23
2.8 Weighted cosine similarity method .....	25
<b>3 Cognitive distance based on journals.....</b>	<b>28</b>
3.1 Data collection process .....	28
3.2 Correlation between publication profiles of research groups together and panel.....	29
a) Pearson's correlation coefficient and Spearman's rank-order correlation coefficient	29

b) Top-Down correlation coefficient .....	31
3.3 Journal similarity matrix .....	32
3.4 Journal overlay map creation .....	33
3.5 Barycenter method .....	34
a) Barycenter calculation .....	34
b) Euclidean distance calculation between barycenters .....	37
c) Barycenter overlay map .....	38
3.6 Similarity-adapted publication vector method .....	40
a) Similarity-adapted publication vector calculation .....	40
b) Euclidean distance between similarity-adapted publication vectors .....	42
3.7 Weighted cosine similarity method .....	43
<b>4 Heat map .....</b>	<b>46</b>
<b>5 Programming code in Python .....</b>	<b>48</b>
<b>References .....</b>	<b>52</b>
<b>Appendix A .....</b>	<b>54</b>
<b>Appendix B .....</b>	<b>62</b>
<b>Appendix C .....</b>	<b>70</b>

## List of Tables

Table 1: Publication statistics of Biology research groups (2004-2010) .....	3
Table 2. Publication statistics of Biology panel members .....	4
Table 3. Euclidean distances between barycenter of Biology individual research groups, panel members, research groups together and panel using WoS SCs VOS map.....	17
Table 4. Euclidean distances between SAPVs of Biology individual groups, panel members, research groups together and panel in WoS SCs similarity matrix .....	22
Table 5. WCS value of Biology individual research groups, panel members, research groups together and panel using WoS SCs similarity matrix.....	27
Table 6. WCD value between Biology individual research groups, panel members, groups and panel using WoS SCs similarity matrix .....	27
Table 7. Pearson and Spearman correlation between three methods using data from Biology individual research groups and panel members.....	27
Table 8. Euclidean distances between barycenter of Biology individual research groups, panel members, research groups together and panel using the journal VOS map.....	38
Table 9. Euclidean distances between SAPV of Biology individual research groups, panel members, research groups together and panel using the journal similarity matrix.....	43
Table 10. WCS value of the Biology groups, panel members, panel and research groups together using the journal similarity matrix .....	45
Table 11. WCD value of the Biology groups, panel members, panel and research groups together using the journal similarity matrix .....	45

## List of Figures

Figure 1. Excerpt of Biology research groups and panel members_WoS SCs.xlsx file .....	5
Figure 2. Excerpt of the Biology panel and research groups together_WoS SCs.xlsx file.....	6
Figure 3. Excerpt of the Biology Panel and groups together_WoS SCs - joined.xlsx file .....	6
Figure 4. Log-log plot of the number of publications (log-log scale) per WoS SC for the panel (vertical axis) and research groups together (horizontal axis) of the Biology department.....	7
Figure 5. Excerpt of the map10.paj file .....	9
Figure 6. Transformation of WoS SCs similarity matrix to Kamada-Kawai map and VOS map .....	10
Figure 7. VOSviewer message before choosing Kamada Kawai map or VOS map data.....	11
Figure 8. Excerpt of WoS SCs Kamada-Kawai map data from map10.paj file.....	11
Figure 9. Excerpt of WoS SCs VOS map data from the map10.paj file.....	11
Figure 10. Excerpt of WoS SCs VOS map .....	12
Figure 11. BIOL-B research group's publication overlay map in WoS SCs .....	13
Figure 12. Barycenter coordinates of the Biology individual research groups, panel members, research groups together and panel using the WoS SCs VOS map.....	15
Figure 13. Excerpt of Euclidean distances matrix of barycenter of the Biology individual research groups, panel members, research groups together and panel using WoS SCs VOS map .....	16
Figure 14. Barycenter overlay map of Biology individual research groups, panel members (PM), research groups together and panel in WoS SCs.....	18
Figure 15. Barycenter map of Biology individual research groups, panel members (PM), research groups together and panel in WoS SCs (zoomed).....	18
Figure 16. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together (groups) with their confidence regions.....	19
Figure 17. Excerpt of WoS SCs similarity matrix .....	20

Figure 18. Excerpt of SAPV of the Biology individual research groups, panel members, research groups together and panel using WoS SCs similarity matrix .....	21
Figure 19. Excerpt of pairwise Euclidean distance matrix between SAPVs of Biology individual research groups, panel members, research groups together and panel together using WoS SCs similarity matrix.....	22
Figure 20. Excerpt of BIOL-B.csv file .....	24
Figure 21. Location of the SAPV of BIOL-B in the WoS SCs similarity matrix.....	24
Figure 22. Excerpt of WCS value matrix of the Biology individual research groups, panel members, research groups together and panel using WoS SCs similarity matrix.....	26
Figure 23. Scatter plot of the correlation between barycenter, SAPV and WCS.....	28
Figure 24. Excerpt of Biology research groups and panel_journal title.xlsx file.....	28
Figure 25. Excerpt of the Biology panel and research groups together_journals title.xlsx file .....	30
Figure 26. Excerpt of the Biology panel and research groups together_journals title-joined.xlsx file.	30
Figure 27. Log-log plot of the number of publications (log-log scale) per journals for the panel (horizontal axis) and research groups together (vertical axis) of the Biology department....	31
Figure 28. Excerpt of the journals VOS map data .....	33
Figure 29. Journal overlay map of the BIOL-B research group .....	33
Figure 30. Excerpt of short form to full journal titles .....	35
Figure 31. Excerpt of journal name change.xlsx file .....	36
Figure 32. Barycenter coordinates of the Biology individual research groups, panel members, research groups together, and panel using journal VOS map.....	37
Figure 33. Excerpt of Euclidean distances matrix of the barycenter of the Biology groups, panel members, research groups together and panel using the journal VOS map .....	37
Figure 34. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together) .....	39

Figure 35. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together (zoomed) .....	39
Figure 36. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together with their confidence regions .....	40
Figure 37. Excerpt of SAPV of the Biology research groups, research groups together, panel members and panel using journal similarity matrix.....	41
Figure 38. Excerpt of pairwise Euclidean distances matrix between the SAPV of the Biology individual research groups, panel members, panel and research groups together using the journal similarity matrix.....	42
Figure 39. Excerpt of WCS value matrix of the Biology individual research groups, panel members, groups and panel using the journal similarity matrix .....	44
Figure 40. Excerpt of the dissimilarities/distances between panel members and individual research groups according to each of the six approaches .....	47
Figure 41. Heat map with hierarchical clustering based on correlation coefficient between six approaches in the Biology department .....	48
Figure 42. WoS SCs overlay map of BIOL-A research group's publications.....	54
Figure 43. WoS SCs overlay map of BIOL-B research group's publications.....	54
Figure 44. WoS SCs overlay map of BIOL-C research group's publications.....	55
Figure 45. WoS SCs overlay map of BIOL-D research group's publications.....	55
Figure 46. WoS SCs overlay map of BIOL-E research group's publications .....	56
Figure 47. WoS SCs overlay map of BIOL-F research group's publications .....	56
Figure 48. WoS SCs overlay map of BIOL-G research group's publications.....	57
Figure 49. WoS SCs overlay map of BIOL-H research group's publications.....	57
Figure 50. WoS SCs overlay map of BIOL-I research group's publications .....	58
Figure 51. WoS SCs overlay map of Biology research groups' publications.....	58



Figure 52. WoS SCs overlay map of PM1's publications .....	59
Figure 53. WoS SCs overlay map of PM2's publications .....	59
Figure 54. WoS SCs overlay map of PM3's publications .....	60
Figure 55. WoS SCs overlay map of PM4's publications .....	60
Figure 56. WoS SCs overlay map of PM5's publications .....	61
Figure 57. WoS SCs overlay map of panel's publications .....	61
Figure 58. SAPV of the BIOL-A research group's publications in WoS SCs similarity matrix.....	62
Figure 59. SAPV of the BIOL-B research group's publications in WoS SCs similarity matrix.....	62
Figure 60. SAPV of the BIOL-C research group's publications in WoS SCs similarity matrix.....	63
Figure 61. SAPV of the BIOL-D research group's publications in WoS SCs similarity matrix.....	63
Figure 62. SAPV of the BIOL-E research group's publications in WoS SCs similarity matrix.....	64
Figure 63. SAPV of the BIOL-F research group's publications in WoS SCs similarity matrix.....	64
Figure 64. SAPV of the BIOL-G research group's publications in WoS SCs similarity matrix.....	65
Figure 65. SAPV of the BIOL-H research group's publications in WoS SCs similarity matrix.....	65
Figure 66. SAPV of the BIOL-I research group's publications in WoS SCs similarity matrix.....	66
Figure 67. SAPV of the Biology research group's publications in WoS SCs similarity matrix.....	66
Figure 68. SAPV of the PM1's publications in WoS SCs similarity matrix .....	67
Figure 69. SAPV of the PM2's publications in WoS SCs similarity matrix .....	67
Figure 70. SAPV of the PM3's publications in WoS SCs similarity matrix .....	68
Figure 71. SAPV of the PM4's publications in WoS SCs similarity matrix .....	68
Figure 72. SAPV of the PM5's publications in WoS SCs similarity matrix .....	69

Figure 73. SAPV of the panel publications in WoS SCs similarity matrix .....	69
Figure 74. Journal overlay map of BIOL-A research group's publications .....	70
Figure 75. Journal overlay map of BIOL-B research group's publications.....	70
Figure 76. Journal overlay map of BIOL-C research group's publications.....	71
Figure 77. Journal overlay map of BIOL-D research group's publications .....	71
Figure 78. Journal overlay map of BIOL-E research group's publications.....	72
Figure 79. Journal overlay map of BIOL-F research group's publications .....	72
Figure 80. Journal overlay map of BIOL-G research group's publications .....	73
Figure 81. Journal overlay map of BIOL-H research group's publications .....	73
Figure 82. Journal overlay map of BIOL-I research group's publications .....	74
Figure 83. Journal overlay map of Biology research groups' publications .....	74
Figure 84. Journal overlay map of PM1's publications .....	75
Figure 85. Journal overlay map of PM2's publications .....	75
Figure 86. Journal overlay map of PM3's publications .....	76
Figure 87. Journal overlay map of PM4's publications .....	76
Figure 88. Journal overlay map of PM5's publications .....	77
Figure 89. Journal overlay map of the panel's publications .....	77

## 1 Introduction

We study the problem of composing an expert panel, such that the individual panel members' expertise covers the specific subdomains in the discipline where the units of assessment (in our case: research groups) have publications. We explore expertise overlap between panel and research groups through publishing in the same or similar Web of Science subject categories (WoS SCs) and journals. We use the data collected in the framework of completed research evaluations by the University of Antwerp (Belgium) through site visits by the expert panel members. We specifically focus on the situation where the expert panel needs to evaluate all the research groups of a department.

Research evaluations carried out at the University of Antwerp are organized by its Department of Research Affairs and Innovation (ADOC). At the start of a research evaluation, a department – typically encompassing several research groups – is invited to suggest potential panel chairs in addition to those suggested by the ADOC. Preferably, chairs are appointed as full professor, have an excellent publication record, have experience in research evaluations, are editors or board members of important journals, and possess academic management experience. The ADOC verifies whether proposed panel chairs and members have no prior involvement (i.e. no prior joint affiliations, no co-publications, no common projects) with the assessed research groups, and further checks if they are scholars with a prominent publication record in recent years, a proven track record of training young researchers, and sufficient experience in research policy, preferably in academic leadership positions. Furthermore, proposed panel chairs and members are preferably not affiliated with any Flemish institution of higher education and have no formal links to the University of Antwerp. The department that is being evaluated is also allowed to suggest potential panel members, but it should be noted that it is eventually the chair's prerogative to decide on the final composition of the panel.

The combined expertise of all panel members is to cover all subdomains in the discipline that is being evaluated and the panel is preferably balanced in terms of gender and nationality. When a sufficient number of professors have agreed to be on the panel, the university's research council ratifies the panel composition. Furthermore, all research groups belonging to a specific department (e.g., Biology) are to be evaluated by the same panel and the language of communication is English. Following the Dutch Standard Evaluation Protocol (VSNU,

2003; VSNU, KNAW, & NWO, 2014), the peer panels assess the quality, the productivity, the relevance and the viability of each research group.

These evaluations consider the entire research groups' scientific activity for a specific period, typically 8 years preceding the year of evaluation. All articles, letters, notes, proceeding papers, and reviews by the research groups published during the reference period are included in the evaluation. In this report, we consider only the publications that are index in Science Citation Index Expanded (SCIE ) and Social Sciences Citation Index (SSCI) of WoS.

Research groups at the University of Antwerp (Belgium) consist of professors (of all ranks), research and teaching assistants, and researchers (PhD students and postdocs). A research group consists either of one professor assisted by junior and/or senior researchers, or of a group of professors and a number of researchers linked to them.

An expert panel typically consists of independent specialists, and is multidisciplinary and/or interdisciplinary in its composition; each of the members are recognized experts in at least one of the fields addressed by the department under evaluation. However, the degree to which the expertise of the panel (members) overlaps with the expertise of the research groups has not been quantified to date. The goal is therefore to present informetric methodologies to assess the congruence of panel expertise and research interests in the units under assessment. As such, we present a bibliometric analysis of the overlap of expertise between research groups in the Departments of Biology and the respective expert panels based on research evaluations carried out at the University of Antwerp.

In this technical report, we present the Biology department's research groups and panel members. We describe our methods step by step. This report is divided into four parts. Firstly, we describe the technical steps for all of our three methods (barycenter, similarity-adapted publication vector, and weighted cosine similarity) using WoS SCs (Section 2). Secondly, we present the three methods using journals (section 3). In the third and fourth part, we present a heat map of spearman rank-order correlation coefficient between each pair of the six methods (section 4) and the programming code for the main methods used respectively (section 5). Finally, we present overlay maps and location of similarity adapted publication vector of Biology individual research groups, all research groups together, panel members and panel (all panel members together) in WoS SCs and journals in the appendix.

## 2 Cognitive distance based on Web of Science subject categories

### 2.1 Data collection process

We collect data from the 2011 assessment of the nine research groups of the Department of Biology, University of Antwerp. First, from ADOC, we collect all the WoS accession numbers of the publications of each research group. We replace the name of the research groups with code names BIOL-A, BIOL-B etc.

#### *a) Research groups data retrieval*

We remove the prefix ‘WOS:’ from the accession numbers and use a Python script to put ‘OR’ in between the accession numbers to create a long search string. We do a basic search in WoS with the accession numbers of each research group, keeping the time span to all years and searching SCIE and SSCI. We use the ‘Analyze Results’ option in the WoS, and rank the records by WoS SCs with the minimum set to 1. We save the resulting list as ‘analyze.txt’ and subsequently save a copy of the file named ‘[Research group code]\_WoS SCs.txt’, for example ‘BIOL-A\_WoS SCs.txt’ and keep both files.

**Table 1: Publication statistics of Biology research groups (2004-2010)**

<b>Group code</b>	<b>Number of Publications</b>	<b>Number of Journals</b>	<b>Number of WoS SCs</b>
BIOL-A	168	53	26
BIOL-B	58	33	13
BIOL-C	212	75	36
BIOL-D	176	68	26
BIOL-E	169	69	28
BIOL-F	58	35	18
BIOL-G	280	139	55
BIOL-H	67	42	25
BIOL-I	86	52	24
<b>All groups</b>	<b>1158</b>	<b>372</b>	<b>90</b>

Table 1 lists the publication profile of the Biology research groups during the eight years preceding their evaluation. The Biology research groups generated 1158 publications in 372 journals. Members of two research groups co-authored 113 publications, while member of three research groups co-authored three publications. In total, their publications are distributed over 90 WoS SCs.

We combine the search sets for each research group from the search history of the WoS, and get the data for the publications of the department as a whole. In this way, any publication that has been co-authored by members of two or more research groups is counted only once. We use the ‘Analyze Results’ option in the WoS, and rank the record by WoS SCs with the minimum set to 1. We save the resulting list as ‘analyze.txt’ and subsequently save a copy of the file named ‘Groups together\_WoS SCs.txt’.

### ***b) Panel members data retrieval***

The Biology panel was composed of five panel members (including the chair). We have obtained the names and curricula vitae of the panel members from the Department of Research Affairs. We replace the original name of each panel member with a code name: PM1, PM2 etc. We perform an advanced search for each panel member in WoS through checking the SCIE and SSCI. All the publications of the individual panel members up to the year of assessment (2011) were taken into account. We use the ‘Analyze Results’ option in the WoS, and rank the record by WoS SCs with the minimum set to 1. We save the resulting list as ‘analyze.txt’ and subsequently save a copy of the file named ‘[PM code]\_WoS SCs.txt’ for example, ‘PM1\_WoS SCs.txt’.

**Table 2. Publication statistics of Biology panel members**

<b>Panel code</b>	<b>Number of Publications</b>	<b>Number of Journals</b>	<b>Number of WoS SCs</b>
PM1	146	48	20
PM2	117	49	24
PM3	76	35	15
PM4	185	49	13
PM5	262	76	28
<b>Panel</b>	<b>786</b>	<b>217</b>	<b>54</b>

Table 2 lists the publication profile of the Biology panel members. The combined publication output of the Biology panel members consists of 786 publications, none of which is co-authored publications between panel members. The number of publications per panel member ranges from 76 to 262. In total, these publications appeared in 217 different journals and are assigned to 54 different WoS SCs.

	A	B	C	D	E
1	<b>Web of Science Categories</b>	<b>records</b>	<b>% of 1156</b>		
2	ECOLOGY	256	22.203		
3	ENVIRONMENTAL SCIENCES	215	18.647		
4	ZOOLOGY	170	14.744		
5	PLANT SCIENCES	136	11.795		
6	EVOLUTIONARY BIOLOGY	113	9.801		
7	MARINE FRESHWATER BIOLOGY	93	8.066		
8	TOXICOLOGY	72	6.245		
9	BEHAVIORAL SCIENCES	68	5.898		
10	BIOLOGY	67	5.811		
11	BIOCHEMISTRY MOLECULAR BIOLOGY	62	5.377		
12	BIODIVERSITY CONSERVATION	60	5.204		
<span>BIOL-A</span> / <span>BIOL-B</span> / <span>BIOL-C</span> / <span>BIOL-D</span> / <span>BIOL-E</span> / <span>BIOL-F</span> / <span>BIOL-G</span> / <span>BIOL-H</span> / <span>BIOL-I</span> / <b>Groups</b> / <span>PM1</span> / <span>PM2</span>					

**Figure 1. Excerpt of Biology research groups and panel members\_WoS SCs.xlsx file**

We combine the search sets for each panel member from the search history of the WoS, and get the result for the panel as a whole. In this way, any co-authored publication between two or more panel members is counted only once. Again, we use the ‘Analyze Results’ option in the WoS, and rank the record by WoS SCs with the minimum set to 1. We save the resulting list as ‘analyze.txt’ and subsequently save a copy of the file named ‘Panel\_WoS SCs.txt’.

The downloaded data files, ‘[Research group code]\_WoS SCs. txt’, ‘[PM code]\_WoS SCs. txt’, ‘Groups\_WoS SCs.txt’ and ‘Panel\_WoS SCs.txt’, have been exported to an MS Excel file. The sheets in the Excel file contain data on and are named after the research groups’ code names (BIOL-A, BIOL-B, BIOL-C, etc.), the panel members’ code names, (PM1, PM2, PM3, etc.), Panel together and Groups together. The Excel file is saved as ‘Biology research groups and panel\_WoS SCs.xlsx’ (Figure 1).

## **2.2 Correlation between publication profiles of research groups together and panel**

### ***a) Pearson’s correlation coefficient and Spearman’s rank-order correlation coefficient***

We determine the correlation between the publication output of research groups together and panel, using Pearson’s correlation coefficient and Spearman’s rank-order correlation coefficient for the numbers of publications per WoS SC. We make an Excel file ‘Biology panel and research groups together\_WoS SCs.xlsx’ (Figure 2) and export data from ‘Panel\_WoS SCs.txt’ and ‘Groups together\_WoS SCs.txt’ in two different sheets.

A Python script ‘join-sheets.py’ is used to take the data of the two sheets and join it into one. We run the program as:

```
python join-sheets.py "Biology panel and research groups together_WoS SCs.xlsx"
```

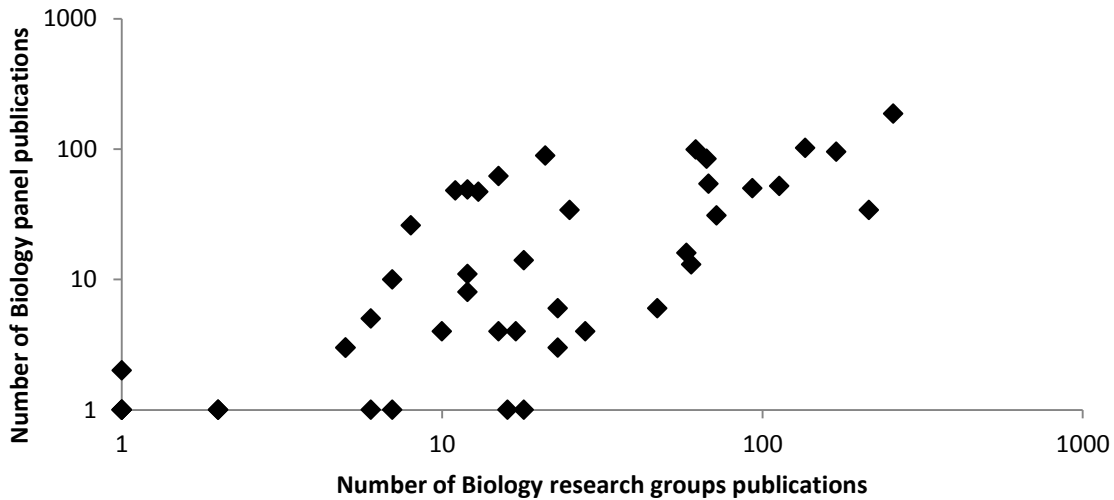
	A	B	C		A	B	C
1	<b>Web of Science Categories</b>	<b>records</b>	<b>% of 1156</b>	1	<b>Web of Science Categories</b>	<b>records</b>	<b>% of 792</b>
2	ECOLOGY	256	22.203	2	ECOLOGY	187	23.791
3	ENVIRONMENTAL SCIENCES	215	18.647	3	PLANT SCIENCES	102	12.977
4	ZOOLOGY	170	14.744	4	BIOCHEMISTRY MOLECULAR BIOLOGY	99	12.595
5	PLANT SCIENCES	136	11.795	5	ZOOLOGY	95	12.087
6	EVOLUTIONARY BIOLOGY	113	9.801	6	ENTOMOLOGY	89	11.323
7	MARINE FRESHWATER BIOLOGY	93	8.066	7	BIOLOGY	84	10.687
8	TOXICOLOGY	72	6.245	8	ENDOCRINOLOGY METABOLISM	62	7.888
9	BEHAVIORAL SCIENCES	68	5.898	9	BEHAVIORAL SCIENCES	54	6.87
10	BIOLOGY	67	5.811	10	EVOLUTIONARY BIOLOGY	52	6.616
11	BIOCHEMISTRY MOLECULAR BIOLOGY	62	5.377	11	MARINE FRESHWATER BIOLOGY	50	6.361
12	BIODIVERSITY CONSERVATION	60	5.204	12	FISHERIES	49	6.234
13	GENETICS HEREDITY	58	5.03	13	PHYSIOLOGY	48	6.107
14	FORESTRY	47	4.076	14	CELL BIOLOGY	47	5.98
15	GEOSCIENCES MULTIDISCIPLINARY	28	2.428	15	MULTIDISCIPLINARY SCIENCES	34	4.326
16	ENGINEERING ENVIRONMENTAL	26	2.255	16	ENVIRONMENTAL SCIENCES	34	4.326
17	MULTIDISCIPLINARY SCIENCES	25	2.168	17	TOXICOLOGY	31	3.944
18	BIOTECHNOLOGY APPLIED MICROBIOLOGY	23	1.995	18	BIOPHYSICS	26	3.308
19	AGRONOMY	23	1.995	19	GENETICS HEREDITY	16	2.036
20	ENTOMOLOGY	21	1.821	20	NEUROSCIENCES	14	1.781
21	CHEMISTRY ANALYTICAL	21	1.821	21	BIODIVERSITY CONSERVATION	13	1.654
22	NEUROSCIENCES	18	1.561	22	VETERINARY SCIENCES	11	1.399
23	BIOCHEMICAL RESEARCH METHODS	18	1.561	23	PHARMACOLOGY PHARMACY	10	1.272
24	METEOROLOGY ATMOSPHERIC SCIENCES	17	1.474	24	GEOGRAPHY PHYSICAL	8	1.018

Figure 2. Excerpt of the Biology panel and research groups together\_WoS SCs.xlsx file

	A	B	C	D
1		<b>Web of Science Categories</b>	<b>records_x</b>	<b>records_y</b>
2	<b>0</b>	ECOLOGY	256	187
3	<b>1</b>	ENVIRONMENTAL SCIENCES	215	34
4	<b>2</b>	ZOOLOGY	170	95
5	<b>3</b>	PLANT SCIENCES	136	102
6	<b>4</b>	EVOLUTIONARY BIOLOGY	113	52
7	<b>5</b>	MARINE FRESHWATER BIOLOGY	93	50
8	<b>6</b>	TOXICOLOGY	72	31
9	<b>7</b>	BEHAVIORAL SCIENCES	68	54
10	<b>8</b>	BIOLOGY	67	84
11	<b>9</b>	BIOCHEMISTRY MOLECULAR BIOLOGY	62	99
12	<b>10</b>	BIODIVERSITY CONSERVATION	60	13
13	<b>11</b>	GENETICS HEREDITY	58	16
14	<b>12</b>	FORESTRY	47	6
15	<b>13</b>	GEOSCIENCES MULTIDISCIPLINARY	28	4
16	<b>14</b>	ENGINEERING ENVIRONMENTAL	26	0
17	<b>15</b>	MULTIDISCIPLINARY SCIENCES	25	34
18	<b>16</b>	BIOTECHNOLOGY APPLIED MICROBIOLOGY	23	6

Figure 3. Excerpt of the Biology Panel and groups together\_WoS SCs - joined.xlsx file





**Figure 4. Log-log plot of the number of publications (log-log scale) per WoS SC for the panel (vertical axis) and research groups together (horizontal axis) of the Biology department**

This produces a new Excel file called ‘Biology panel and groups together\_WoS SCs-joined.xlsx’ (Figure 3). To calculate the correlation, the value zero was kept on the corresponding WoS SCs in which either the panel or the groups had no publications (but not both). Using the data from the file, we calculate correlation coefficient in SPSS (Statistical Package for the Social Sciences) and find value ( $r = 0.78$ ,  $\rho = 0.53$ ). Figure 4 shows a log-log plot of the number of publications per WoS SCs for the Biology panel and research groups together.

### ***b) Top-Down correlation coefficient***

In some cases, the panel publications belong to a WoS SC in which the research groups have not published or vice versa, i.e. there are many zeroes on both sides. Since traditional correlation coefficients like Pearson’s and Spearman’s are not well-suited to zero-inflated data (i.e., data with a large amounts of zeroes), we adopt the top-down correlation coefficient (Iman & Conover, 1987). This correlation coefficient was found to be an adequate rank correlation coefficient for zero-inflated data (Huson, 2007). For a full description of the Top-down correlation coefficient we refer to Iman and Conover (1987). This coefficient places emphasis on the higher ranked data by computing the correlation using Savage scores derived from the ranked data.

Savage scores are calculated as follows:

$$S_i = \sum_{j=i}^n 1/j \quad (1)$$

where  $i$  is an item's rank among a set of  $n$  items. For instance, if  $n = 3$ , the three Savage scores are  $S_1 = 1 + \frac{1}{2} + \frac{1}{3}$ ,  $S_2 = \frac{1}{2} + \frac{1}{3}$ , and  $S_3 = \frac{1}{3}$ . The Top-down correlation coefficient is calculated as:

$$r_{td} = \left( \sum_{i=1}^n S_{R_i} S_{Q_i} - n \right) / (n - S_1) \quad (2)$$

where  $S$  is the Savage score,  $R_i$  and  $Q_i$  are the ranks of the data in the two samples, and  $n$  is the sample size. In case of ties, we use the average Savage score.

We use a Python script 'calc\_topdowncorr.py' (all core logic is in topdowncorr.py, see section 5) for top-down correlation taking into account formulas (1) and (2). We reuse the 'Biology panel and groups together\_WoS SCs - joined.xlsx' (Figure 3) file, but keep the zeros in the WoS SCs where neither the panel nor the research groups have publications. We run the program as:

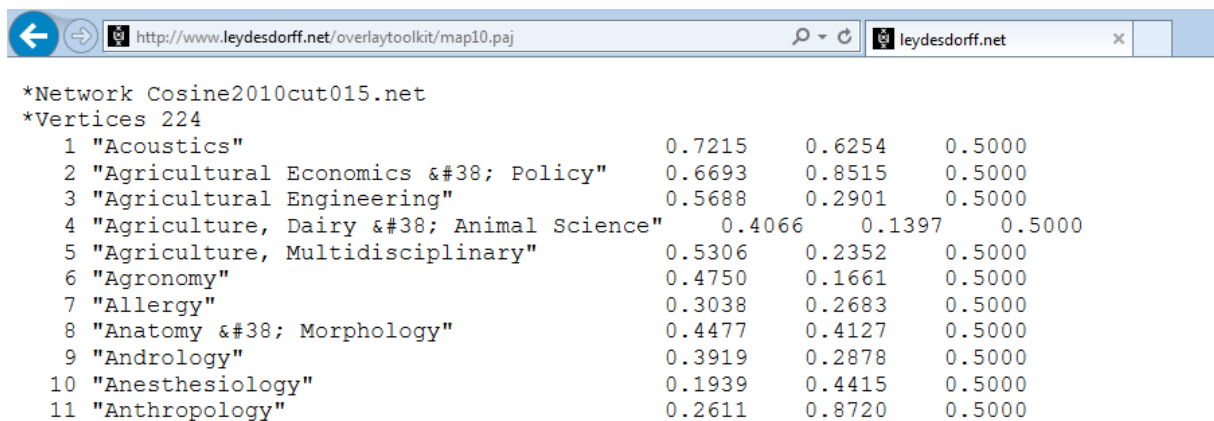
```
python calc_topdowncorr.py "Biology panel and research groups together_WoS SCs-joined.xlsx"
```

The outcome shows that the top-down correlation between Biology research groups together and the panel's profile in the WoS SCs is low (0.29).

In our opinion, the correlations are an insufficient measure in this case, as the similarity of WoS SCs is not taken into account here. This is reminiscent of the way diversity is sometimes studied using only the dimensions of variety and balance. As discussed by Stirling (2007), the additional dimension of disparity – the opposite concept of similarity – is needed to provide a complete picture. Likewise, a comparison of publication profiles based on WoS SCs that does not consider WoS SC similarity might yield distorted results.

## 2.3 Web of Science subject categories similarity matrix

We download the global map of science based on WoS SCs data made available at <http://www.leydesdorff.net/overlaytoolkit/map10.paj>. These authors (Leydesdorff & Rafols, 2009; Rafols, Porter, & Leydesdorff, 2010; Leydesdorff, Carley, & Rafols, 2013) created a matrix of citing to cited WoS SCs based on the SCIE and SSCI, which was subsequently normalized in the citing direction. Only cosine values > 0.15 were retained. The result is a symmetric N×N similarity matrix (here, N=224). If we interpret it as an adjacency matrix, we see that it is equivalent to a weighted network, in which similar categories are linked (the higher the link weight, the stronger the similarity). The file ‘map10.paj’ contains this weighted network of WoS SCs.



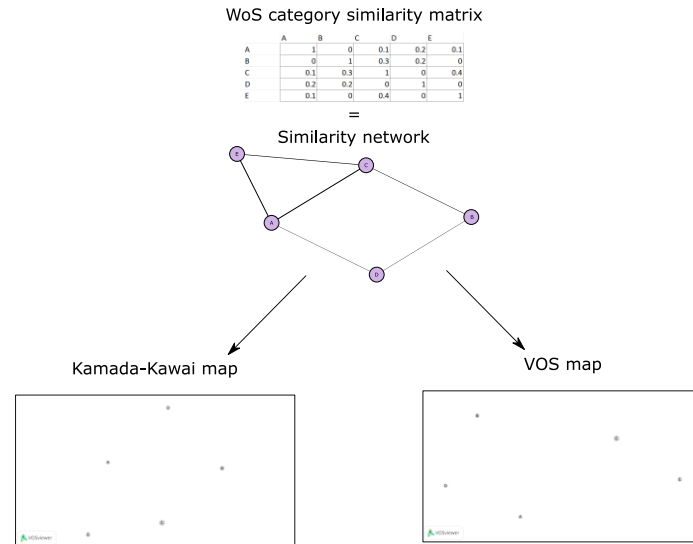
```
*Network Cosine2010cut015.net
*Vertices 224
 1 "Acoustics" 0.7215 0.6254 0.5000
 2 "Agricultural Economics &#38; Policy" 0.6693 0.8515 0.5000
 3 "Agricultural Engineering" 0.5688 0.2901 0.5000
 4 "Agriculture, Dairy &#38; Animal Science" 0.4066 0.1397 0.5000
 5 "Agriculture, Multidisciplinary" 0.5306 0.2352 0.5000
 6 "Agronomy" 0.4750 0.1661 0.5000
 7 "Allergy" 0.3038 0.2683 0.5000
 8 "Anatomy &#38; Morphology" 0.4477 0.4127 0.5000
 9 "Andrology" 0.3919 0.2878 0.5000
10 "Anesthesiology" 0.1939 0.4415 0.5000
11 "Anthropology" 0.2611 0.8720 0.5000
```

Figure 5. Excerpt of the map10.paj file

We download the ‘map10.paj’ (Figure 5) file and open the file in Pajek (available at <http://mrvar.fdv.uni-lj.si/pajek>) and save the network as ‘map10.net’. The information in the network file can be visualized. The subfield of bibliometric mapping is dedicated to the visualization, clustering and interpretation of similarity matrices or networks like the one we use. Many different algorithms or layout techniques have been developed for this purpose. We have used two layout techniques:

- i) Kamada-Kawai (Kamada & Kawai, 1989) is a spring-based layout algorithm for networks, which is implemented in, among others, Pajek (de Nooy, Mrvar, & Batagelj, 2012). Kamada-Kawai is the algorithm used by Rafols et al., (2010).

- ii) VOS (van Eck & Waltman, 2007) stands for ‘visualization of similarities’ and is a variant of multidimensional scaling (Borg & Groenen, 2005; van Eck, Waltman, Dekker, & van den Berg, 2010). It is implemented in VOSviewer and in recent versions of Pajek.



**Figure 6. Transformation of WoS SCs similarity matrix to Kamada-Kawai map and VOS map**

Figure 6 shows the transformation of WoS SC similarity matrix to Kamada-Kawai and VOS map. It provides an overview of the relations between similarity matrix, network and the two maps. Since the source data include all research fields included in the SCI and SSCI, the resulting maps are global maps of science (as opposed to local maps of science, which focus on one or a few disciplines).

We run VOSviewer (<http://www.vosviewer.com>) and click on ‘Create’ from the action tab. It offers to create a map based on a network. We select this option and in the next step through Pajek tab, we choose the ‘map10.net’ file and click on the next button. It prompts us to choose whether we want to use the coordinates that are in the file or want to calculate new ones (Figure 7).

We choose ‘Yes’ to keep using the Kamada-Kawai coordinates. We save the map as ‘Kamada-Kawai.txt’ file, export the data to an Excel file, and save as ‘WoS SCs\_Kamada-Kawai map.xlsx’ (Figure 8). Again, we run VOSviewer and click on ‘Create’ from the action tab. It offers to create a map based on a network. We select this option and in the next step through the Pajek tab, we choose the ‘map10.net’ file and click on the next button. It again

prompts us to choose whether we want to use the coordinates that are in the file or want to calculate new ones (Figure 7). We choose ‘No’ to let VOSviewer calculate the coordinates according to its own VOS algorithm (Figure 9).

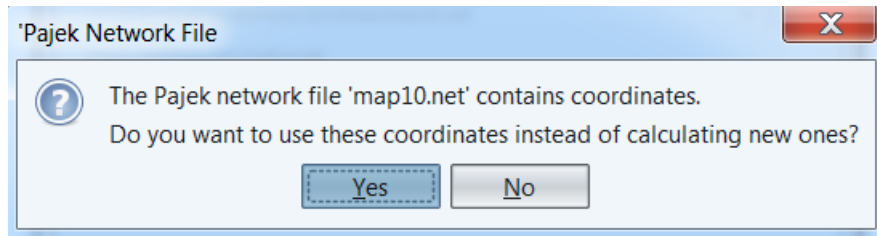


Figure 7. VOSviewer message before choosing Kamada Kawai map or VOS map data

	A	B	C	D	E	F
1	id	label	x	y	weight	cluster
2	1	Acoustics	0.7215	0.6254	4.596	2
3	2	Agricultural Economics, Policy	0.6693	0.8515	5.484	3
4	3	Agricultural Engineering	0.5688	0.2901	15.708	4
5	4	Agriculture, Dairy, Animal Science	0.4066	0.1397	6.208	1
6	5	Agriculture, Multidisciplinary	0.5306	0.2352	15.422	4
7	6	Agronomy	0.475	0.1661	10.28	4
8	7	Allergy	0.3038	0.2683	8.034	1
9	8	Anatomy, Morphology	0.4477	0.4127	28.204	1
10	9	Andrology	0.3919	0.2878	16.946	1
11	10	Anesthesiology	0.1939	0.4415	9.528	1

Figure 8. Excerpt of WoS SCs Kamada-Kawai map data from map10.paj file

	A	B	C	D	E	F
1	id	label	x	y	weight	cluster
2	1	Acoustics	-0.6842	-0.1274	4.596	2
3	2	Agricultural Economics, Policy	1.285	-0.1948	5.484	3
4	3	Agricultural Engineering	-0.6952	0.0854	15.708	4
5	4	Agriculture, Dairy, Animal Science	-0.2417	0.2277	6.208	1
6	5	Agriculture, Multidisciplinary	-0.4803	0.1691	15.422	4
7	6	Agronomy	-0.5502	0.1844	10.28	4
8	7	Allergy	0.104	0.2443	8.034	1
9	8	Anatomy, Morphology	0.0029	0.1922	28.204	1
10	9	Andrology	-0.0119	0.2208	16.946	1
11	10	Anesthesiology	0.3086	0.1703	9.528	1

Figure 9. Excerpt of WoS SCs VOS map data from the map10.paj file

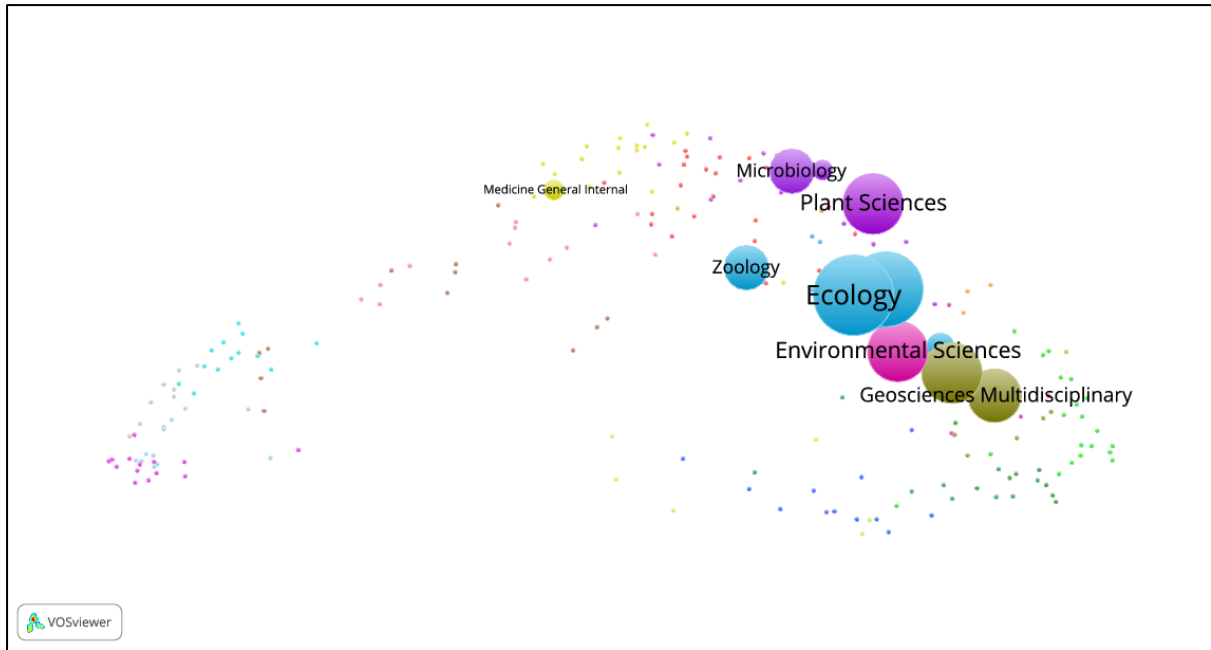
	A	B	C	D	E	F
1	id	label	x	y	weight	cluster
2	1	Acoustics	0.4329	-0.1442	4.596	2
3	2	Agricultural Economics Policy	-0.9319	-0.2762	5.484	3
4	3	Agricultural Engineering	0.699	0.0866	15.708	4
5	4	Agriculture, Dairy Animal Science	0.2792	0.3684	6.208	1
6	5	Agriculture, Multidisciplinary	0.5097	0.2382	15.422	4
7	6	Agromony	0.5917	0.2442	10.28	4
8	7	Allergy	-0.0432	0.5132	8.034	1
9	8	Anatomy Morphology	0.1085	0.3599	28.204	1
10	9	Andrology	0.0906	0.427	16.946	1
11	10	Anesthesiology	-0.2826	0.3844	9.528	1

**Figure 10. Excerpt of WoS SCs VOS map**

However, we have observed the coordinates of the VOS map that we derived from the map10.paj file is different that the VOS map available at <http://www.leydesdorff.net/overlay> toolkit while creating overlap map (Figure 10). We use this VOS map (Figure 10). as this map is readily available and applied for creating overlay maps (Leydesdorff, Carley, et al., 2013; Rafols et al., 2010). The details of obtaining this VOS map have been discussed in the next section. In this technical report, calculations of barycenters, Euclidean distance comparisons, and visual explorations are based on the VOS map of WoS SCs (Figure 10).

## 2.4 Web of Science subject categories overlay map creation

During data collection (see section 2.1, the resulting files are downloaded using the default name ‘analyze.txt’. We download the ‘WC10.exe’ program from <http://www.leydesdorff.net/overlay> toolkit. This file ‘analyze.txt’ transformed by the mini-program ‘WC10.exe’ to ‘WC10.vec’ for upload into Pajek as a vector, and generate files like ‘vos4.csv’, ‘vos6.csv’, and ‘vos19.csv’ for use in VOSviewer (with 4, 6 or 19 base colors for the clusters, respectively). We keep ‘analyze.txt’ and ‘WC10.exe’ in a folder and run the exe file. The program ‘WC10.exe’ generates three map files: ‘vos4.csv’, ‘vos6.csv’, and ‘vos19.csv’. We open the ‘vos19.csv’ in VOSviewer. For example, Figure 11 shows BIOL-B research group’s publications overlay map in WoS SCs.



**Figure 11. BIOL-B research group’s publication overlay map in WoS SCs**

The ‘vos4.csv’, ‘vos6.csv’, and ‘vos19.csv’ map files contain the VOS map as mentioned in the previous section. We save the map data to an Excel file, and save as ‘WoS SCs\_VOS\_map.xlsx’ (Figure 10).

We prepare overlay maps for each research group, each panel member, research groups together and panel (see the Appendix A).

## **2.5 Bootstrapping and confidence intervals**

The barycenter (discussed in section 2.6 and 3.5) and Similarity-adapted vector (SAPV) methods (discussed in section 2.7 and 3.6) determine cognitive distance, on the basis of the WoS SCs/journals in which the groups and panel members have published. In the same way, Weighted cosine similarity method (discussed in section 2.8 and 3.7) determine similarity on the basis of the WoS SCs/journals in which the groups and panel members have published. However, such information is not entirely deterministic; it is, for instance, dependent on the database used as well as environmental factors like the speed with which a journal processes a submission. It logically follows that small differences in Euclidean distances or similarity bear little meaning.

To study this problem in a more systematic way, we employ a bootstrapping approach in order to determine 95 % confidence intervals (CIs) to each Euclidean distance (both between barycenters and SAPVs) and similarity. If two CIs do not overlap, the difference between the distances is statistically significant at the 0.05 level. Although it is possible for overlapping CIs to have a statistically significant difference between the corresponding distances, the difference between the distances is less likely to have practical meaning.

Bootstrapping (Efron & Tibshirani, 1998) is a simulation-based method for estimating standard error and confidence intervals. Bootstrapping depends on the notion of a bootstrap sample. To determine a bootstrap sample for a panel member or research group with  $N$  publications, we randomly sample with replacement  $N$  publications from its set of publications. In other words, the same publication can be chosen multiple times. Some publications in the original data set will not occur in the bootstrap data set, whereas others will occur once, twice or even more times. From the bootstrap sample, one can calculate a bootstrap replication, in our case a barycenter using formula (3), an SAPV using formula (5), and WCS using formula (7).

By generating a large amount of independent bootstrap samples (in our case 1000) and each time calculating the bootstrap replication, we can approximate the variability within the data set. Since we have a two-sample problem (distance between two entities; Efron & Tibshirani, 1998, Ch. 8), we calculate the distances between pairs of bootstrap replications, from which we obtain a CI using a bootstrap percentile approach (Efron & Tibshirani, 1998, Ch. 13). In the case of WCS, we generate 1000 independent bootstrap sample for both entities and calculate the similarity between them using formula 7. A more detailed explanation and implementation of our method is available on Github (<http://nbviewer.jupyter.org/gist/rafguns/6fa3460677741e356538337003692389> and <http://nbviewer.jupyter.org/gist/rafguns/faff8dc090b67a783b85d488f88952ba>).

## **2.6 Barycenter method**

### ***a) Barycenter calculation***

The barycenter of a set of points (here: WoS SCs) with associated weights (here: number of publications) is defined as the point  $C = (C_1, C_2)$ , where



$$C_1 = \frac{\sum_{j=1}^N m_j L_{j,1}}{T} ; C_2 = \frac{\sum_{j=1}^N m_j L_{j,2}}{T} \quad (3)$$

Here,  $L_{j,1}$  and  $L_{j,2}$  are the horizontal and vertical coordinates of WoS SC  $j$  on the map,  $m_j$  is the number of publications in WoS SC  $j$ , and  $T = \sum_{j=1}^N m_j$  is the total number of publications of the entity (panel member, research group). Note that  $T$  is larger than the total number of publications as we use full counting of WoS SCs: if a publication appears in a journal belonging to two categories, it will be counted twice. For further elaboration on the barycenter method, we refer to (Jin & Rousseau, 2001; Rousseau, 1989; Verleysen & Engels, 2013, 2014).

Formula (3) is implemented in a Python script ‘barycenter-categories.py’ (the actual barycenter calculation is done in the barycenter function, see section 5) that takes as input the map file (‘WoS SC\_VOS\_map.xlsx’, Figure 10) and the weights (number of publications) per WoS SC (‘Biology research groups and panel\_WoS SCs.xlsx’, Figure 1), and calculates a barycenter for each entity (Figure 12). We run the program as:

```
python barycenter-categories.py "WoS SC_VOS_map.xlsx" "Biology research groups and panel_WoS SCs.xlsx"
```

	A	B	C
1		<b>x</b>	<b>y</b>
2	<b>BIOL-A</b>	0.149479	0.230169
3	<b>BIOL-B</b>	0.523074	0.105991
4	<b>BIOL-C</b>	0.433215	0.12318
5	<b>BIOL-D</b>	0.560874	0.113229
6	<b>BIOL-E</b>	0.207052	0.164653
7	<b>BIOL-F</b>	0.399432	0.353316
8	<b>BIOL-G</b>	0.289973	0.194414
9	<b>BIOL-H</b>	0.561755	0.073293
10	<b>BIOL-I</b>	0.461581	0.301902
11	<b>Groups</b>	0.357824	0.174109
12	<b>PM1</b>	0.488921	0.172867
13	<b>PM2</b>	0.124493	0.196331
14	<b>PM3</b>	0.413504	0.345149
15	<b>PM4</b>	0.365586	0.214172
16	<b>PM5</b>	0.247693	0.277106
17	<b>Panel</b>	0.303642	0.237966

**Figure 12. Barycenter coordinates of the Biology individual research groups, panel members, research groups together and panel using the WoS SCs VOS map**

This program calculates the barycenter and generates an output file ‘Biology research groups and panel\_WoS SCs-barycenter.xlsx’. Figure 12 shows the barycenter coordinates of the Biology individual research groups, panel members, research groups together and panel.

**b) Euclidean distance between barycenters**

Subsequently, we determine the Euclidean distance between the barycenters of different entities: individual research groups, research groups together, panel members and panel. The Euclidean distance between points C = (C<sub>1</sub>, C<sub>2</sub>) and D = (D<sub>1</sub>, D<sub>2</sub>) is calculated as follows:

$$d = \sqrt{(C_1 - D_1)^2 + (C_2 - D_2)^2}. \tag{4}$$

We use the implementation of Euclidean distance in `scipy.spatial.dist`. We note that the Python script ‘barycenter-categories.py’ executes both formula (3) and (4). The distances thus obtained should be interpreted as having arbitrary units on a ratio scale (Egghe & Rousseau, 1990). This means that there is a fixed meaningful zero (distance zero in the map), and distances can be compared in terms of percentage or fraction (e.g. the distance between A and B is 1.5 times larger than the distance between C and D).

	A	B	C	D	E	F	G	H	I	J	K	L
1		<b>BIOL-A</b>	<b>BIOL-B</b>	<b>BIOL-C</b>	<b>BIOL-D</b>	<b>BIOL-E</b>	<b>BIOL-F</b>	<b>BIOL-G</b>	<b>BIOL-H</b>	<b>BIOL-I</b>	<b>Groups</b>	<b>PM1</b>
2	<b>BIOL-A</b>	0	0.393692	0.303237	0.427692	0.087218	0.278643	0.144972	0.441114	0.32024	0.215756	0.344245
3	<b>BIOL-B</b>	0.393692	0	0.091488	0.038486	0.32142	0.276509	0.249309	0.050649	0.205335	0.178739	0.075092
4	<b>BIOL-C</b>	0.303237	0.091488	0	0.128046	0.229934	0.232602	0.159977	0.137881	0.180959	0.090981	0.074645
5	<b>BIOL-D</b>	0.427692	0.038486	0.128046	0	0.357539	0.289319	0.282805	0.039945	0.213206	0.21198	0.093455
6	<b>BIOL-E</b>	0.087218	0.32142	0.229934	0.357539	0	0.269451	0.088099	0.366279	0.289175	0.151068	0.281988
7	<b>BIOL-F</b>	0.278643	0.276509	0.232602	0.289319	0.269451	0	0.192954	0.323669	0.080659	0.183974	0.201421
8	<b>BIOL-G</b>	0.144972	0.249309	0.159977	0.282805	0.088099	0.192954	0	0.29755	0.202492	0.070825	0.200112
9	<b>BIOL-H</b>	0.441114	0.050649	0.137881	0.039945	0.366279	0.323669	0.29755	0	0.249593	0.22749	0.123368
10	<b>BIOL-I</b>	0.32024	0.205335	0.180959	0.213206	0.289175	0.080659	0.202492	0.249593	0	0.16461	0.1319
11	<b>Groups</b>	0.215756	0.178739	0.090981	0.21198	0.151068	0.183974	0.070825	0.22749	0.16461	0	0.131102
12	<b>PM1</b>	0.344245	0.075092	0.074645	0.093455	0.281988	0.201421	0.200112	0.123368	0.1319	0.131102	0
13	<b>PM2</b>	0.042063	0.408691	0.317271	0.444224	0.088429	0.316601	0.165491	0.454243	0.353234	0.234388	0.365183
14	<b>PM3</b>	0.287975	0.263063	0.222842	0.274782	0.274228	0.016271	0.194887	0.309651	0.064666	0.179874	0.188066
15	<b>PM4</b>	0.216698	0.191064	0.113372	0.219834	0.166087	0.143202	0.078152	0.241514	0.130045	0.040807	0.130068
16	<b>PM5</b>	0.108854	0.324214	0.241063	0.353465	0.119572	0.169802	0.092873	0.374398	0.21532	0.150788	0.262786
17	<b>Panel</b>	0.15436	0.256062	0.173104	0.28588	0.121261	0.149938	0.045646	0.306169	0.17039	0.083746	0.196383

**Figure 13. Excerpt of Euclidean distances matrix of barycenter of the Biology individual research groups, panel members, research groups together and panel using WoS SCs VOS map**

**Table 3. Euclidean distances between barycenter of Biology individual research groups, panel members, research groups together and panel using WoS SCs VOS map**

	Groups	BIOL-A	BIOL-B	BIOL-C	BIOL-D	BIOL-E	BIOL-F	BIOL-G	BIOL-H	BIOL-I
Panel	0.084	0.154	0.256	0.173	0.286	0.121	0.150	0.046	0.306	0.170
PM1	0.131	0.344	<b><u>0.075</u></b>	<b><u>0.075</u></b>	<b><u>0.093</u></b>	0.282	0.201	0.200	<b><u>0.123</u></b>	<b><u>0.132</u></b>
PM2	0.234	<b><u>0.042</u></b>	0.409	0.317	0.444	<b><u>0.088</u></b>	0.317	0.165	0.454	0.353
PM3	0.180	<b><u>0.288</u></b>	0.263	0.223	0.275	0.274	<b><u>0.016</u></b>	0.195	0.310	<b><u>0.065</u></b>
PM4	0.041	0.217	0.191	<b><u>0.113</u></b>	0.220	<b><u>0.166</u></b>	0.143	<b><u>0.078</u></b>	0.242	0.130
PM5	0.151	0.109	0.324	0.241	0.353	<b><u>0.120</u></b>	0.170	<b><u>0.093</u></b>	0.374	0.215

For each research group we determined the panel member at the shortest distance. Average of shortest distance is 0.073 (SD 0.030). The number in the row of this panel member is indicated in bold and underlined. Distances whose confidence intervals overlap with that of the shortest distance are in bold (same column).

From the matrix of Euclidean distances, which includes distances between all entity pairs (Figure 13), we extract Table 3 containing only the distances between the research groups and research groups together on the one hand and the panel and panel members on the other, for the convenience of analysis.

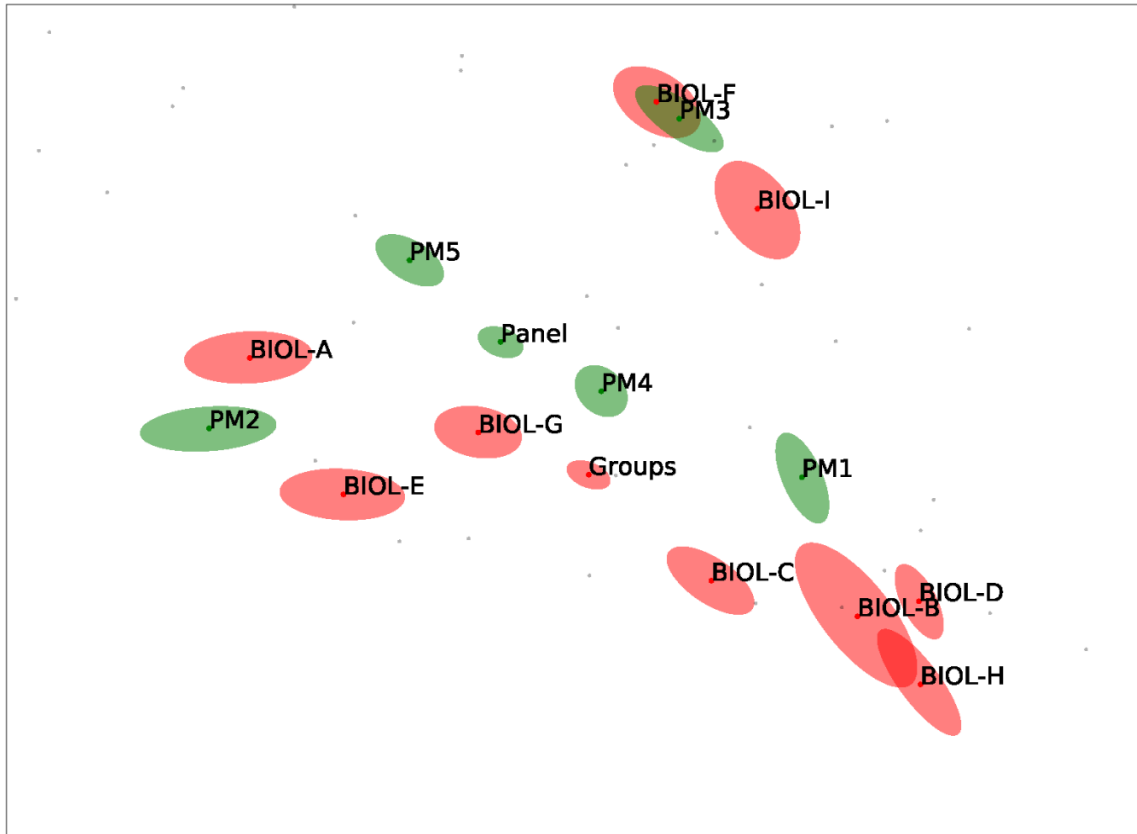
In Table 3, for each research group we find the shortest distance to one of the panel members, and underline and bold it. In addition, the average and standard deviation of the shortest distances are calculated. The confidence intervals (discussed in section 2.5) are included through the typography of the values.

### ***c) Barycenter overlay map***

We take the ‘WoS SC\_VOS\_map.xlsx’ (Figure 10) file and manually input the Biology individual groups, panel members, research groups together and panel’s coordinates (shown in Figure 12) after the 224 WoS SCs. We fill up the ‘weight’ column with 20 (we can put other numbers too) to highlight the size of the bubble.

In the ‘cluster’ column, we assign 1 to all the 224 WoS SCs, 2 to the research groups together, 3 to all research groups, 4 to the panel, and 5 to individual panel members. We save the file as ‘Barycenter overlap map of Biology department.csv’. After that, we open the file with VOSviewer to visualize the barycenters (Figure 14). Figure 15 shows a zoomed in version of Figure 14. We also create the barycenter overlap map of Biology department and include the confidence regions of the respective barycenter of panel, panel members (PM), research groups and research groups together using the WoS SCs VOS\_map (Figure 16).





**Figure 16. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together (groups) with their confidence regions**

The bootstrap replications of barycenters are also used to add a 95% confidence region for each barycenter to the maps. For each barycenter we have a cloud of 1000 points (bootstrapped barycenters) surrounding it. The confidence region is an ellipse that covers 95% of the bootstrapped barycenters. The larger the confidence region, the less stable the barycenter is. Although the CI of the distance between two barycenters and their confidence regions are related, the two should not be conflated. In particular, we stress that overlapping confidence regions as seen in Figure 16 (figure with overlapping regions in it) does not correspond to overlap between CIs for distances.

The maps were plotted using Matplotlib (<http://matplotlib.org>). First, the base map was plotted using the pre-existing coordinates. Next, the barycenters were added as slightly larger red or green points. Finally, a partially translucent confidence region (ellipse) was calculated and superimposed on the map. Calculation of the confidence region was done using an implementation by Kington (2014). We briefly outline what elements determine the location and placement of such a confidence ellipse. The center of the ellipse is simply the mean of all

bootstrapped barycenters. The width and height of the ellipse (or its axes) depend on the variance in the cloud of points. Finally, the orientation of the ellipse is obtained from the largest eigenvector.

## 2.7 Similarity-adapted publication vector method

### a) Similarity-adapted publication vector calculation

A similarity-adapted publication vector (SAPV) is determined as the vector  $C = (C_1, C_2, \dots, C_N)$ , where:

$$C_k = \frac{\sum_{j=1}^N s_{kj} m_j}{\sum_{i=1}^N \sum_{j=1}^N s_{ij} m_j} \quad (5)$$

where  $s_{kj}$  denotes the similarity value between the  $k$ -th and the  $j$ -th WoS SC, and  $m_j$  is the number of publications in WoS SC  $j$ . The numerator of formula (5) is equal to the  $k$ -th element of  $S * M$ , the multiplication of the similarity matrix  $S$  and the column matrix of publications  $M = (m_j)_j$ . The denominator is the L1-norm of the unnormalized vector.

We take the ‘map10.net’ file (see section 2.3) and with a Python script, we transform the network back into the adjacency matrix and save it as ‘WoS SCs similarity matrix.xlsx’ (Figure 17).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Acoustics	al Economics	tural Engin	Dairy Anire,	Multidi	Agronomy	Allergy	my Morpl	Andrology	esthesiol	onthropolog	rea Studie	omy Astro
2	Acoustics	1	0	0	0	0	0	0	0	0	0	0	0	0
3	Agricultural Economics Policy	0	1	0	0	0	0	0	0	0	0	0	0.185	0
4	Agricultural Engineering	0	0	1	0.16	0.445	0.319	0	0	0	0	0	0	0
5	Agriculture, Dairy Animal Science	0	0	0.16	1	0.413	0	0	0.184	0.247	0	0	0	0
6	Agriculture, Multidisciplinary	0	0	0.445	0.413	1	0.615	0	0.159	0	0	0	0	0
7	Agronomy	0	0	0.319	0	0.615	1	0	0	0	0	0	0	0
8	Allergy	0	0	0	0	0	0	1	0	0	0	0	0	0
9	Anatomy Morphology	0	0	0	0.184	0.159	0	0	1	0.478	0.23	0	0	0
10	Andrology	0	0	0	0.247	0	0	0	0.478	1	0	0	0	0
11	Anesthesiology	0	0	0	0	0	0	0	0.23	0	1	0	0	0
12	Anthropology	0	0	0	0	0	0	0	0	0	0	1	0.277	0
13	Area Studies	0	0.185	0	0	0	0	0	0	0	0	0.277	1	0
14	Astronomy Astrophysics	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 17. Excerpt of WoS SCs similarity matrix

A python script ‘sa-vector-categories.py’ is used that takes as input the WoS SCs similarity matrix (Figure 17) and the number of publications of Biology individual research groups and panel members per WoS SC (weights) (Figure 1), and calculates SAPVs for all entities. The

calculation of SAPVs is carried out by the sa\_vector function, (see section 5). We run the program as:

```
python sa-vector-categories.py "WoS SC_similarity matrix.xlsx" "Biology research groups and Panel_WoS SCs.xlsx"
```

This program calculates the SAPV of each entity and stores the result in an output file named ‘Biology research groups and panel\_WoS SCs-sa-vectors.xlsx’ (Figure 18).

	A	B	C	D	E	F	G	H	I	J	K	
1		<b>Acoustics</b>	<b>Agriculture</b>	<b>Agriculture</b>	<b>Agriculture</b>	<b>Agriculture</b>	<b>Agriculture</b>	<b>Agronomy</b>	<b>Allergy</b>	<b>Anatomy</b>	<b>Andrology</b>	<b>Anesthesiology</b>
2	<b>BIOL-A</b>	8.01E-05	0	0.003924	0.004305	0.008805	0.009057	0.001766	0.022889	0.010256	0.004448	0.000299
3	<b>BIOL-B</b>	0	0	0.012559	0.001433	0.013802	0.018159	0.001355	0.010492	0.003112	0.00029	0.00029
4	<b>BIOL-C</b>	0	0	0.017456	0.00334	0.013315	0.010276	0.002474	0.010659	0.007656	0.000595	0.000595
5	<b>BIOL-D</b>	0	0	0.018461	0.001336	0.020314	0.02775	0.000666	0.008696	0.00344	0.000283	0.000283
6	<b>BIOL-E</b>	0.000183	0	0.00755	0.004103	0.0098	0.010133	0.001393	0.019045	0.007156	0.003878	0.003878
7	<b>BIOL-F</b>	0	0	0.010468	0.004111	0.015662	0.019589	0.003959	0.019633	0.012622	0.000301	0.000301
8	<b>BIOL-G</b>	0.000166	9.31E-05	0.008589	0.004103	0.011867	0.013052	0.003305	0.016774	0.007495	0.001103	0.001103
9	<b>BIOL-H</b>	0.000194	0	0.016255	0.00081	0.014438	0.016931	0.001213	0.008149	0.001976	0.000195	0.000195
10	<b>BIOL-I</b>	8.55E-05	6.6E-05	0.012737	0.004013	0.017674	0.021527	0.002676	0.017229	0.010937	0.00033	0.00033
11	<b>Groups</b>	9.79E-05	2.71E-05	0.01116	0.003492	0.013056	0.014387	0.002174	0.015673	0.007514	0.001704	0.001704
12	<b>PM1</b>	5.43E-05	0	0.011903	0.001578	0.016602	0.023309	0.001424	0.014427	0.006772	0.000996	0.000996
13	<b>PM2</b>	0	0	0.004527	0.004897	0.008478	0.008598	0.002499	0.021574	0.008268	0.004864	0.004864
14	<b>PM3</b>	0	0	0.009165	0.004402	0.013523	0.01507	0.002859	0.020271	0.012944	0.000508	0.000508
15	<b>PM4</b>	0	0	0.009634	0.002399	0.014155	0.016523	0.001282	0.01918	0.005692	0.000852	0.000852
16	<b>PM5</b>	5.33E-05	0	0.005823	0.004909	0.008492	0.005272	0.00335	0.020981	0.013668	0.003585	0.003585
17	<b>Panel</b>	2.74E-05	0	0.007526	0.003926	0.011271	0.011721	0.002487	0.019727	0.010145	0.002615	0.002615

Figure 18. Excerpt of SAPV of the Biology individual research groups, panel members, research groups together and panel using WoS SCs similarity matrix

### b) Euclidean distance between similarity-adapted publication vectors

Subsequently, we determine the Euclidean distances between different entities SAPV: individual research groups, research groups together, panel members, and panel. The Euclidean distance between vectors  $a$  and  $b$  in  $\mathbf{R}^N$  is:

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + \dots + (a_N - b_N)^2} \quad (6)$$

Again, we use the implementation of Euclidean distance in scipy.spatial.dist. We note that the python script ‘sa-vector-categories.py’ executes both formula (5) and (6), and calculates Euclidean distances between the SAPV in an output file ‘Biology research groups and panel\_WoS SCs-sa-vectors.xlsx’ (Figure 19).

	A	B	C	D	E	F	G	H	I	J	K	L
1		<b>BIOL-A</b>	<b>BIOL-B</b>	<b>BIOL-C</b>	<b>BIOL-D</b>	<b>BIOL-E</b>	<b>BIOL-F</b>	<b>BIOL-G</b>	<b>BIOL-H</b>	<b>BIOL-I</b>	<b>Groups</b>	<b>PM1</b>
2	<b>BIOL-A</b>	0	0.093105	0.070788	0.08881	0.02996	0.060302	0.03688	0.105507	0.063177	0.038572	0.057301
3	<b>BIOL-B</b>	0.093105	0	0.076214	0.038223	0.071395	0.122019	0.061945	0.023583	0.117828	0.062402	0.05241
4	<b>BIOL-C</b>	0.070788	0.076214	0	0.065834	0.060321	0.079276	0.057407	0.075467	0.074093	0.042048	0.058939
5	<b>BIOL-D</b>	0.08881	0.038223	0.065834	0	0.070464	0.105469	0.060033	0.045451	0.09883	0.054258	0.036492
6	<b>BIOL-E</b>	0.02996	0.071395	0.060321	0.070464	0	0.079376	0.02405	0.082405	0.078766	0.02696	0.046441
7	<b>BIOL-F</b>	0.060302	0.122019	0.079276	0.105469	0.079376	0	0.074336	0.132512	0.015265	0.066553	0.075582
8	<b>BIOL-G</b>	0.03688	0.061945	0.057407	0.060033	0.02405	0.074336	0	0.074863	0.073095	0.018969	0.032619
9	<b>BIOL-H</b>	0.105507	0.023583	0.075467	0.045451	0.082405	0.132512	0.074863	0	0.127639	0.0725	0.066176
10	<b>BIOL-I</b>	0.063177	0.117828	0.074093	0.09883	0.078766	0.015265	0.073095	0.127639	0	0.063606	0.070922
11	<b>Groups</b>	0.038572	0.062402	0.042048	0.054258	0.02696	0.066553	0.018969	0.0725	0.063606	0	0.02846
12	<b>PM1</b>	0.057301	0.05241	0.058939	0.036492	0.046441	0.075582	0.032619	0.066176	0.070922	0.02846	0
13	<b>PM2</b>	0.016547	0.091207	0.072987	0.088713	0.022664	0.070668	0.037513	0.103511	0.072519	0.040618	0.059815
14	<b>PM3</b>	0.062346	0.129389	0.081992	0.11416	0.083045	0.015444	0.080481	0.139203	0.02338	0.072301	0.084459
15	<b>PM4</b>	0.048338	0.06974	0.078645	0.067385	0.041335	0.082423	0.028031	0.084645	0.081366	0.041149	0.04025
16	<b>PM5</b>	0.038966	0.111364	0.065298	0.105145	0.057781	0.045829	0.060783	0.119727	0.052201	0.054268	0.076296
17	<b>Panel</b>	0.019933	0.088623	0.058405	0.08096	0.03465	0.046912	0.033717	0.099907	0.048846	0.029541	0.048954

**Figure 19. Excerpt of pairwise Euclidean distance matrix between SAPVs of Biology individual research groups, panel members, research groups together and panel together using WoS SCs similarity matrix**

From the calculated matrix of pairwise Euclidean distances between SAPVs of Biology groups, panel members, groups together, and panel together (Figure 18) we extract Table 4 containing only the distances between the research groups and groups together on the one hand and the panel and panel members on the other, for the convenience of analysis. In Table 4, for each research group we find the shortest distance to one of the panel members, and underline and bold it. In addition, the average and standard deviation of the shortest distances are calculated. The confidence intervals (discussed in section 2.5) are included through the typography of the values.

**Table 4. Euclidean distances between SAPVs of Biology individual groups, panel members, research groups together and panel in WoS SCs similarity matrix**

	<b>Groups</b>	<b>BIOL-A</b>	<b>BIOL-B</b>	<b>BIOL-C</b>	<b>BIOL-D</b>	<b>BIOL-E</b>	<b>BIOL-F</b>	<b>BIOL-G</b>	<b>BIOL-H</b>	<b>BIOL-I</b>
Panel	0.030	0.020	0.089	0.058	0.081	0.035	0.047	0.034	0.100	0.049
PM1	0.028	0.057	<b><u>0.052</u></b>	<b><u>0.059</u></b>	<b><u>0.036</u></b>	0.046	0.076	<b><u>0.033</u></b>	<b><u>0.066</u></b>	0.071
PM2	0.041	<b><u>0.017</u></b>	0.091	0.073	0.089	<b><u>0.023</u></b>	0.071	<b><u>0.038</u></b>	0.104	0.073
PM3	0.072	0.062	0.129	0.082	0.114	0.083	<b><u>0.015</u></b>	0.080	0.139	<b><u>0.023</u></b>
PM4	0.041	0.048	<b><u>0.070</u></b>	0.079	0.067	0.041	0.082	<b><u>0.028</u></b>	<b><u>0.085</u></b>	0.081
PM5	0.054	0.039	0.111	<b><u>0.065</u></b>	0.105	0.058	0.046	0.061	0.120	0.052

For each research group we determined the panel member at the shortest distance. Average of shortest distance is 0.035 (SD 0.019). The number in the row of this panel member is indicated in bold and underlined. Distances whose confidence intervals overlap with that of the shortest distance are in bold (same column).



### ***c) Similarity-adapted publication vector overlay map***

Results of the SAPV approach cannot be visualized easily since an SAPV has  $N$  coordinates itself. However, visualization is possible if one expands the similarity matrix with one extra row and column, containing the SAPV's coordinates. The expanded  $(N + 1) \times (N + 1)$  matrix can then be visualized using, for instance, VOSviewer. Note that this approach works well for visualizing the location of one SAPV but cannot be used for multiple SAPVs at the same time, for two reasons:

- Adding extra rows/columns affects the layout algorithm and may distort the original base map. The effect of one extra point turns out to be negligible.
- It is unclear what similarity score should be assigned to two SAPVs.

We determine SAPVs of all entities (Figure 18). We take the WoS SCs similarity matrix Excel file (Figure 17). We copy BIOL-B's SAPV and paste at the bottom row and last column of the matrix file, thereby expanding the matrix to dimensions  $(N + 1) \times (N + 1)$ . We save the file as 'BIOL-B\_similarity matrix.xlsx'. A python script 'excel2network.py' is used to convert '[Research group code]\_similarity matrix.xlsx' files to Pajek network files (which can then be used in Pajek or VOSviewer). We run the program as:

```
python excel2network.py "BIOL-B_similarity matrix.xlsx" Sheet1
```

This program yields an output file named 'BIOL-B\_similarity matrix.net'. We create a map based on the network file using VOSviewer. It is not possible to easily locate BIOL-B in the map due to many different cluster colors. Therefore, we save the map data as 'BIOL-B.txt' file. In the text file, we can identify BIOL-B, but cannot easily change cluster number of all the WoS SCs in the file that is necessary to highlight the BIOL-B's location in the overlay map. Therefore, we import the data from the 'BIOL-B.txt' file to 'BIOL-B.xlsx' file.

In the 'BIOL-B.xlsx' file we first identify the BIOL-B label and assigned 20 (we can put other numbers too) for weight. In the cluster column, we assign 1 to all the WoS SCs and 2 to BIOL-B and save as CSV file (Figure 20). We open the file with VOSviewer to visualize the SAPV 'location' of BIOL-B (Figure 21).

A	B	C	D	E	F
id	label	x	y	weight	cluster
140	Ornithology	-0.481	0.1537	7.6898	1
141	Instruments Instrumentation	-0.8893	-0.1957	20.632	1
142	Tropical Medicine	0.2373	0.1688	14.3475	1
143	Robotics	-0.6917	-0.2703	7.066	1
144	Computer Science, Software Engineering	-0.6678	-0.2769	9.337	1
145	Biotechnology Applied Microbiology	-0.3508	0.1678	29.5586	1
146	Social Work	1.2395	-0.0435	9.858	1
147	Ethnic Studies	1.3785	-0.0985	13.541	1
148	Pathology	0.0791	0.2086	25.1297	1
149	BIOL-B	-0.6212	0.1361	20	3
150	Agricultural Engineering	-0.6926	0.086	15.7206	1
151	Endocrinology Metabolism	0.0472	0.2006	23.1127	1
152	Management	1.2979	-0.1727	8.864	1
153	Medicine, Research Experimental	0.0113	0.1972	35.3388	1
154	Materials Science, Composites	-0.9678	-0.1482	12.636	1
155	Respiratory System	0.1722	0.2091	15.4521	1
156	Materials Science, Multidisciplinary	-0.9875	-0.1297	21.304	1

Figure 20. Excerpt of BIOL-B.csv file

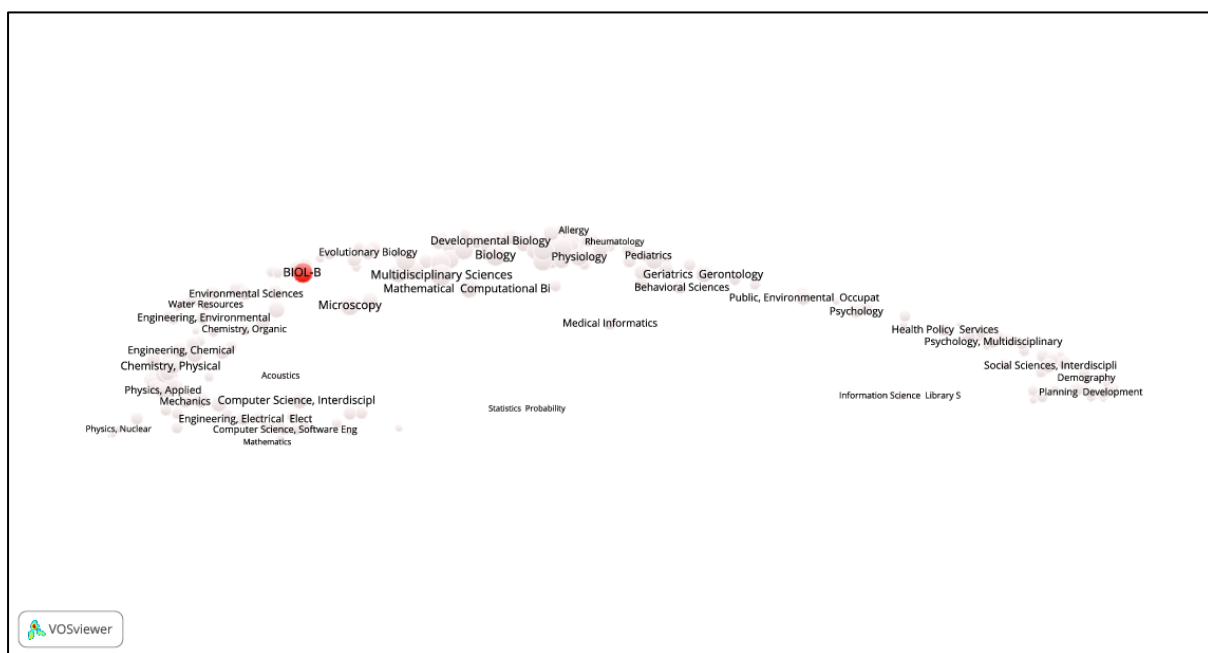


Figure 21. Location of the SAPV of BIOL-B in the WoS SCs similarity matrix

We repeat the above-mentioned process to create separate maps for each research group, each panel member, research groups together and panel (see Appendix B).

## 2.8 Weighted cosine similarity method

We consider a weighted similarity method (generalized cosine similarity). The weighted similarity between panel member (PM)  $k$  and research group  $m$ , according to Zhou et al. (2012) is:

$$\begin{aligned} & \frac{\sum_{i=1}^N M_i^k (\sum_{j=1}^N R_j^m S_{ji})}{\sqrt{\left(\sum_{i=1}^N M_i^k (\sum_{j=1}^N M_j^k S_{ji})\right) \cdot \left(\sum_{i=1}^N R_i^m (\sum_{j=1}^N R_j^m S_{ji})\right)}} \\ &= \frac{(M^k)^t * S * R^m}{\sqrt{(M^k)^t * S * M^k} \cdot \sqrt{(R^m)^t * S * R^m}} \end{aligned} \quad (7)$$

The numerator is nothing but the matrix multiplication:  $(M^k)^t * S * R^m$ , where  $^t$  denotes matrix transposition,  $S$  is the WoS SCs similarity matrix,  $M^k$  denotes the column matrix of publications of panel member  $k$  and  $R^m$  denotes the column matrix of publications of research group  $m$ . Similarly, the two products under the square root in the denominator are:  $(M^k)^t * S * M^k$  and  $(R^m)^t * S * R^m$ . The result is the similarity between panel member  $k$  and research group  $m$ .

This value is calculated for each panel member and each research group. Weighted cosine similarity (WCS) is implemented in Python as a fairly straightforward set of matrix operations (see section 5, `weighted_cosine`). A python script ‘`cosine-categories.py`’ is used that takes as input the similarity matrix (‘`WoS SCs_similarity matrix.xlsx`’, see Figure 17) and the weights (number of publications) per WoS SC (‘`Biology research groups and panel_WoS SCs.xlsx`’, see Figure 1), and calculates the weighted cosine similarity between all entities. We run the program as:

```
python cosine-categories.py "WoS SC_similarity matrix.xlsx" "Biology research groups and panel_WoS SCs.xlsx"
```

This program calculates the WCS value in an output file as ‘`Biology research groups and panel_WoS SCs-cosine.xlsx`’ (Figure 22).

	A	B	C	D	E	F	G	H	I	J	K
1		<b>BIOL-A</b>	<b>BIOL-B</b>	<b>BIOL-C</b>	<b>BIOL-D</b>	<b>BIOL-E</b>	<b>BIOL-F</b>	<b>BIOL-G</b>	<b>BIOL-H</b>	<b>BIOL-I</b>	<b>Groups</b>
2	<b>BIOL-A</b>	1	0.695992	0.560737	0.630906	0.94551	0.634034	0.924108	0.606112	0.600162	0.894697
3	<b>BIOL-B</b>	0.695992	1	0.714285	0.89077	0.797504	0.436381	0.867361	0.963794	0.455074	0.889908
4	<b>BIOL-C</b>	0.560737	0.714285	1	0.708582	0.678423	0.505577	0.699997	0.789864	0.527651	0.810478
5	<b>BIOL-D</b>	0.630906	0.89077	0.708582	1	0.719653	0.571878	0.799796	0.860889	0.619949	0.871254
6	<b>BIOL-E</b>	0.94551	0.797504	0.678423	0.719653	1	0.506227	0.955667	0.741748	0.490989	0.93715
7	<b>BIOL-F</b>	0.634034	0.436381	0.505577	0.571878	0.506227	1	0.578359	0.362017	0.98136	0.671642
8	<b>BIOL-G</b>	0.924108	0.867361	0.699997	0.799796	0.955667	0.578359	1	0.802422	0.570234	0.969613
9	<b>BIOL-H</b>	0.606112	0.963794	0.789864	0.860889	0.741748	0.362017	0.802422	1	0.383748	0.850261
10	<b>BIOL-I</b>	0.600162	0.455074	0.527651	0.619949	0.490989	0.98136	0.570234	0.383748	1	0.675636
11	<b>Groups</b>	0.894697	0.889908	0.810478	0.871254	0.93715	0.671642	0.969613	0.850261	0.675636	1
12	<b>PM1</b>	0.780407	0.889291	0.673629	0.94754	0.804355	0.722736	0.886293	0.816899	0.741496	0.930619
13	<b>PM2</b>	0.969328	0.685816	0.539724	0.607378	0.972378	0.545189	0.909806	0.597332	0.514308	0.872733
14	<b>PM3</b>	0.638791	0.34953	0.47242	0.466661	0.488732	0.976675	0.538284	0.281627	0.943865	0.62467
15	<b>PM4</b>	0.863798	0.773161	0.551596	0.73048	0.865505	0.561967	0.927933	0.689041	0.548273	0.876224
16	<b>PM5</b>	0.813544	0.537682	0.683224	0.457739	0.746211	0.738721	0.72345	0.533301	0.670274	0.772988
17	<b>Panel</b>	0.948055	0.748543	0.684937	0.718485	0.911234	0.782138	0.926288	0.68328	0.747396	0.941503

**Figure 22. Excerpt of WCS value matrix of the Biology individual research groups, panel members, research groups together and panel using WoS SCs similarity matrix**

From the calculated WCS value matrix (Figure 22), we extract Table 5 containing only the WCS value of the research groups and groups on the one hand and the panel and panel members on the other, for the convenience of analysis. The confidence intervals (discussed in section 2.5) are included through the typography of the values.

Since our barycenter method (see section 2.6) and SAPV method (see section 2.7) are distance-based rather than similarity-based, we use  $1 - \text{WCS}$  as values to obtain dissimilarity values: weighted cosine dissimilarity (WCD) in Table 6, which can more easily be compared with the other two approaches. For the sake of simplicity, the results are shown under the WCS method.

We calculate the Pearson's correlation coefficient ( $r$ ) and the Spearman rank-order correlation coefficient ( $\rho$ ) between the three methods: barycenter, SAPV and WCS. These calculations are based on all Euclidean distances between barycenter and SAPV of individual research groups and panel members, and WCS value of individual research groups and panel members only. Although there are co-publications between groups, the barycenter distances between panel and combined group and separate groups, and combined groups and individual panel member can be (or at least are) considered independent, and have been included in the correlation calculation. The results are shown in Table 7.

**Table 5. WCS value of Biology individual research groups, panel members, research groups together and panel using WoS SCs similarity matrix**

	Groups	BIOL-A	BIOL-B	BIOL-C	BIOL-D	BIOL-E	BIOL-F	BIOL-G	BIOL-H	BIOL-I
Panel	0.942	0.948	0.749	0.685	0.718	0.911	0.782	0.926	0.683	0.747
PM1	0.931	0.780	<b><u>0.889</u></b>	<b><u>0.674</u></b>	<b><u>0.948</u></b>	0.804	0.723	<b><u>0.886</u></b>	<b><u>0.817</u></b>	0.741
PM2	0.873	<b><u>0.969</u></b>	0.686	0.540	0.607	<b><u>0.972</u></b>	0.545	<b><u>0.910</u></b>	0.597	0.514
PM3	0.625	0.639	0.350	0.472	0.467	0.489	<b><u>0.977</u></b>	0.538	0.282	<b><u>0.944</u></b>
PM4	0.876	0.864	<b><u>0.773</u></b>	0.552	0.730	0.866	0.562	<b><u>0.928</u></b>	0.689	0.548
PM5	0.773	0.814	0.538	<b><u>0.683</u></b>	0.458	0.746	0.739	0.723	0.533	0.670

For each research group we determine the panel member at the highest similarity. The number in the row corresponding to this panel member is indicated in bold and underlined. Similarities whose confidence intervals overlap with that of the highest similarities are in bold (same column).

**Table 6. WCD value between Biology individual research groups, panel members, groups and panel using WoS SCs similarity matrix**

	Groups	BIOL-A	BIOL-B	BIOL-C	BIOL-D	BIOL-E	BIOL-F	BIOL-G	BIOL-H	BIOL-I
Panel	0.058	0.052	0.251	0.315	0.282	0.089	0.218	0.074	0.317	0.253
PM1	0.069	0.220	<b><u>0.111</u></b>	<b><u>0.326</u></b>	0.052	0.196	0.277	<b><u>0.114</u></b>	<b><u>0.183</u></b>	0.259
PM2	0.127	<b><u>0.031</u></b>	0.314	0.460	0.393	<b><u>0.028</u></b>	0.455	<b><u>0.090</u></b>	0.403	0.486
PM3	0.375	0.361	0.650	0.528	0.533	0.511	<b><u>0.023</u></b>	0.462	0.718	<b><u>0.056</u></b>
PM4	0.124	0.136	<b><u>0.227</u></b>	0.448	0.270	0.134	0.438	<b><u>0.072</u></b>	0.311	0.452
PM5	0.227	0.186	0.462	<b><u>0.317</u></b>	<b><u>0.542</u></b>	0.254	0.261	0.277	0.467	0.330

The lowest similarity between a group and a panel member is underlined and printed in bold.

In Table 7, the upper triangle refers to Pearson's correlations while the lower triangle refers to Spearman correlations. Table 7 and Figure 23 show there is moderate correlation between barycenter in one hand and SAPV and WCS on the other, while the correlation is strong between SAPV and WCS methods.

**Table 7. Pearson and Spearman correlation between three methods using data from Biology individual research groups and panel members**

	Pearson	Barycenter	SAPV	WCS
Spearman				
Barycenter		1	0.69	0.54
SAPV		0.67	1	0.93
WCS		0.61	0.90	1

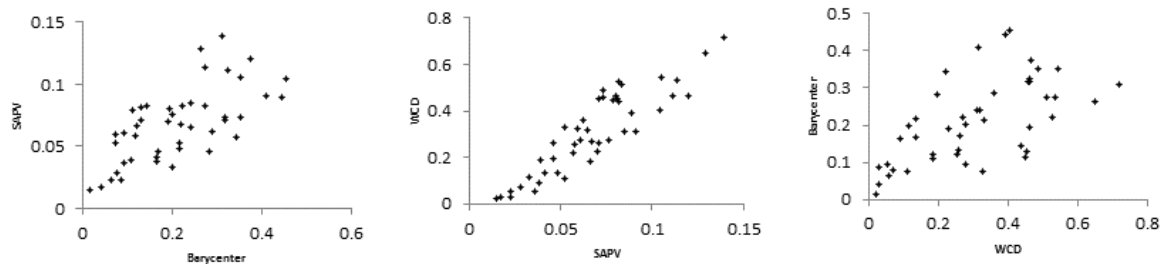


Figure 23. Scatter plot of the correlation between barycenter, SAPV and WCS

### 3 Cognitive distance based on journals

#### 3.1 Data collection process

For collecting journal data, after the search result (see section 2.1) we use the ‘Analyze Results’ option in the WoS, and rank the record by Source title (hereafter journal title) and set the minimum record count (threshold) to one. We repeat this procedure for each of the research groups and panel members. We save the record as ‘analyze.txt’ and subsequently rename the file to ‘[Research group code]\_ journals title.txt’, for example ‘BIOL-B\_journals title.txt’. For panel members we rename to ‘[Panel member code]\_ journals title.txt’, for example ‘PM2\_ journals title.txt’.

	A	B	C
1	<b>Source Titles</b>	<b>records</b>	<b>% of 1156</b>
2	ENVIRONMENTAL POLLUTION	40	3.469
3	BIOLOGICAL JOURNAL OF THE LINNEAN SOCIETY	33	2.862
4	JOURNAL OF EXPERIMENTAL BIOLOGY	26	2.255
5	AQUATIC TOXICOLOGY	23	1.995
6	ENVIRONMENTAL SCIENCE TECHNOLOGY	22	1.908
7	JOURNAL OF EVOLUTIONARY BIOLOGY	19	1.648
8	ENVIRONMENTAL TOXICOLOGY AND CHEMISTRY	19	1.648
9	NEW PHYTOLOGIST	16	1.388
10	BEHAVIORAL ECOLOGY	16	1.388
11	GLOBAL CHANGE BIOLOGY	15	1.301
12	ENVIRONMENT INTERNATIONAL	15	1.301
13	TREE PHYSIOLOGY	14	1.214
14	FUNCTIONAL ECOLOGY	14	1.214
15	SCIENCE OF THE TOTAL ENVIRONMENT	13	1.127

Navigation: Groups BIOL A BIOL B BIOL C BIOL D BIOL E BIOL F BIOL G BIOL H BIOL I

Figure 24. Excerpt of Biology research groups and panel\_journal title.xlsx file

We combine the search sets for each research group and panel member from the search history of the WoS, and get the result for the research groups as a whole and the panel. In this way, any publication that has been co-authored by members of two or more research groups or by two or more panel members is counted only once. We save the resulting list as 'analyze.txt' and save a copy of the file as 'Groups together\_journals titles.txt' for the groups as a whole, and as 'Panel\_journals title.txt' for the panel.

All downloaded data files are exported to an MS Excel file. The downloaded data files, '[Research group code]\_journals title.txt', '[PM code]\_journals title.txt', 'Groups together\_journals title.txt' and 'Panel\_journals title.txt' have been exported to an MS Excel file. The sheets in the Excel file contain data on and are named after the research groups' code names (BIOL-A, BIOL-B, BIOL-C, etc.), the panel members' code names, (PM1, PM2, PM3, etc.), Panel members together and Groups together. The Excel file is saved as 'Biology research groups and panel\_journals title.xlsx' (Figure 24).

Publication statistics for each research groups and panel members have shown in the Table 1 and Table 2 respectively.

### **3.2 Correlation between publication profiles of research groups together and panel**

#### ***a) Pearson's correlation coefficient and Spearman's rank-order correlation coefficient***

We determine the correlation between the publication output of groups and panel, using Pearson's correlation coefficient and Spearman's rank-order correlation coefficient for the numbers of publications per journal. We make an Excel file 'Biology panel and groups together\_journals title.xlsx' and export data from 'Panel\_journals titles. txt' and 'Groups together\_journals title.txt' in two different sheets (Figure 25).

We reuse the Python script 'join-sheets.py' (see section 2.2) to take the data of the two sheets and join it into one. We run the program as:

```
python join-sheets.py "Biology panel and research groups together_journals title.xlsx"
```

It produces a new Excel file named ‘Biology panel and research groups together\_journals title-joined.xlsx’ (Figure 26). To calculate correlation, the value zero was kept on the corresponding journals in which either the panel or the groups had no publications (but not both). We calculate correlation using SPSS and obtain value ( $r = 0.26$ ,  $\rho = -0.28$ ). A log-log plot of the number of publications per journal for the Biology panel and research groups together is shown in Figure 27.

	A	B	C
1	<b>Source Titles</b>	<b>records</b>	<b>% of 1156</b>
2	ENVIRONMENTAL POLLUTION	40	3.469
3	BIOLOGICAL JOURNAL OF THE LINNEAN SC	33	2.862
4	JOURNAL OF EXPERIMENTAL BIOLOGY	26	2.255
5	AQUATIC TOXICOLOGY	23	1.995
6	ENVIRONMENTAL SCIENCE TECHNOLOGY	22	1.908
7	JOURNAL OF EVOLUTIONARY BIOLOGY	19	1.648
8	ENVIRONMENTAL TOXICOLOGY AND CHEM	19	1.648
9	NEW PHYTOLOGIST	16	1.388
10	BEHAVIORAL ECOLOGY	16	1.388
11	GLOBAL CHANGE BIOLOGY	15	1.301
12	ENVIRONMENT INTERNATIONAL	15	1.301
13	TREE PHYSIOLOGY	14	1.214
14	FUNCTIONAL ECOLOGY	14	1.214
15	SCIENCE OF THE TOTAL ENVIRONMENT	13	1.127
16	PHYSIOLOGIA PLANTARUM	13	1.127
17	JOURNAL OF ANATOMY	13	1.127
18	ANIMAL BIOLOGY	13	1.127
19	OECOLOGIA	12	1.041
20	HYDROBIOLOGIA	12	1.041

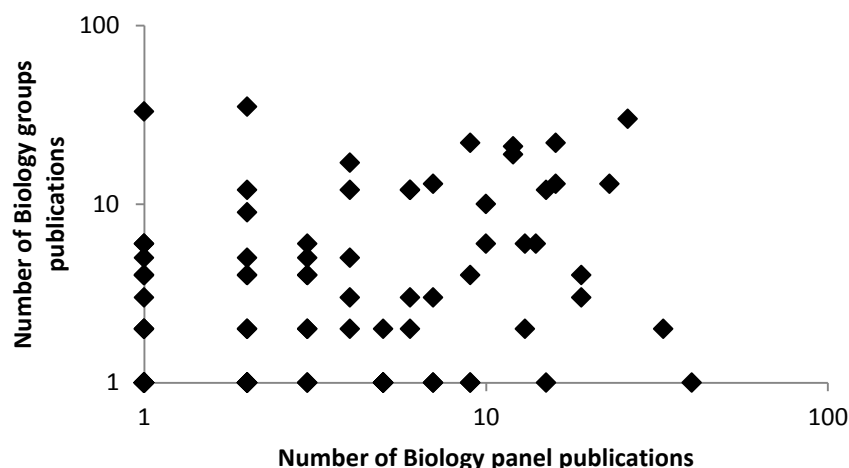
	A	B	C
1	<b>Source Titles</b>	<b>records</b>	<b>% of 792</b>
2	EXPERIMENTAL AND APPLIED ACAROLOG	35	4.453
3	GENERAL AND COMPARATIVE ENDOCRIN	33	4.198
4	JOURNAL OF EXPERIMENTAL BIOLOGY	30	3.817
5	PROCEEDINGS OF THE ROYAL SOCIETY B I	22	2.799
6	NEW PHYTOLOGIST	22	2.799
7	CELL AND TISSUE RESEARCH	22	2.799
8	OECOLOGIA	21	2.672
9	FISH PHYSIOLOGY AND BIOCHEMISTRY	19	2.417
10	ANIMAL BEHAVIOUR	19	2.417
11	EXPERIMENTAL APPLIED ACAROLOGY	18	2.29
12	NATURE	17	2.163
13	OIKOS	13	1.654
14	BEHAVIORAL ECOLOGY	13	1.654
15	AQUATIC TOXICOLOGY	13	1.654
16	JOURNAL OF FISH BIOLOGY	12	1.527
17	JOURNAL OF ENDOCRINOLOGY	12	1.527
18	GLOBAL CHANGE BIOLOGY	12	1.527
19	ENTOMOLOGIA EXPERIMENTALIS ET APPI	12	1.527
20	ANNALS OF BOTANY	12	1.527

Figure 25. Excerpt of the Biology panel and research groups together\_journals title.xlsx file

	A	B	C	D
1		<b>Source Titles</b>	<b>records_x</b>	<b>records_y</b>
2	<b>0</b>	ENVIRONMENTAL POLLUTION	40	1
3	<b>1</b>	BIOLOGICAL JOURNAL OF THE LINNEAN SOCIETY	33	2
4	<b>2</b>	JOURNAL OF EXPERIMENTAL BIOLOGY	26	30
5	<b>3</b>	AQUATIC TOXICOLOGY	23	13
6	<b>4</b>	ENVIRONMENTAL SCIENCE TECHNOLOGY	22	0
7	<b>5</b>	JOURNAL OF EVOLUTIONARY BIOLOGY	19	3
8	<b>6</b>	ENVIRONMENTAL TOXICOLOGY AND CHEMISTRY	19	4
9	<b>7</b>	NEW PHYTOLOGIST	16	22
10	<b>8</b>	BEHAVIORAL ECOLOGY	16	13
11	<b>9</b>	GLOBAL CHANGE BIOLOGY	15	12
12	<b>10</b>	ENVIRONMENT INTERNATIONAL	15	1
13	<b>11</b>	TREE PHYSIOLOGY	14	0
14	<b>12</b>	FUNCTIONAL ECOLOGY	14	6
15	<b>13</b>	SCIENCE OF THE TOTAL ENVIRONMENT	13	0
16	<b>14</b>	PHYSIOLOGIA PLANTARUM	13	6
17	<b>15</b>	JOURNAL OF ANATOMY	13	0
18	<b>16</b>	ANIMAL BIOLOGY	13	2
19	<b>17</b>	OECOLOGIA	12	21

Figure 26. Excerpt of the Biology panel and research groups together\_journals title-joined.xlsx file





**Figure 27. Log-log plot of the number of publications (log-log scale) per journals for the panel (horizontal axis) and research groups together (vertical axis) of the Biology department**

***b) Top-Down correlation coefficient***

In some cases, the panel has published in a journal where the research groups have not or vice versa , i.e. there are many zeroes on both sides. Since traditional correlation coefficients like Pearson’s and Spearman’s are not well-suited to zero-inflated data (i.e., data with a large amounts of zeroes), we adopt the top-down correlation coefficient (Iman & Conover, 1987). This correlation coefficient was found to be an adequate rank correlation coefficient for zero-inflated data (Huson, 2007). For a full description of the top-down correlation coefficient we refer to Iman and Conover (1987). This coefficient places emphasis on the higher ranked data by computing the correlation using Savage scores derived from the ranked data.

We reuse the formula 1 and 2 (details in section 2.2b) and the python script “calc\_topdowncorr.py” (all core logic is in topdowncorr.py, see section 5). We reuse the ‘Biology research groups and panel\_journal title- joined.xlsx’ (Figure 26) file, but keep the zeros in the WoS SCs where neither the panel nor the research groups have publications. We run the program as:

We keep the program file with ‘Biology research groups and panel\_ journals title-joined.xlsx’ (Figure 26) and run the program as:

```
python calc_topdowncorr.py "Biology research groups and panel_ journals title-joined.xlsx"
```

The outcome shows that the top-down correlation between Biology research groups together and the panel based on the journals in which they publish is very low (0.15). We conclude that both Spearman and top-down correlation suggest that there is low correlation between the journal publication portfolios of panel and research groups together.

In our opinion, the correlations are an insufficient measure in this case, as the similarity of journals is not taken into account here. This is reminiscent of the way diversity is sometimes studied using only the dimensions of variety and balance. As discussed by Stirling (2007), the additional dimension of disparity – the opposite concept of similarity – is needed to provide a complete picture. Likewise, a comparison of publication profiles based on journals that does not consider journal similarity might yield distorted results.

### **3.3 Journal similarity matrix**

Journal similarity data were received as a NET file (file name cosine.net) from Loet Leydesdorff in the context of the joint paper (Rahman, Guns, Leydesdorff, & Engels, 2016). While we did not construct this similarity matrix ourselves, we briefly outline the main steps that were taken to create it. The data was harvested from Clarivate Analytics's (formerly Thomson Reuters') Journal Citation Reports (JCR) of the Science and Social Science Editions 2011. An aggregated journal-journal citation matrix of 10,675 journals<sup>1</sup> was constructed with a grand total of 35,295,459 citations over the entire matrix, which was subsequently normalized in the citing direction. The similarities between journals are calculated using the cosine similarity between their citing distributions respectively (see Leydesdorff, Rafols, & Chen (2013) for details). The resulting journal similarity matrix can be considered as an adjacency matrix, and thus is equivalent to a weighted network where similar journals are linked and link weights increase with similarity strength.

The size of the file 'cosine.net' is around 1 gigabyte. First, we compress the file using gzip to 'cosine.net.gz'. After compression, the file is 291 megabytes. Next, we use a Python script 'load\_ndim\_data.py' to produce a file 'matrix.h5', which contains the network's adjacency matrix and is used further on. We use the gzipped network file 'cosine.net.gz' as input and run:

---

<sup>1</sup> The Science and Social Science Editions 2011 contain 8281 and 2943 journals respectively. Of these journals, 549 are contained in both databases.

python load\_ndim\_data.py cosine.net.gz

This way, we store the adjacency matrix in HDF5 (Hierarchical Data Format version 5), which was found to be the most efficient way of storing the data in terms of speed and memory requirements. From [http://www.leydesdorff.net/journals11/citing\\_all.txt](http://www.leydesdorff.net/journals11/citing_all.txt), we download the journal VOS map and save it to a file named ‘Journal\_VOS\_map.xlsx’ (Figure 28).

	A	B	C	D	E	F
1	id	label description	x	y	normalized weight	cluster
2	1	4OR-A Quarterly Journal of Operations Research	0.6434	0.2171	0.01	1000
3	2	AAOHN JOURNAL	-0.0632	-0.2391	0.01	1000
4	3	AAPG BULLETIN	0.0281	0.6925	0.01	1000
5	4	AAPS Journal	-0.6656	0.1736	0.01	1000
6	5	AAPS PHARMSCITECH	-0.4438	0.3727	0.01	1000
7	6	AATCC REVIEW	0.1818	0.6869	0.01	1000
8	7	ABDOMINAL IMAGING	-0.6488	-0.3543	0.01	1000
9	8	ABHANDLUNGEN AUS DEM MATHEMATISCHEN SEMINAR DER UNIVERSITÄT ZÜRICH	0.9848	0.617	0.01	1000
10	9	Abstract and Applied Analysis	0.8847	0.595	0.01	1000
11	10	ACADEMIC EMERGENCY MEDICINE	-0.6345	-0.5654	0.01	1000

Figure 28. Excerpt of the journals VOS map data

### 3.4 Journal overlay map creation

During data collection, the resulting files were downloaded using the default name ‘analyze.txt’ (see section 3.1). We downloaded the ‘Analyze.exe’ program, as well as the file ‘citing.dbf’ from <http://www.leydesdorff.net/journals11>.

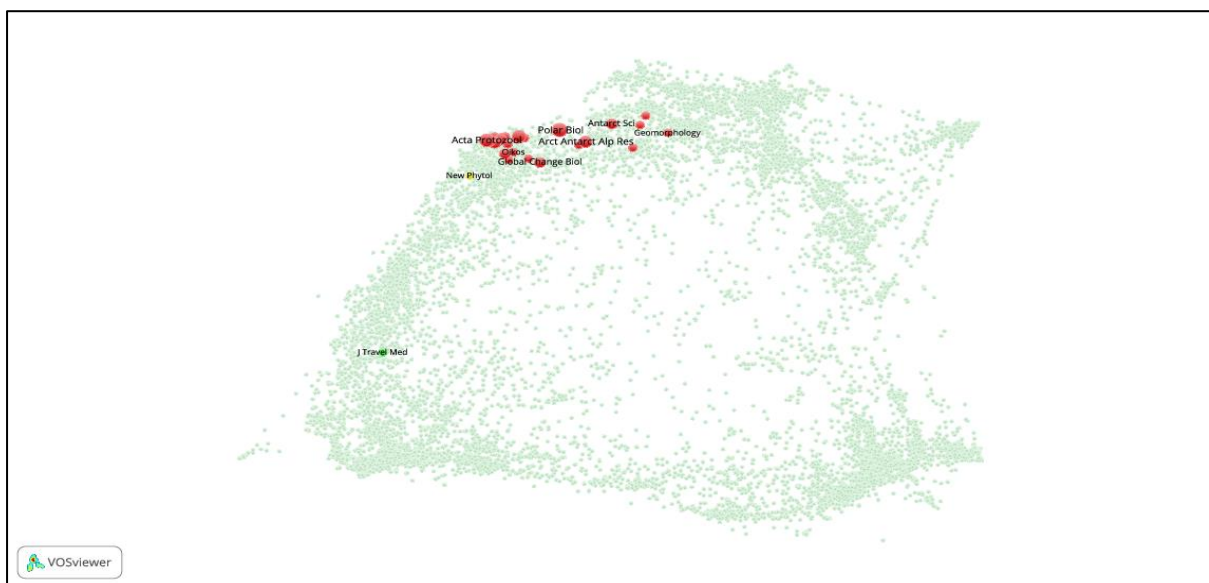


Figure 29. Journal overlay map of the BIOL-B research group

For each entity (Individual research groups, panel members, research groups together and panel), we save the corresponding ‘analyze.txt’ file in the concerned folder and run the program ‘Analyze.exe’. ‘Analyze.exe’ reads ‘analyze.txt’, and generates an output file ‘citing.txt’. We open the latter in VOSviewer to obtain an overlay map. For example, Figure 29 shows the journal overlay map of the BIOL-B research group.

We prepare separate journal overlay maps for each research group, each panel member, research groups together and panel (see Appendix C).

### 3.5 Barycenter method

#### *a) Barycenter calculation*

We recall the formula 3. The barycenter is defined as the point  $C = (C_1, C_2)$ , where

$$C_1 = \frac{\sum_{j=1}^N m_j L_{j,1}}{T} ; C_2 = \frac{\sum_{j=1}^N m_j L_{j,2}}{T}$$

Here,  $L_{j,1}$  and  $L_{j,2}$  are the horizontal and vertical coordinates of journal  $j$  on the map,  $m_j$  is the number of publications in journal  $j$ , and  $T = \sum_{j=1}^N m_j$  is the total number of publications of the entity.

Based on formula 3, Python script ‘journals-barycenter.py’ is used. This script takes ‘Journals\_VOS\_map.xlsx’ (Figure 28) and ‘Biology research groups and panel\_journals title.xlsx’ (Figure 24) as input. We run the program as follows

```
python journals-barycenter.py "Journals_VOS_map.xlsx" "Biology research groups and panel journals title.xlsx"
```

At this point, we notice that our program indicates that the journal titles in our Biology data do not match with the journal titles of the VOS map. We find that in the journal similarity matrix, the journal titles are written in short form while our downloaded data from WoS contains the full titles. In ‘citing.dbf’ (available at <http://www.leydesdorff.net/journals11>) both shortened and full titles are available. In addition, we have received 487 records from Loet Leydesdorff that were not included in the ‘citing.dbf’ file. Based on ‘citing.dbf’ and the additional data, we make a separate file ‘translation table.xlsx’ (Figure 30). We use the full title of the journals for matching.

	A	B
1	<b>CITEDJ</b>	<b>TITLE</b>
2	INDOGER FORSCH	INDOGERMANISCHE FORSCHUNGEN
3	SOCIETY	SOCIETY
4	ECOHEALTH	ECOHEALTH
5	GLASS PHYS CHEM	GLASS PHYSICS AND CHEMISTRY
6	ANN UROL	ANNALES D UROLOGIE
7	POPULATION	POPULATION
8	ADULT EDUC QUART	ADULT EDUCATION QUARTERLY
9	DEV DISABIL RES REV	DEVELOPMENTAL DISABILITIES RESEARCH REVIEWS
10	ANTIBIOTIQUES	ANTIBIOTIQUES
11	FUJITSU SCI TECH J	FUJITSU SCIENTIFIC & TECHNICAL JOURNAL
12	AGRARFORSCHUNG	AGRARFORSCHUNG

**Figure 30. Excerpt of short form to full journal titles**

We modify the program to accommodate the translation table. We rerun the program. This time our program indicates that there are some journals that do not match with any journal in the VOS map. This turns out to be due to name or organizational changes over time; indeed, journals are not static entities. More specifically, possible reasons are:

- The journal title is changed, shortened or extended;
- Two or more journals merge into a new journal;
- One journal splits into two or more new journals;
- A journal is excluded from the WoS, discontinued, or not listed during the construction of the aggregated journal-journal citation matrix.

We have developed the following guidelines to handle these uniformly:

- If journal A is renamed to B then treat both as equivalent.
- If journals A1 and A2 are merged into journal B, we treat both A1 and A2 as equivalent to B.
- If journal X splits into multiple journals, we look up which research groups or panel members have publications in journal X and determine which of the new journals best corresponds to the specialty of the authors, then change all occurrences of the journals in the WoS exported data with the best fitting latter journals.
- If a journal is discontinued or excluded from WoS, or not included in the aggregated journal-journal citation matrix and there is no equivalent for some other reason, then it is removed from the sample.

For each journal that is not found in the map, we search the title in the WoS and Journal Citation Reports, and consult its website as well as the ISSN database (www.issn.org) to identify the reasons behind the title change. Subsequently, based on the abovementioned guidelines we make a separate MS Excel file ‘Journal name change.xlsx’ (Figure 31) to translate ‘old’ titles to ‘correct’ titles.

We keep the ‘Biology research groups and panel\_journals title.xlsx’ (Figure 24), ‘Journals\_VOS\_map.xlsx’ (Figure 28), ‘translate.xlsx’ (Figure 30), and ‘Journal name change.xlsx’ (Figure 31) files in a folder. A modified Python script ‘journals-barycenter.py’ is used that takes the ‘Journal name change.xlsx’ file into account.

	A	B
1	<b>OLD TITLES</b>	<b>CORRECT TITLES</b>
159	JOURNAL OF PHYSICS E SCIENTIFIC INSTRUMENTS	MEASUREMENT SCIENCE AND TECHNOLOGY
160	ZEITSCHRIFT FUR PHYSIK C PARTICLES AND FIELDS	EUROPEAN PHYSICAL JOURNAL C
161	ZEITSCHRIFT FUR PHYSIK B CONDENSED MATTER	EUROPEAN PHYSICAL JOURNAL B
162	JOURNAL OF PHYSICS C SOLID STATE PHYSICS	JOURNAL OF PHYSICS: CONDENSED MATTER
163	JOURNAL OF PHYSICAL CHEMISTRY	JOURNAL OF PHYSICAL CHEMISTRY A
164	PHILOSOPHICAL MAGAZINE B PHYSICS OF CONDENSED MATTER STAT	PHILOSOPHICAL MAGAZINE
165	EUROPEAN PHYSICAL JOURNAL	EUROPEAN PHYSICAL JOURNAL A
166	CHEMISCHE BERICHTE RECUEIL	EUROPEAN JOURNAL OF ORGANIC CHEMISTRY
167	CHEMISCHE BERICHTE	EUROPEAN JOURNAL OF INORGANIC CHEMISTRY
168	CANADIAN JOURNAL OF APPLIED SPECTROSCOPY	CANADIAN JOURNAL OF ANALYTICAL SCIENCES AND SPECTROSCOPY
169	INORGANICA CHIMICA ACTA ARTICLES AND LETTERS	INORGANICA CHIMICA ACTA
170	AMERICAN JOURNAL OF MEDICAL GENETICS	AMERICAN JOURNAL OF MEDICAL GENETICS PART A
171	ANALYTICAL LETTERS PART A CHEMICAL ANALYSIS	ANALYTICAL LETTERS
172	JOURNAL OF THE SOUTH AFRICAN CHEMICAL INSTITUTE	SOUTH AFRICAN JOURNAL OF CHEMISTRY-SUID-AFRIKAANSE TYDSKRIF VIR CHEMIE
173	GEMS GEMOLOGY	GEMS & GEMOLOGY
174	BULLETIN DE LA SOCIETE CHIMIQUE DE FRANCE	EUROPEAN JOURNAL OF ORGANIC CHEMISTRY
175	INORGANICA CHIMICA ACTA LETTERS	INORGANICA CHIMICA ACTA
176	JOURNAL OF GEOPHYSICAL RESEARCH ATMOSPHERES	JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES
177	PHILOSOPHICAL MAGAZINE B PHYSICS OF CONDENSED MATTER STAT	PHILOSOPHICAL MAGAZINE
178	INORGANICA CHIMICA ACTA ARTICLES	INORGANICA CHIMICA ACTA

**Figure 31. Excerpt of journal name change.xlsx file**

We run the program as follows:

```
python journals-barycenter.py "Journals_VOS_map.xlsx" "Biology research groups and panel journals title.xlsx"
```

This program calculates the barycenter coordinates of Biology individual research groups, panel members, groups and panel in the journals VOS map in an output file ‘Biology research groups and panel\_journals title -barycenters.xlsx’ (Figure 32).

	A	B	C
1		<b>x</b>	<b>y</b>
2	<b>BIOL-A</b>	-0.38607	0.31135
3	<b>BIOL-B</b>	-0.37724	0.574245
4	<b>BIOL-C</b>	-0.23525	0.414805
5	<b>BIOL-D</b>	-0.33602	0.510718
6	<b>BIOL-E</b>	-0.29353	0.357211
7	<b>BIOL-F</b>	-0.63502	0.33927
8	<b>BIOL-G</b>	-0.42356	0.404341
9	<b>BIOL-H</b>	-0.30687	0.541683
10	<b>BIOL-I</b>	-0.58107	0.400881
11	<b>PM1</b>	-0.42822	0.459813
12	<b>PM2</b>	-0.37804	0.325685
13	<b>PM3</b>	-0.60792	0.330372
14	<b>PM4</b>	-0.50716	0.476186
15	<b>PM5</b>	-0.57265	0.2557
16	<b>Groups</b>	-0.37383	0.412116
17	<b>Panel</b>	-0.50152	0.366177

Figure 32. Barycenter coordinates of the Biology individual research groups, panel members, research groups together, and panel using journal VOS map

***b) Euclidean distance calculation between barycenters***

Subsequently, we determine the Euclidean distances between the barycenters of different entities: individual research groups, panel members, research groups together and panel. We reuse the formula 4 and determine the Euclidean distance between barycenters. We note that the python script ‘journals-categories.py’ executes both formula 3 and 4.

	A	B	C	D	E	F	G	H	I	J	K
1		<b>BIOL- A</b>	<b>BIOL- B</b>	<b>BIOL- C</b>	<b>BIOL- D</b>	<b>BIOL- E</b>	<b>BIOL- F</b>	<b>BIOL- G</b>	<b>BIOL- H</b>	<b>BIOL- I</b>	<b>PM1</b>
2	<b>BIOL- A</b>	0	0.263043	0.18289	0.205554	0.103287	0.250507	0.100264	0.24357	0.214566	0.15433
3	<b>BIOL- B</b>	0.263043	0	0.213493	0.075723	0.232618	0.348804	0.176105	0.077533	0.267585	0.125275
4	<b>BIOL- C</b>	0.18289	0.213493	0	0.139118	0.08193	0.406836	0.188596	0.145695	0.346091	0.198144
5	<b>BIOL- D</b>	0.205554	0.075723	0.139118	0	0.159282	0.344661	0.137762	0.04253	0.268532	0.105314
6	<b>BIOL- E</b>	0.103287	0.232618	0.08193	0.159282	0	0.341963	0.138313	0.184955	0.290838	0.169322
7	<b>BIOL- F</b>	0.250507	0.348804	0.406836	0.344661	0.341963	0	0.221243	0.385553	0.081894	0.239366
8	<b>BIOL- G</b>	0.100264	0.176105	0.188596	0.137762	0.138313	0.221243	0	0.18022	0.157544	0.055668
9	<b>BIOL- H</b>	0.24357	0.077533	0.145695	0.04253	0.184955	0.385553	0.18022	0	0.308234	0.146383
10	<b>BIOL- I</b>	0.214566	0.267585	0.346091	0.268532	0.290838	0.081894	0.157544	0.308234	0	0.163814
11	<b>PM1</b>	0.15433	0.125275	0.198144	0.105314	0.169322	0.239366	0.055668	0.146383	0.163814	0
12	<b>PM2</b>	0.016431	0.248561	0.168317	0.189744	0.090205	0.257334	0.090877	0.227421	0.216502	0.143207
13	<b>PM3</b>	0.222658	0.335689	0.382106	0.326266	0.315535	0.028525	0.198642	0.367805	0.075448	0.221464
14	<b>PM4</b>	0.204531	0.162773	0.278745	0.174582	0.244528	0.187334	0.110228	0.210724	0.105514	0.080619
15	<b>PM5</b>	0.194702	0.373709	0.373029	0.347888	0.297011	0.104277	0.210529	0.390417	0.145425	0.250046
16	<b>Groups</b>	0.101507	0.162164	0.1386	0.105601	0.09728	0.271156	0.050335	0.145846	0.207541	0.072341
17	<b>Panel</b>	0.12781	0.242363	0.270674	0.219732	0.208193	0.136177	0.086804	0.262092	0.086782	0.118918

Figure 33. Excerpt of Euclidean distances matrix of the barycenter of the Biology groups, panel members, research groups together and panel using the journal VOS map

From the matrix of Euclidean distances between all entity pairs (Figure 33), we extract Table 8 containing only distances between the research groups and panel on the one hand and the panel and panel members on the other, for the convenience of analysis.

In Table 8, for each research group we find the shortest distances to one of the panel members, and underline and bold it. In addition, the average and standard deviation of the shortest distances are calculated. The confidence intervals (discussed in section 2.5) are included through the typography of the values.

**Table 8. Euclidean distances between barycenter of Biology individual research groups, panel members, research groups together and panel using the journal VOS map**

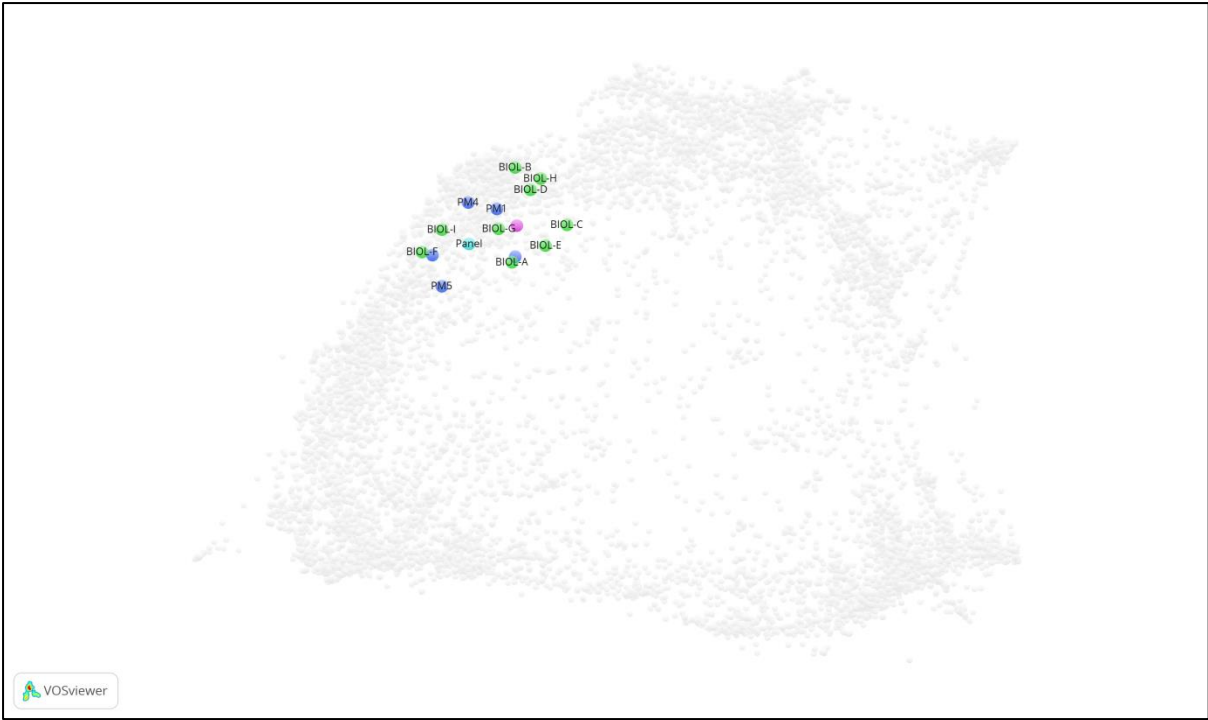
	Groups	BIOL- A	BIOL- B	BIOL-C	BIOL D	BIOL-E	BIOL -F	BIOL- G	BIOL- H	BIOL- I
Panel	0.136	0.128	0.242	0.271	0.220	0.208	0.136	0.087	0.262	0.087
PM1	0.072	0.154	<b><u>0.125</u></b>	<b><u>0.198</u></b>	<b><u>0.105</u></b>	0.169	0.239	<b><u>0.056</u></b>	<b><u>0.146</u></b>	<b><u>0.164</u></b>
PM2	0.087	<b><u>0.016</u></b>	0.249	<b><u>0.168</u></b>	0.190	<b><u>0.090</u></b>	0.257	<b><u>0.091</u></b>	<b><u>0.227</u></b>	0.217
PM3	0.248	0.223	0.336	0.382	0.326	0.316	<b><u>0.029</u></b>	0.199	0.368	<b><u>0.075</u></b>
PM4	0.148	0.205	<b><u>0.163</u></b>	0.279	0.175	0.245	0.187	<b><u>0.110</u></b>	<b><u>0.211</u></b>	<b><u>0.106</u></b>
PM5	0.253	0.195	0.374	0.373	0.348	0.297	<b><u>0.104</u></b>	0.211	0.390	<b><u>0.145</u></b>

For each research group we determine the panel member at the highest similarity. An average shortest distance is 0.090 (SD 0.051). The number in the row corresponding to this panel member is indicated in bold and underlined. Similarities whose confidence intervals overlap with that of the highest similarities are in bold (same column).

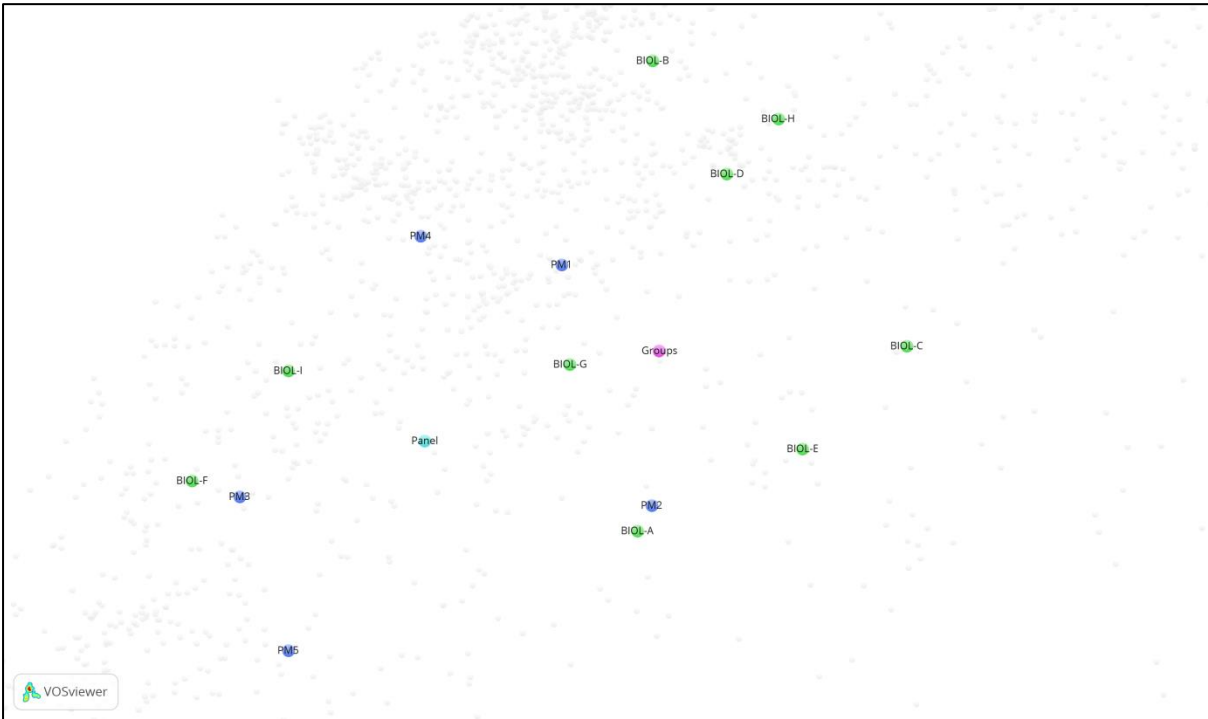
### ***c) Barycenter overlay map***

We take the Journal level\_VOS map (Figure 28) file and manually add the Biology individual groups, panel members, research groups together and panel's coordinates (Figure 32) after the 10,673 journals title. We fill up the 'weight' column with 20 (we can put other numbers too) to highlight the size of the bubble of the added entities. In the 'cluster' column, we assign 1 to all the 10,673 journals, 2 to the research groups together, 3 to all research groups, 4 to the panel, and 5 to individual panel members. We save the map file as 'Barycenter map of Biology department in the journal level.csv'. After that, we open the file with VOSviewer to visualize the barycenters (Figure 34). Figure 35 shows a zoomed in version of Figure 34.

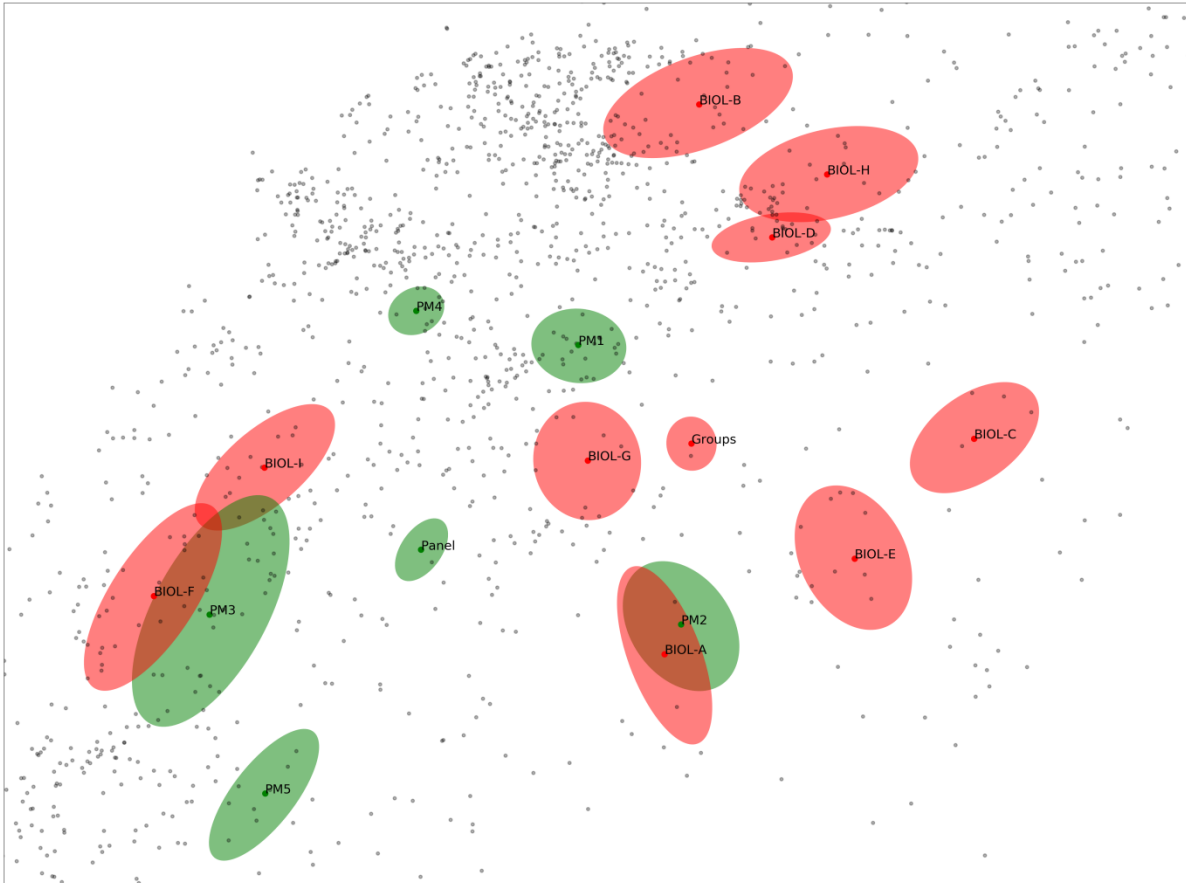




**Figure 34. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together**



**Figure 35. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together (zoomed)**



**Figure 36. Barycenter overlay map of Biology panel, panel members (PM), research groups and research groups together with their confidence regions**

We also create the barycenter overlap map of Biology department and include the confidence regions of the respective barycenter of panel, panel members (PM), research groups and research groups together using the journal VOS map (Figure 36). The bootstrap replications of barycenters are also used to add a 95% confidence region for each barycenter to the maps. In particular, we stress that overlapping confidence regions as seen in Figure 36 (figure with overlapping regions in it) does not correspond to overlap between CIs for distances. For detail process about confidence regions see section 2.6c.

### **3.6 Similarity-adapted publication vector method**

#### ***a) Similarity-adapted publication vector calculation***

Recall formula 5. A similarity-adapted publication vector is determined as the vector  $C = (C_1, C_2, \dots, C_N)$ , where:

$$C_k = \frac{\sum_{j=1}^N s_{kj} m_j}{\sum_{i=1}^N \sum_{j=1}^N s_{ij} m_j}$$

Here,  $s_{kj}$  denotes the similarity value between the  $k$ -th and the  $j$ -th journal, and  $m_j$  is the number of publications in journal  $j$ . The numerator of the formula is equal to the  $k$ -th element of  $S * M$ , the multiplication of the similarity matrix  $S$  and the column matrix of publications  $M = (m_j)_j$ . The denominator is the L1-norm of the unnormalized vector.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		4OR A QU	AAOHN JC	AAPG BUL	AAPS JOU	AAPS PHA	AATCC RE	ABDOMIN	ABHANDL	ABSTRACT	ACADEMI	ACADEMI	ACADEMI	ACADEMI
2	<b>BIOL A</b>	1.89E-07	3.41E-05	3.48E-05	0.00015	1.86E-05	9.32E-05	5.4E-06	2.13E-06	2.02E-05	7.86E-06	1.79E-05	7.41E-06	4.18E-05
3	<b>BIOL B</b>	3.33E-08	2.54E-05	0.000113	0.000112	1.4E-05	6.42E-05	2.2E-06	4.79E-07	1.39E-05	6.4E-06	1.53E-05	6.4E-06	2.26E-05
4	<b>BIOL C</b>	1.44E-06	0.000117	3.63E-05	0.00021	4.27E-05	0.000111	7.98E-06	1.71E-06	1.1E-05	1.55E-05	2.13E-05	2.3E-05	3.89E-05
5	<b>BIOL D</b>	2.52E-06	2.73E-05	6.03E-05	0.000131	1.61E-05	6.83E-05	1.2E-06	4.1E-07	1.64E-05	3.77E-06	1.31E-05	3.92E-06	2.42E-05
6	<b>BIOL E</b>	6.61E-07	5.72E-05	3.64E-05	0.000134	1.69E-05	8.54E-05	3.51E-06	2.65E-06	1.73E-05	7.82E-06	1.8E-05	1.31E-05	3.21E-05
7	<b>BIOL F</b>	1.11E-07	2.18E-05	2.4E-05	0.000231	2.56E-05	8.09E-05	3.76E-06	8.75E-07	1.99E-05	6.85E-06	1.4E-05	7.39E-06	3.65E-05
8	<b>BIOL G</b>	6.8E-07	4.31E-05	3.36E-05	0.000146	1.76E-05	6.99E-05	7.27E-06	1.36E-06	1.84E-05	1.68E-05	2.53E-05	2.6E-05	3.57E-05
9	<b>BIOL H</b>	1.68E-06	3.42E-05	8.99E-05	0.000112	1.67E-05	7.16E-05	1.44E-06	6.58E-07	1.42E-05	4.51E-06	1.3E-05	5.7E-06	2.12E-05
10	<b>BIOL I</b>	6.28E-08	2.03E-05	2.58E-05	0.000225	2.77E-05	8.1E-05	3.31E-06	7.44E-07	2E-05	5.34E-06	1.25E-05	5.05E-06	3.35E-05
11	<b>PM1</b>	6.49E-08	2.94E-05	4.71E-05	0.000157	1.74E-05	7.37E-05	2.2E-06	6.61E-07	1.97E-05	6.69E-06	1.7E-05	6.32E-06	3E-05
12	<b>PM2</b>	1.72E-07	3.03E-05	3.87E-05	0.00016	1.93E-05	8.75E-05	4.56E-06	1.75E-06	2.04E-05	1.03E-05	1.99E-05	8.94E-06	3.59E-05
13	<b>PM3</b>	9.83E-09	2.76E-05	2.54E-05	0.000243	3.59E-05	9.31E-05	5.88E-06	1.42E-06	1.84E-05	1.22E-05	1.77E-05	8.23E-06	4.65E-05
14	<b>PM4</b>	1.07E-07	2.44E-05	3.25E-05	0.000139	1.57E-05	7.11E-05	1.94E-06	3.73E-06	2.17E-05	5.5E-06	1.47E-05	5.61E-06	2.85E-05
15	<b>PM5</b>	9.28E-09	4.27E-05	2.23E-05	0.000242	2.55E-05	8.29E-05	1.23E-05	4.38E-06	1.57E-05	1.35E-05	1.97E-05	1.09E-05	5.14E-05
16	<b>Groups</b>	8.97E-07	4.57E-05	4.13E-05	0.000162	2.25E-05	8.25E-05	4.83E-06	1.41E-06	1.72E-05	9.9E-06	1.84E-05	1.35E-05	3.4E-05
17	<b>Panel</b>	6.39E-08	3.27E-05	3.26E-05	0.000192	2.21E-05	8.05E-05	6.3E-06	2.69E-06	1.87E-05	9.86E-06	1.8E-05	8.3E-06	3.96E-05

**Figure 37. Excerpt of SAPV of the Biology research groups, research groups together, panel members and panel using journal similarity matrix**

A Python script ‘sim\_adapted\_pub\_vectors\_journals.py’ (the calculation of SAPVs is carried out by the sa\_vector function, see section 5) is used that takes as input the similarity matrix and the weights (number of publications) per journals (‘Biology groups and panel\_ journals title.xlsx’, Figure 24). This script calculates SAPVs for all entities. We keep ‘matrix.h5’ (see section 3.3), ‘cosine.net.gz’ (see section 3.3), ‘translate.xlsx’ (Figure 30), ‘Journal name change.xlsx’ (Figure 31) and ‘Biology research groups and panel\_ journals title.xlsx’ in a folder. We run the program as:

```
python sim_adapted_pub_vectors_journals.py matrix.h5 "Biology research groups and panel_journals title.xlsx"
```

This program calculates the SAPV (Figure 37) of each entity and stores the results in an output file named ‘Biology research groups and panel\_ journals title-SA-Vector.xlsx’ (Figure 38).

**b) Euclidean distance between similarity-adapted publication vectors**

Subsequently, we determine the Euclidean distances between the SAPV of different entities: individual research groups, panel members, research groups together and panel. We reuse formula 6. Again, we use the implementation of Euclidean distance in `scipy.spatial.dist`. It is mentionable that the Python script ‘`sim_adapted_pub_vectors_journals.py`’ executes both formulas 5 and 6. Although the matrix and vectors are large, the calculation of SAPV and distances is relatively fast, due to the use of efficient matrix procedures implemented in NumPy (<http://www.numpy.org>) and SciPy (<http://www.scipy.org>).

	A	B	C	D	E	F	G	H	I	J	K
1		<b>BIOL A</b>	<b>BIOL B</b>	<b>BIOL C</b>	<b>BIOL D</b>	<b>BIOL E</b>	<b>BIOL F</b>	<b>BIOL G</b>	<b>BIOL H</b>	<b>BIOL I</b>	<b>PM1</b>
2	<b>BIOL A</b>	0	0.009062	0.014256	0.009233	0.005527	0.011554	0.004659	0.010508	0.011669	0.006567
3	<b>BIOL B</b>	0.0090621	0	0.014226	0.008144	0.008788	0.014752	0.007305	0.004819	0.014447	0.008678
4	<b>BIOL C</b>	0.0142556	0.014226	0	0.012852	0.012372	0.014017	0.013416	0.012395	0.013878	0.013236
5	<b>BIOL D</b>	0.0092333	0.008144	0.012852	0	0.009152	0.010432	0.007695	0.007784	0.009795	0.004302
6	<b>BIOL E</b>	0.0055274	0.008788	0.012372	0.009152	0	0.013866	0.004126	0.008679	0.013702	0.008177
7	<b>BIOL F</b>	0.0115545	0.014752	0.014017	0.010432	0.013866	0	0.012628	0.015466	0.002	0.008039
8	<b>BIOL G</b>	0.0046593	0.007305	0.013416	0.007695	0.004126	0.012628	0	0.007991	0.012445	0.006305
9	<b>BIOL H</b>	0.0105077	0.004819	0.012395	0.007784	0.008679	0.015466	0.007991	0	0.015105	0.009397
10	<b>BIOL I</b>	0.0116694	0.014447	0.013878	0.009795	0.013702	0.002	0.012445	0.015105	0	0.007695
11	<b>PM1</b>	0.0065668	0.008678	0.013236	0.004302	0.008177	0.008039	0.006305	0.009397	0.007695	0
12	<b>PM2</b>	0.0029693	0.009496	0.014537	0.009172	0.005008	0.011536	0.004676	0.010742	0.011659	0.006429
13	<b>PM3</b>	0.011325	0.014595	0.013314	0.010791	0.013692	0.002731	0.012554	0.015285	0.003791	0.008393
14	<b>PM4</b>	0.0063011	0.008996	0.015703	0.008622	0.00649	0.013511	0.004395	0.009822	0.013301	0.007268
15	<b>PM5</b>	0.009027	0.013278	0.011971	0.011959	0.011565	0.008206	0.010992	0.013881	0.009293	0.009539
16	<b>Groups</b>	0.0051377	0.007431	0.010537	0.00554	0.005223	0.009506	0.004169	0.007617	0.00928	0.004069
17	<b>Panel</b>	0.0049296	0.009588	0.012171	0.007428	0.007503	0.007459	0.006161	0.01045	0.007798	0.004292

**Figure 38. Excerpt of pairwise Euclidean distances matrix between the SAPV of the Biology individual research groups, panel members, panel and research groups together using the journal similarity matrix**

From the calculated matrix of pairwise Euclidean distances between SAPVs of Biology individual research groups, panel members, research groups together and panel (Figure 38), we extract Table 9 containing only the distances between the research groups and research groups together on the one hand and the panel and panel members on the other, for the convenience of analysis.

In Table 9, for each research group we find the shortest distances to one of the panel members, and underline and bold those. In addition, the average and standard deviation of the shortest distances are calculated. We use the average and standard deviation of the shortest distances as a comparative measure. The confidence intervals (discussed in section 2.5) are included through the typography of the values.

**Table 9. Euclidean distances between SAPV of Biology individual research groups, panel members, research groups together and panel using the journal similarity matrix**

	Groups	BIOL- A	BIOL- B	BIOL-C	BIOL D	BIOL-E	BIOL -F	BIOL- G	BIOL- H	BIOL- I
Panel	0.004	0.005	0.010	0.012	0.007	0.008	0.007	0.006	0.010	0.008
PM1	0.004	0.007	<b><u>0.009</u></b>	<b><u>0.013</u></b>	<b><u>0.004</u></b>	0.008	0.008	<b><u>0.006</u></b>	<b><u>0.009</u></b>	0.008
PM2	0.005	<b><u>0.003</u></b>	<b><u>0.010</u></b>	<b><u>0.015</u></b>	0.009	<b><u>0.005</u></b>	0.012	<b><u>0.005</u></b>	<b><u>0.011</u></b>	0.012
PM3	0.009	0.011	0.015	<b><u>0.013</u></b>	0.011	0.014	<b><u>0.003</u></b>	0.013	0.015	<b><u>0.004</u></b>
PM4	0.007	0.006	<b><u>0.009</u></b>	0.016	0.009	<b><u>0.006</u></b>	0.014	<b><u>0.004</u></b>	<b><u>0.010</u></b>	0.013
PM5	0.009	0.009	0.013	<b><u>0.012</u></b>	0.012	0.012	0.008	0.011	0.014	0.009

For each research group we determine the panel member at the highest similarity. An average shortest distance is 0.006 (SD 0.003). The number in the row corresponding to this panel member is indicated in bold and underlined. Similarities whose confidence intervals overlap with that of the highest similarities are in bold (same column).

### 3.7 Weighted cosine similarity method

Recall the formula 7. We consider a weighted similarity method (generalized cosine similarity). The weighted similarity between panel member (PM) k and research group m, according to Zhou et al. (2012) is:

$$\frac{\sum_{i=1}^N M_i^k (\sum_{j=1}^N R_j^m s_{ji})}{\sqrt{(\sum_{i=1}^N M_i^k (\sum_{j=1}^N M_j^k s_{ji})) \cdot (\sum_{i=1}^N R_i^m (\sum_{j=1}^N R_j^m s_{ji}))}}$$

$$= \frac{(M^k)^t * S * R^m}{\sqrt{(M^k)^t * S * M^k} \cdot \sqrt{(R^m)^t * S * R^m}}$$

The numerator is nothing but the matrix multiplication:  $(M^k)^t * S * R^m$ , where t denotes matrix transposition, S is the journal similarity matrix,  $M^k$  denotes the column matrix of publications of panel member PMk and  $R^m$  denotes the column matrix of publications of research group m. Similarly, the two products under the square root in the denominator are:  $(M^k)^t * S * M^k$  and  $(R^m)^t * S * R^m$ . The result is the similarity between panel member PMk and research group m.

This value is calculated for each panel member and each research group. Similarity-weighted cosine is implemented in Python as a fairly straightforward set of matrix operations (see section 5, `weighted_cosine`).

We keep ‘`matrix.h5`’ (see section 3.3), ‘`cosine.net.gz`’ (see section 3.3), ‘`translate.xlsx`’ (Figure 30), ‘`Journal name change.xlsx`’ (Figure 31) and ‘`Biology research groups and panel_journals title.xlsx`’ in a folder. A python script ‘`cosine-journals.py`’ is used that takes as input the similarity matrix and the weights (number of publications) per journals (‘`Biology research groups and panel_journals title.xlsx`’, Figure 24), and calculates SAPVs for all entities, as well as the pairwise distances between them. We run the program as:

```
python cosine-journals.py matrix.h5 "Biology research groups and panel_journals
title.xlsx"
```

This program calculates the WCS value in an output file as ‘`Biology research groups and panel_journals title-cosine.xlsx`’ (Figure 39).

	A	B	C	D	E	F	G	H	I	J	K
1		<b>BIOL A</b>	<b>BIOL B</b>	<b>BIOL C</b>	<b>BIOL D</b>	<b>BIOL E</b>	<b>BIOL F</b>	<b>BIOL G</b>	<b>BIOL H</b>	<b>BIOL I</b>	<b>PM1</b>
2	<b>BIOL A</b>	1	0.477135	0.239199	0.445682	0.748114	0.327303	0.811947	0.443214	0.311397	0.608978
3	<b>BIOL B</b>	0.477135	1	0.26262	0.575027	0.435991	0.261389	0.599852	0.755236	0.262527	0.56729
4	<b>BIOL C</b>	0.239199	0.26262	1	0.320861	0.50809	0.246878	0.377221	0.469075	0.242844	0.266897
5	<b>BIOL D</b>	0.445682	0.575027	0.320861	1	0.465078	0.615811	0.601186	0.670374	0.626184	0.890472
6	<b>BIOL E</b>	0.748114	0.435991	0.50809	0.465078	1	0.25965	0.849798	0.527657	0.255121	0.558339
7	<b>BIOL F</b>	0.327303	0.261389	0.246878	0.615811	0.25965	1	0.348518	0.263602	0.974342	0.728757
8	<b>BIOL G</b>	0.811947	0.599852	0.377221	0.601186	0.849798	0.348518	1	0.65092	0.344583	0.726245
9	<b>BIOL H</b>	0.443214	0.755236	0.469075	0.670374	0.527657	0.263602	0.65092	1	0.265649	0.613234
10	<b>BIOL I</b>	0.311397	0.262527	0.242844	0.626184	0.255121	0.974342	0.344583	0.265649	1	0.727727
11	<b>PM1</b>	0.608978	0.56729	0.266897	0.890472	0.558339	0.728757	0.726245	0.613234	0.727727	1
12	<b>PM2</b>	0.816051	0.430277	0.199062	0.452574	0.896183	0.329606	0.823954	0.429091	0.313805	0.620458
13	<b>PM3</b>	0.325402	0.270032	0.242843	0.566426	0.247812	0.940286	0.33	0.266648	0.899825	0.66963
14	<b>PM4</b>	0.643094	0.450345	0.174141	0.516003	0.628511	0.308277	0.770398	0.468519	0.302859	0.642316
15	<b>PM5</b>	0.610003	0.341137	0.461361	0.320882	0.462911	0.426667	0.462938	0.374126	0.365631	0.410695
16	<b>Groups</b>	0.798777	0.640529	0.589401	0.763407	0.851621	0.570504	0.908887	0.7289	0.567675	0.831591
17	<b>Panel</b>	0.809259	0.546108	0.371235	0.685253	0.745201	0.636141	0.827808	0.575564	0.603018	0.836503

**Figure 39. Excerpt of WCS value matrix of the Biology individual research groups, panel members, groups and panel using the journal similarity matrix**

From the calculated WCS value matrix of Biology individual research groups, panel members, research groups together and panel in journals (Figure 39), we extract Table 10 containing only the WCS value of the research groups and research groups together on the one hand and the panel and panel members on the other, for the convenience of analysis.

In Table 10, for each research group we find the highest similarity to one of the panel members, and underline and bold those. The confidence intervals (discussed in the section 2.5) are included through the typography of the values. We calculate similarity between two entities based on their publication vectors. We generated 1000 independent bootstrap samples and each time calculated the similarity.

**Table 10. WCS value of the Biology groups, panel members, panel and research groups together using the journal similarity matrix**

	Groups	BIOL-A	BIOL-B	BIOL-C	BIOL-D	BIOL-E	BIOL-F	BIOL-G	BIOL-H	BIOL-I
Panel	0.898	0.809	0.546	0.371	0.685	0.745	0.636	0.828	0.576	0.603
PM1	0.832	0.609	<b><u>0.567</u></b>	0.267	<b><u>0.890</u></b>	0.558	0.729	0.726	<b><u>0.613</u></b>	0.728
PM2	0.781	<b><u>0.816</u></b>	<b>0.430</b>	0.199	0.453	<b><u>0.896</u></b>	0.330	<b><u>0.824</u></b>	0.429	0.314
PM3	0.544	0.325	0.270	0.243	0.566	0.248	<b><u>0.940</u></b>	0.330	0.267	<b><u>0.900</u></b>
PM4	0.689	0.643	<b>0.450</b>	0.174	0.516	0.629	0.308	<b>0.770</b>	<b>0.469</b>	0.303
PM5	0.600	0.610	0.341	<b><u>0.461</u></b>	0.321	0.463	0.427	0.463	0.374	0.366

For each research group we determine the panel member at the highest similarity. The number in the row corresponding to this panel member is indicated in bold and underlined. Similarities whose confidence intervals overlap with that of the highest similarities are in bold (same column).

**Table 11. WCD value of the Biology groups, panel members, panel and research groups together using the journal similarity matrix**

	Groups	BIOL-A	BIOL-B	BIOL-C	BIOL-D	BIOL-E	BIOL-F	BIOL-G	BIOL-H	BIOL-I
Panel	0.102	0.191	0.454	0.629	0.315	0.255	0.364	0.172	0.424	0.397
PM1	0.168	0.391	<b><u>0.433</u></b>	0.733	<b><u>0.110</u></b>	0.442	0.271	0.274	<b><u>0.387</u></b>	0.272
PM2	0.219	<b><u>0.184</u></b>	<b>0.570</b>	0.801	0.547	<b><u>0.104</u></b>	0.670	<b><u>0.176</u></b>	0.571	0.686
PM3	0.456	0.675	0.730	0.757	0.434	0.752	<b><u>0.060</u></b>	0.670	0.733	<b><u>0.100</u></b>
PM4	0.311	0.357	<b>0.550</b>	0.826	0.484	0.371	0.692	<b>0.230</b>	<b>0.531</b>	0.697
PM5	0.400	0.390	0.659	<b><u>0.539</u></b>	0.679	0.537	0.573	0.537	0.626	0.634

The lowest similarity between a group and a panel member is underlined and printed in bold.

Since the barycenter (see section 3.5) and SAPV (see section 3.6) approaches are distance-based rather than similarity-based, we use  $1 - \text{WCS}$  as values to obtain dissimilarity values: weighted cosine dissimilarity (Table 11), denoted as WCD, which can more easily be compared with the other two approaches. For the sake of simplicity, the results are shown under the WCS method.

## 4 Heat map

A heat map with hierarchical clustering is a two-dimensional representation of data where the values are represented by colors and arranging items in a hierarchy based on the similarity between them. It provides an immediate visual summary of information.

We have proposed three methods, each of which can be applied at the level of WoS SCs and journals. This leads to six approaches, as follows:

### WoS SCs

- i) Barycenter
- ii) Similarity-adapted publication vector (SAPV)
- iii) Weighted cosine similarity (WCS)

### Journals

- iv) Barycenter
- v) Similarity-adapted publication vector (SAPV)
- vi) Weighted cosine similarity (WCS)

We calculate Spearman's rank-order correlation coefficient between each pair of the six methods. More specifically, we determine the correlation using the distances between barycenters and between SAPVs, and dissimilarity of individual research groups and panel members using  $1 - \text{WCS}$ . We create an MS Excel file ( Figure 40) containing:

- i) Euclidean distances between barycenters of the Biology individual research groups and panel members at the level of WoS SCs,
- ii) Euclidean distances between barycenters of the Biology individual research groups and panel members at the level of journals,
- iii) Euclidean distances between SAPVs of the Biology individual research groups and panel members at the level of WoS SCs,
- iv) Euclidean distances between SAPVs of the Biology individual research groups and panel members at the level of journals,
- v) WCS value of Biology individual research groups and panel members at the level of WoS SCs,



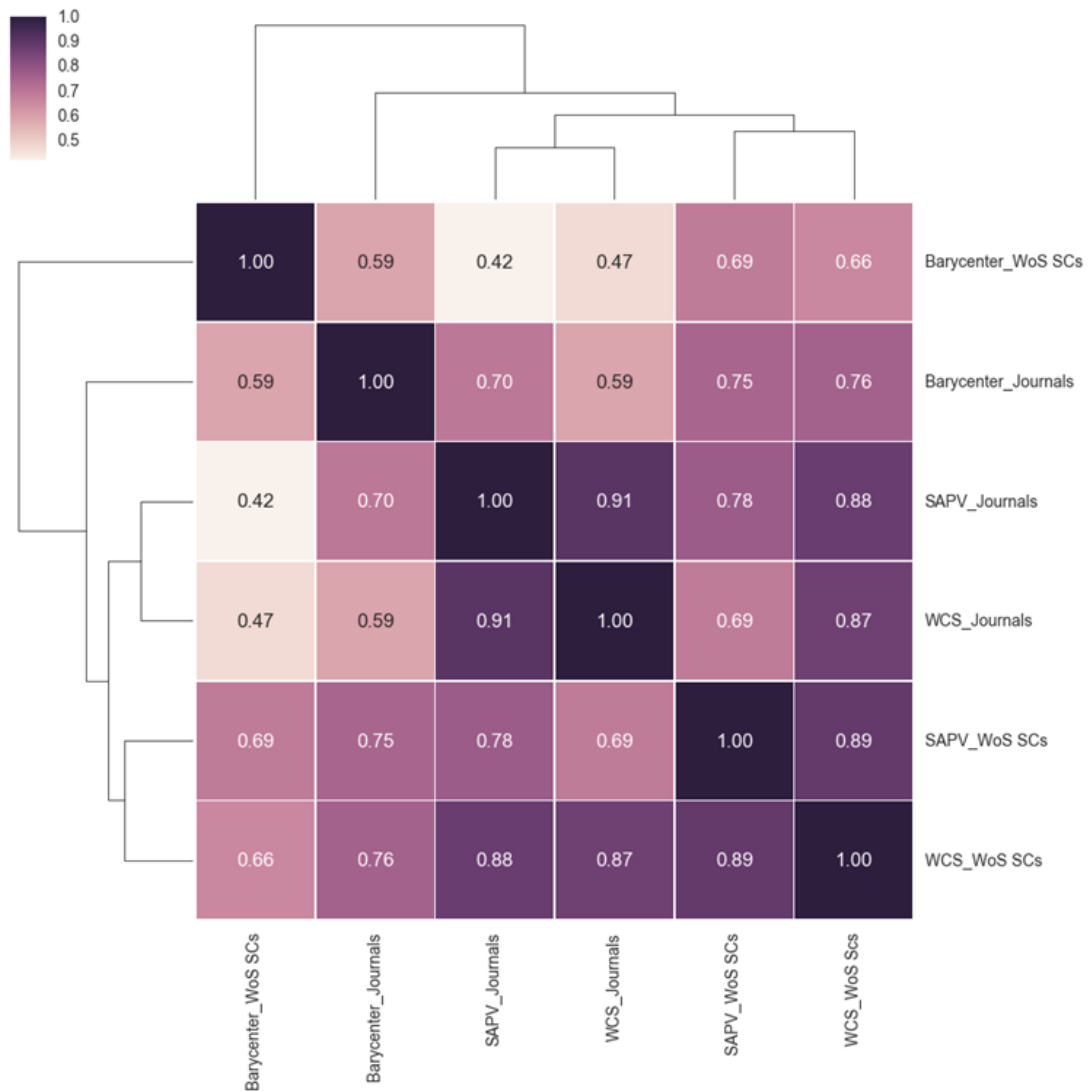
- vi) WCS value of Biology individual research groups and panel members at the level of journals.

	A	B	C	D	E	F
1	Barycenter_ BIOL_WoS SCs	Barycenter_ BIOL_Journals	SAPV_ BIOL_WoS SCs	SAPV_ BIOL_Journals	WCS_ BIOL_WoS SCs	WCS_ PHYS_Journals
2	0.344	0.154	0.057	0.007	0.220	0.391
3	0.075	0.125	0.052	0.009	0.111	0.433
4	0.075	0.198	0.059	0.013	0.326	0.733
5	0.093	0.105	0.036	0.004	0.052	0.110
6	0.282	0.169	0.046	0.008	0.196	0.442
7	0.201	0.239	0.076	0.008	0.277	0.271
8	0.200	0.056	0.033	0.006	0.114	0.274
9	0.123	0.146	0.066	0.009	0.183	0.387
10	0.132	0.164	0.071	0.008	0.259	0.272
11	0.042	0.016	0.017	0.003	0.031	0.184
12	0.409	0.249	0.091	0.009	0.314	0.570
13	0.317	0.168	0.073	0.015	0.460	0.801
14	0.444	0.190	0.089	0.009	0.393	0.547
15	0.088	0.090	0.023	0.005	0.028	0.104
16	0.317	0.257	0.071	0.012	0.455	0.670
17	0.165	0.091	0.038	0.005	0.090	0.176
18	0.454	0.227	0.104	0.011	0.403	0.571

**Figure 40. Excerpt of the dissimilarities/distances between panel members and individual research groups according to each of the six approaches**

We import data from the MS Excel file (Figure 40) to SPSS, and calculate the Spearman's rank-order correlation coefficient between the six approaches.

The heat map with hierarchical clustering (Figure 41) shows that correlations between the two levels of aggregation based on barycenter ( $\rho = 0.59$ ) is moderate, while SAPV ( $\rho = 0.91$ ) and WCS ( $\rho = 0.87$ ) are strong. The correlations between the barycenter methods on the one hand and the SAPV and WCS methods on the other are moderate to low. In addition, correlation between SAPV and WCS in both WoS SCs and journals are very strong. Overall, this suggests that the influence of the 2D reduction is substantial. Moreover, in general WoS SC and journal results correlate strongly. That suggests that the level of aggregation has minor influence for determining cognitive distances.



**Figure 41. Heat map with hierarchical clustering based on correlation coefficient between six approaches in the Biology department**

## 5 Programming code in Python

The essential code to calculate barycenters, similarity-adapted publication vectors, and similarity weighted cosine is as follows:

```
import numpy as np
import pandas as pd

def ensure_symmetric(M):
    m, n = M.shape
    if m != n:
```

```

        raise ValueError("M is not square!")

def barycenter(counts, coords):
    """Calculate the barycenter for the given counts and coordinates"""
    m, n = coords.shape

    if len(counts) != m:
        raise ValueError("'counts' should have the same number of items "
                          "(now: {}) as rows of 'coords' (now: {})".format(
                              len(counts), m))

    # Transposing twice because of broadcasting rules
    a = (coords.T * counts).T
    return a.sum(axis=0) / sum(counts)

def sa_vector(counts, S, normalize=True):
    """Calculate the similarity adapted vector for the given counts and
    similarity matrix S
    """
    ensure_symmetric(S)

    if len(counts) != len(S):
        raise ValueError("'counts' should have the same number of items "
                          "(now: {}) as rows of similarity matrix (now: {})".
                          .format(len(counts), len(S)))

    # Transposing twice because of broadcasting rules
    raw_sa_vector = (S.T * counts).T.sum(axis=0)
    return raw_sa_vector / raw_sa_vector.sum() if normalize else raw_sa_vector

def weighted_cosine(u, v, S):
    """Calculate cosine similarity between vectors u and v, weighted by
    similarity matrix S
    """
    ensure_symmetric(S)
    if len(u) != len(v) != len(S):
        raise ValueError("Vectors or similarity matrix of different length.")

    u = u / np.sum(u)
    v = v / np.sum(v)

    return u.dot(S).dot(v) / np.sqrt(u.dot(S).dot(u) * v.dot(S).dot(v))

```

## Code to calculate top-down correlation, accounting for ties

```
from __future__ import division

import itertools
import numpy as np
from operator import itemgetter

def savage_score(rank, n, endrank=None):
    """Calculate savage score for given rank in list of n items

    If endrank is given, return array of savage scores for all items between
    rank and endrank.

    """
    if rank < 1 or rank > n:
        raise ValueError("rank should be between 1 and n")

    if not hasattr(savage_score, 'lookup') or n != savage_score.n:
        savage_score.n = n
        arr = np.cumsum([1 / i for i in xrange(n, 0, -1)])
        savage_score.lookup = arr[::-1]

    if endrank is not None:
        return savage_score.lookup[rank - 1:endrank - 1]
    else:
        return savage_score.lookup[rank - 1]

def avg_savage_score(start, length, n):
    return np.average(savage_score(start, n, start + length))

def _ties(values):
    """Find ties in list of values"""
    prev = None
    ties = []
    start = 0
    final = object()

    # We add an element 'final' at the end, to ensure that the last entry is
    # also properly handled.
    for rank, value in enumerate(itertools.chain(values, [final]), start=1):
        if value == prev:
            if start == 0: # start of a tie
                start = rank - 1
            else:
                if start != 0: # end of a tie
                    ties.append((start, rank - start))
                    start = 0
        prev = value

    return ties

def savage_scores_with_ties(values):

    def next_tie():
```

```

    try:
        return ties.pop(0)
    except IndexError:
        return -1, -1

n = len(values)
ties = _ties(values)
tierank, tielength = next_tie()

for rank in range(1, n + 1):
    value = values[rank - 1]
    if rank >= tierank and rank < tierank + tielength:
        yield avg_savage_score(tierank, tielength, n), value
    else:
        if rank == tierank + tielength:
            tierank, tielength = next_tie()
        yield savage_score(rank, n), value

def dict_with_savage_scores(d):
    # If d is a ranked list of items, convert it to a dict with decreasing
    # values.
    if isinstance(d, list):
        d = dict(zip(d, range(len(d), 0, -1)))

    d_sorted = sorted(d.iteritems(), reverse=True, key=itemgetter(1))
    items, values = zip(*d_sorted)
    return {item: rank for item, (rank, value)
            in zip(items, savage_scores_with_ties(values))}

def top_down_correlation(R, Q):
    n = len(R)
    assert len(Q) == n

    R_scores = dict_with_savage_scores(R)
    Q_scores = dict_with_savage_scores(Q)

    return (sum(R_scores[item] * Q_scores[item] for item in R_scores) - n) / \
        (n - savage_score(1, n))

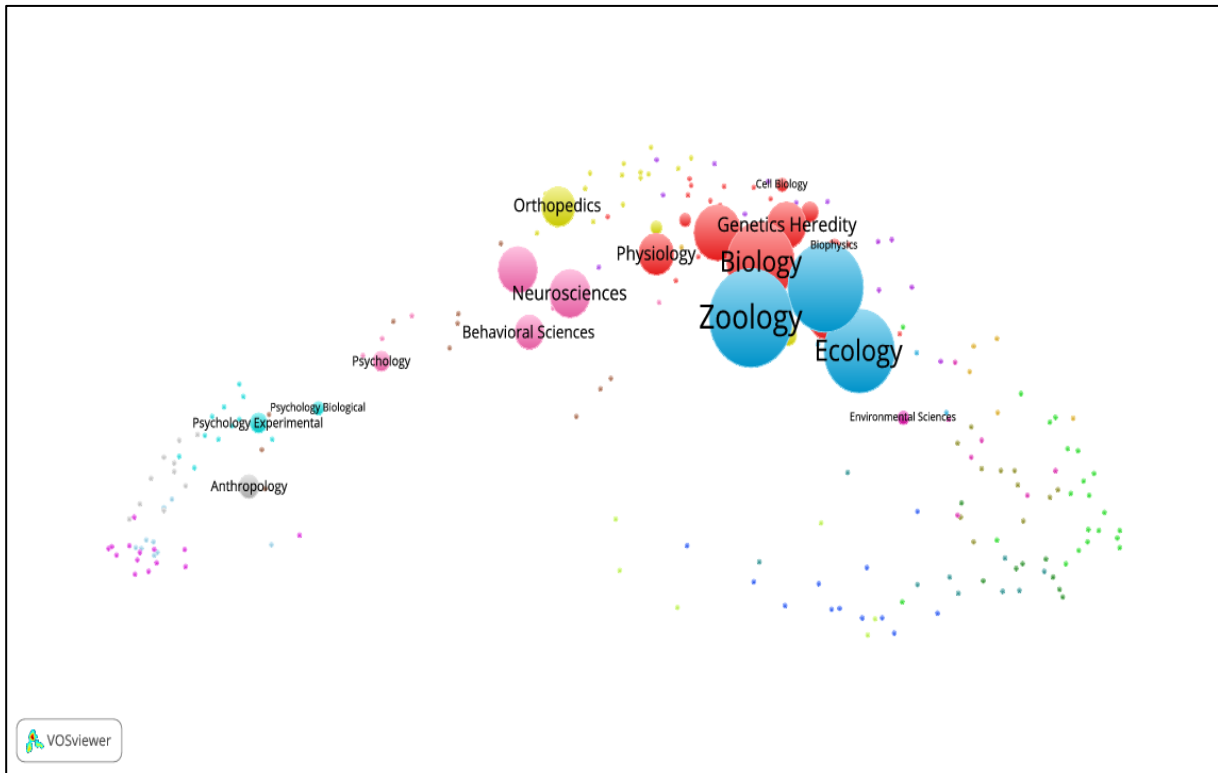
```

## References

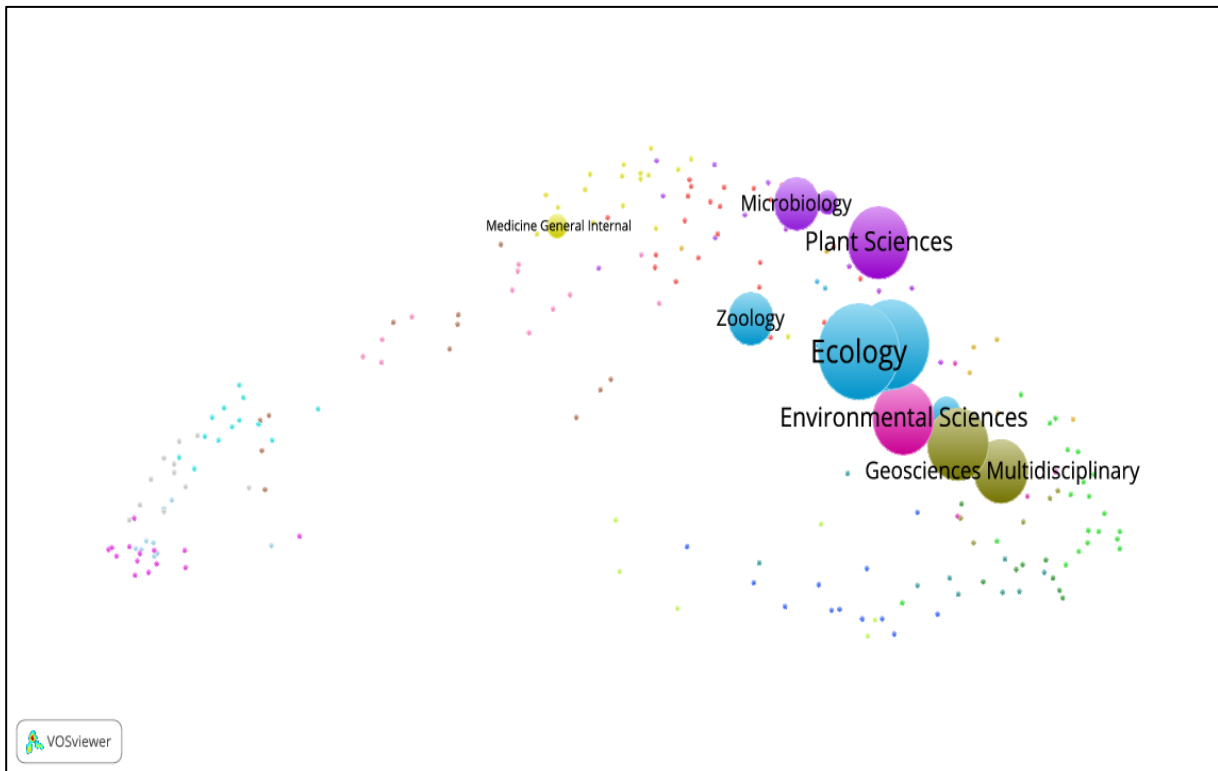
- Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling*. New York, NY: Springer New York.
- de Nooy, W., Mrvar, A., & Batagelj, V. (2012). *Exploratory Social Network Analysis with Pajek* (2nd edition). England ; New York: Cambridge University Press.
- Efron, B., & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.
- Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics*. Elsevier Science Publishers. Retrieved from <https://uhdSPACE.uhasselt.be/dSPACE/handle/1942/587>
- Huson, L. W. (2007). Performance of some correlation coefficients when applied to zero-clustered data. *Journal of Modern Applied Statistical Methods*, 6(2), 530–536.
- Iman, R. L., & Conover, W. J. (1987). A measure of Top-Down correlation. *Technometrics*, 29(3), 351–357.
- Jin, B., & Rousseau, R. (2001). An introduction to the barycentre method with an application to China's mean centre of publication. *Libri*, 51(4), 225–233.
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7–15.
- Kington, J. (2014). Balanced Cross Sections, Shortening Estimates, and the Magnitude of Out-of-Sequence Thrusting in the Nankai Trough Accretionary Prism, Japan. *Figshare*. <https://doi.org/10.6084/m9.figshare.1015774.v1>
- Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, 94(2), 589–593.
- Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2), 348–362.
- Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. *Journal of the American Society for Information Science and Technology*, 64(12), 2573–2586.
- Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871–1887.

- Rahman, A. I. M. J., Guns, R., Leydesdorff, L., & Engels, T. C. E. (2016). Measuring the match between evaluators and evaluatees: cognitive distances between panel members and research groups at the journal level. *Scientometrics*.
- Rousseau, R. (1989). Kinematical statistics of scientific output. Part II: standardized polygonal approach. *Revue Française de Bibliométrie*, 4, 65–77.
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface*, 4(15), 707–719.
- van Eck, N. J., & Waltman, L. (2007). VOS: A New Method for Visualizing Similarities Between Objects. In R. Decker & H.-J. Lenz (Eds.), *Advances in Data Analysis* (pp. 299–306). Springer Berlin Heidelberg.
- van Eck, N. J., Waltman, L., Dekker, R., & van den Berg, J. (2010). A Comparison of Two Techniques for Bibliometric Mapping: Multidimensional Scaling and VOS. *Journal of the American Society for Information Science and Technology*, 61(12), 2405–2416.
- Verleysen, F. T., & Engels, T. C. E. (2013). Measuring internationalisation of book publishing in the social sciences and humanities using the barycentre method. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Horlesberger, & H. Moed (Eds.), *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference (ISSI), 15 - 19 July 2013* (pp. 1170–1176). Vienna, Austria.
- Verleysen, F. T., & Engels, T. C. E. (2014). Barycenter representation of book publishing internationalization in the Social Sciences and Humanities. *Journal of Informetrics*, 8(1), 234–240.
- VSNU. (2003). *Standard Evaluation Protocol 2003-2009 for Public Research Organisations*. Utrecht/den Haag/Amsterdam: VSNU, NWO and KNAW.
- VSNU, KNAW, & NWO. (2014). *Standard evaluation Protocol 2015-2021: protocol for research assessment in The Netherlands*. Utrecht/den Haag/Amsterdam: VSNU, NWO and KNAW.

## Appendix A

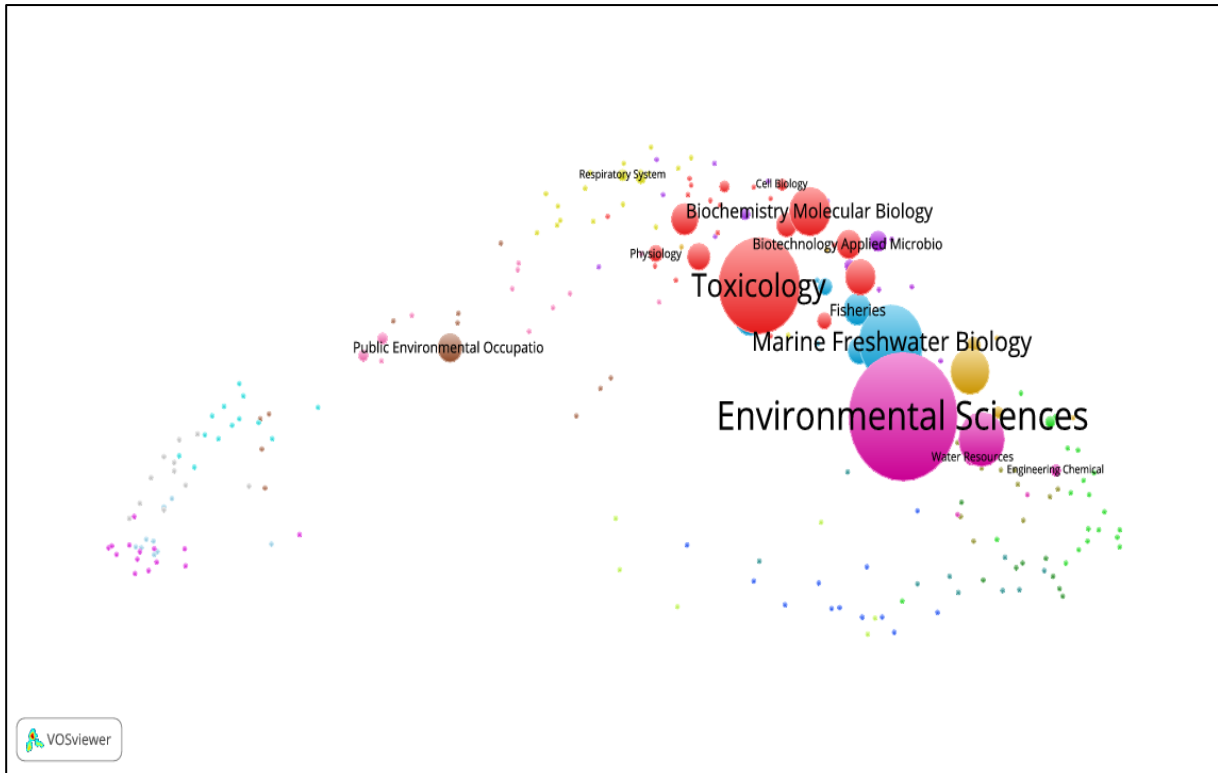


**Figure 42.** WoS SCs overlay map of BIOL-A research group's publications

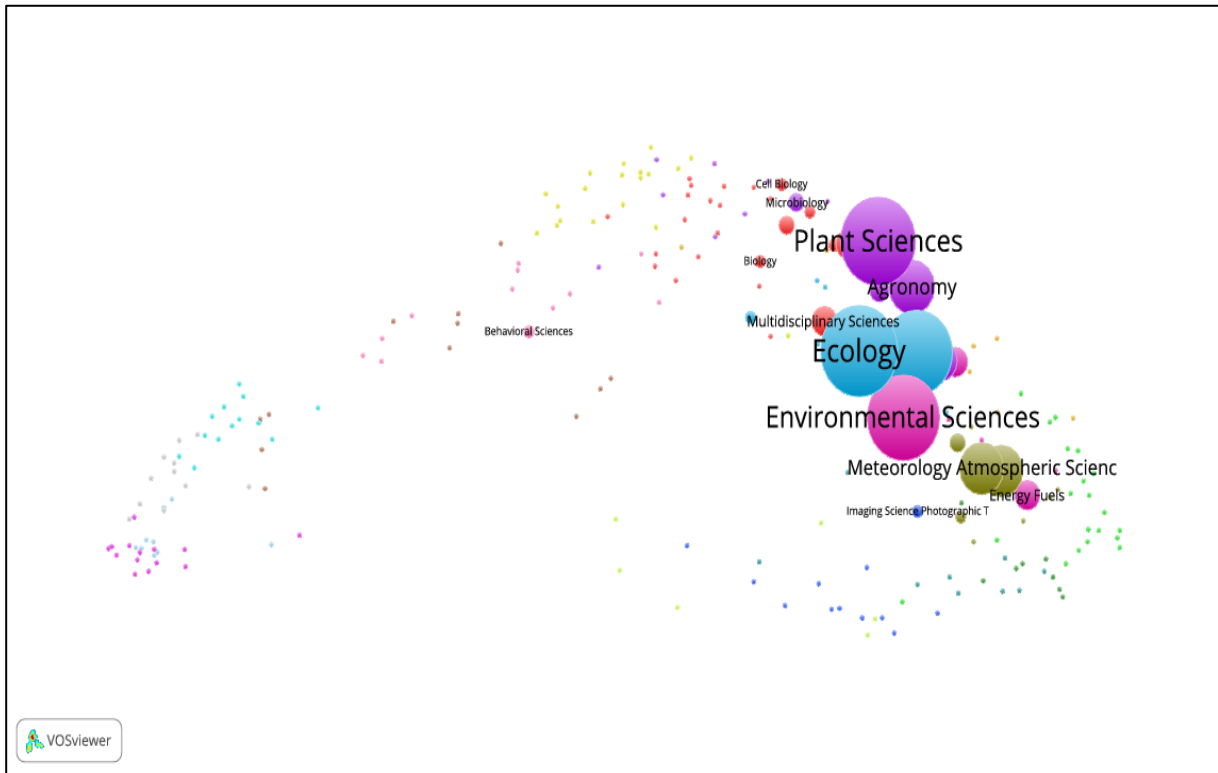


**Figure 43.** WoS SCs overlay map of BIOL-B research group's publications

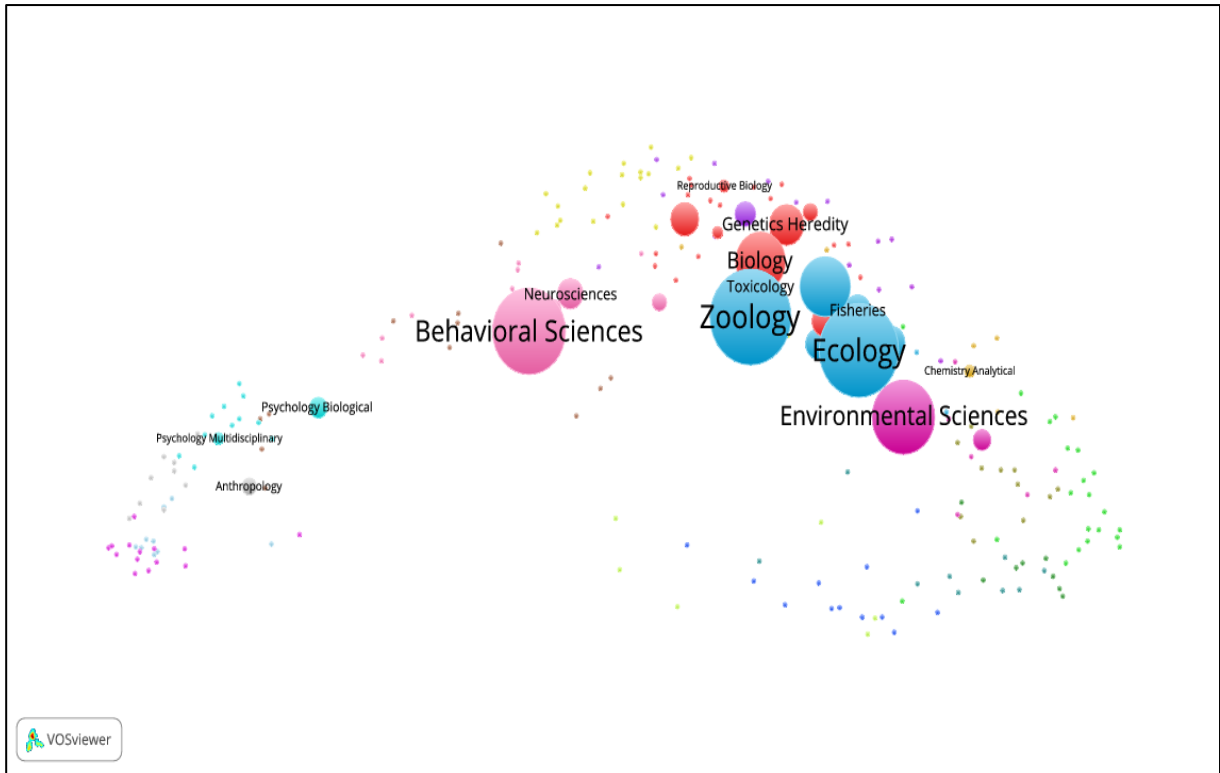




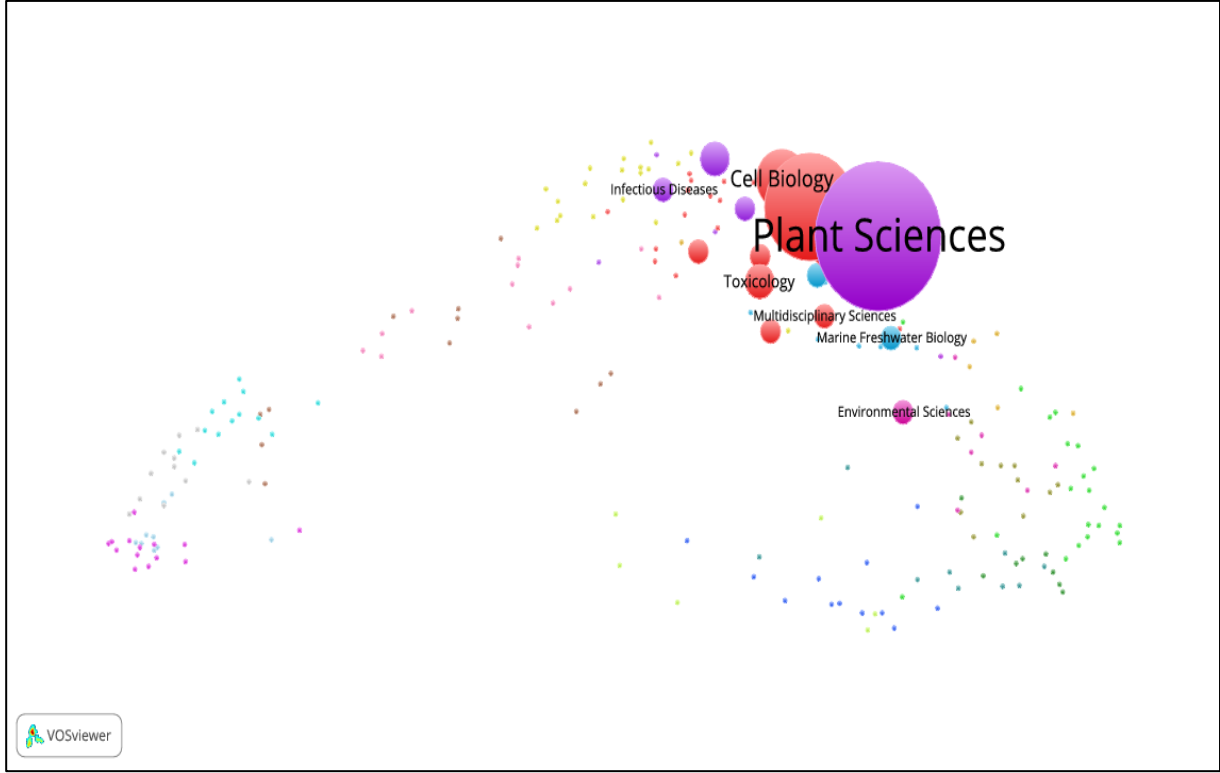
**Figure 44. WoS SCs overlay map of BIOL-C research group's publications**



**Figure 45. WoS SCs overlay map of BIOL-D research group's publications**



**Figure 46. WoS SCs overlay map of BIOL-E research group's publications**



**Figure 47. WoS SCs overlay map of BIOL-F research group's publications**

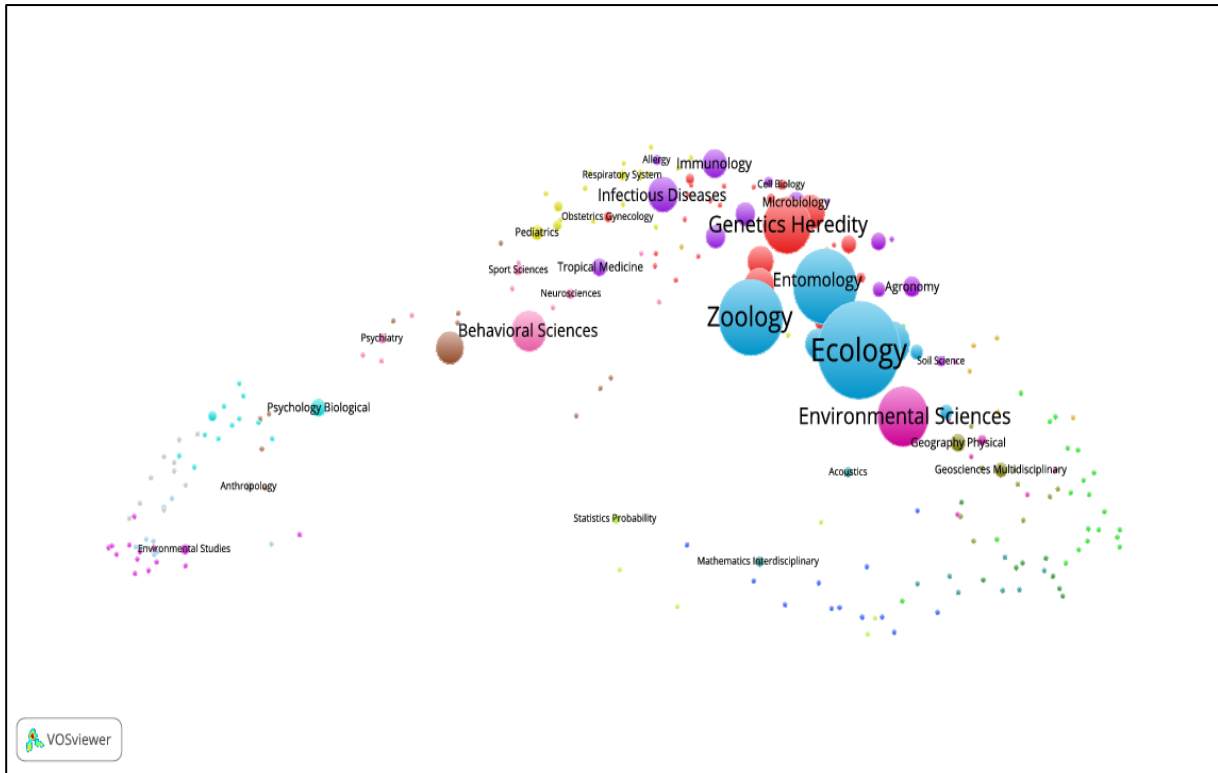


Figure 48. WoS SCs overlay map of BIOL-G research group's publications

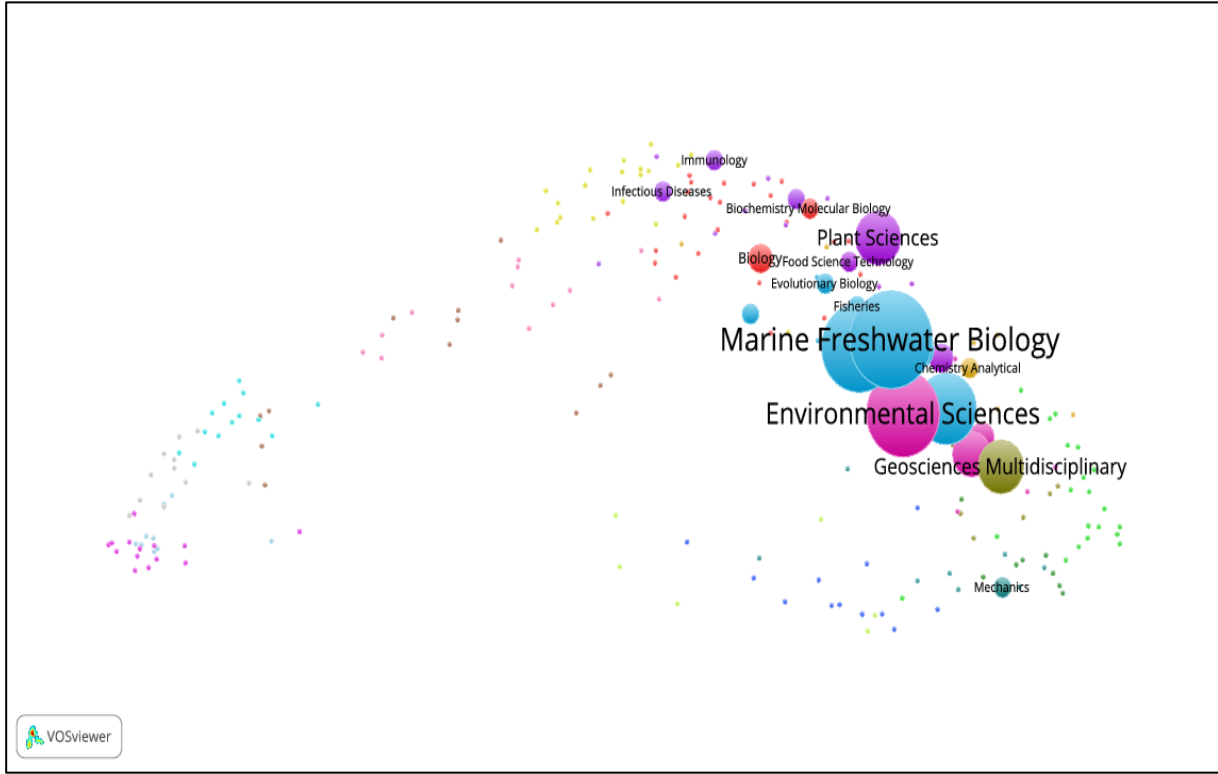
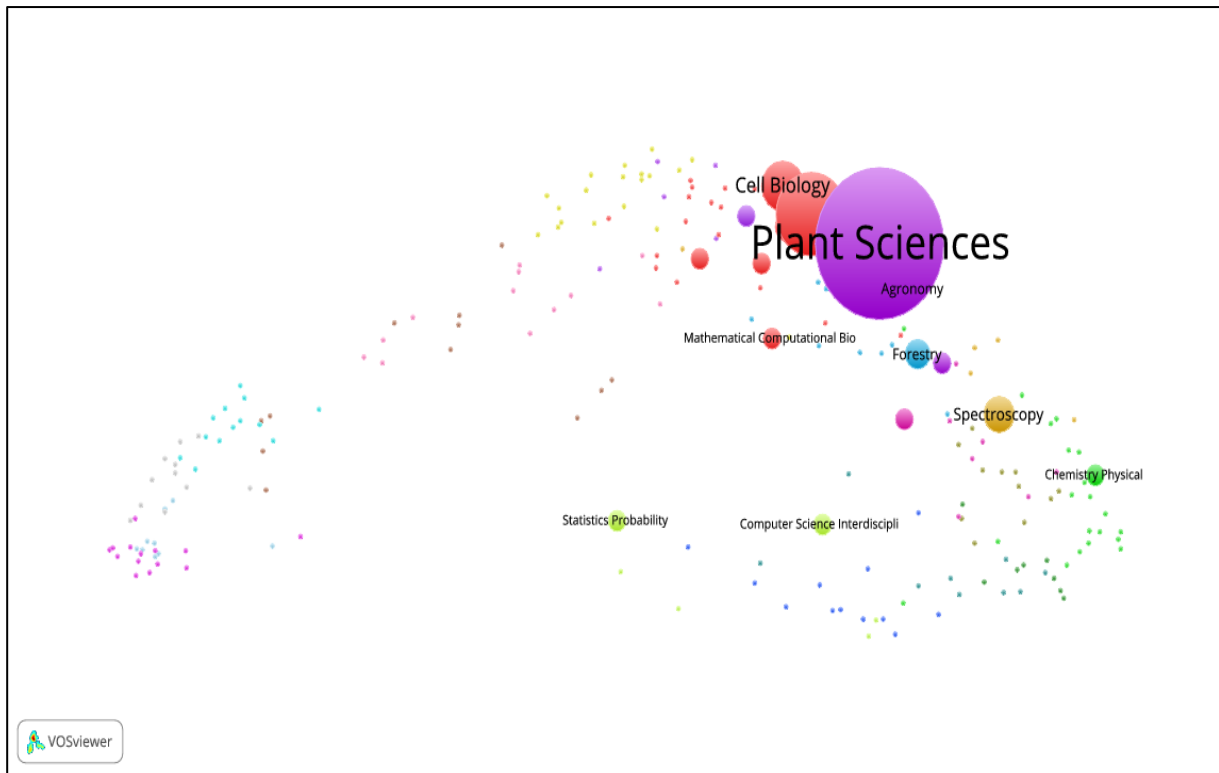
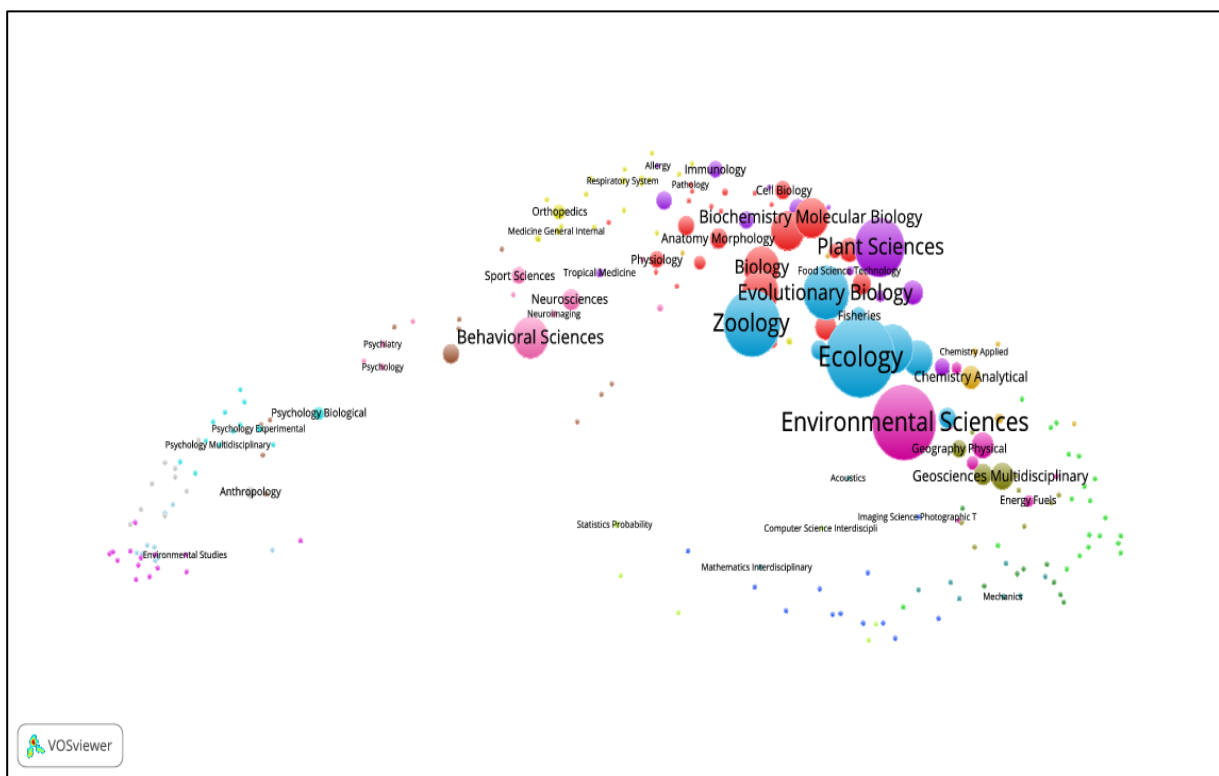


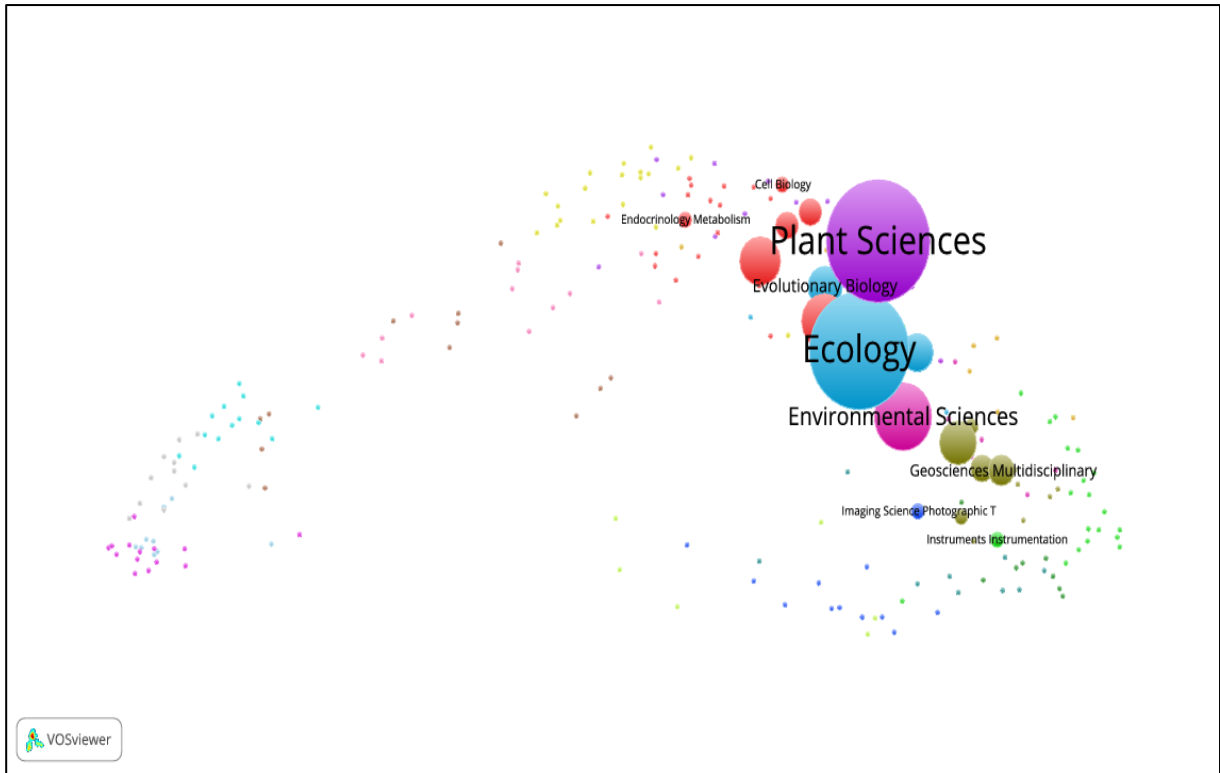
Figure 49. WoS SCs overlay map of BIOL-H research group's publications



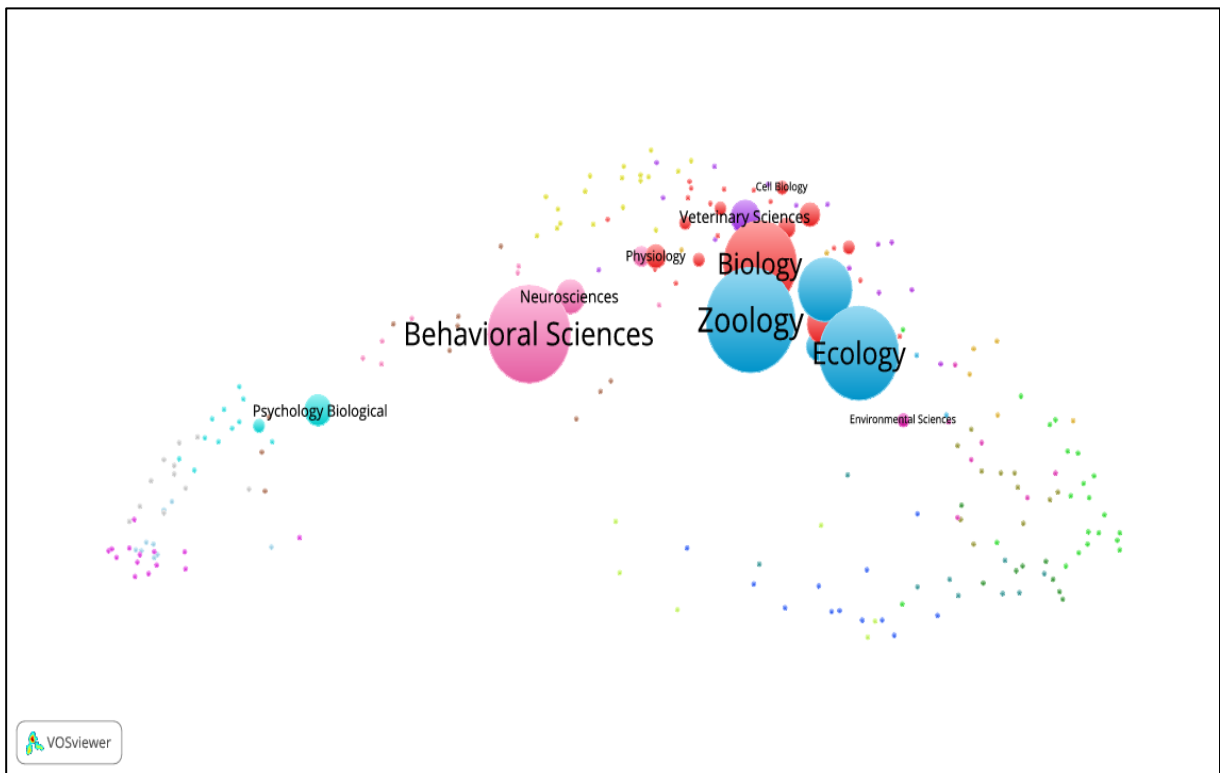
**Figure 50. WoS SCs overlay map of BIOL-I research group's publications**



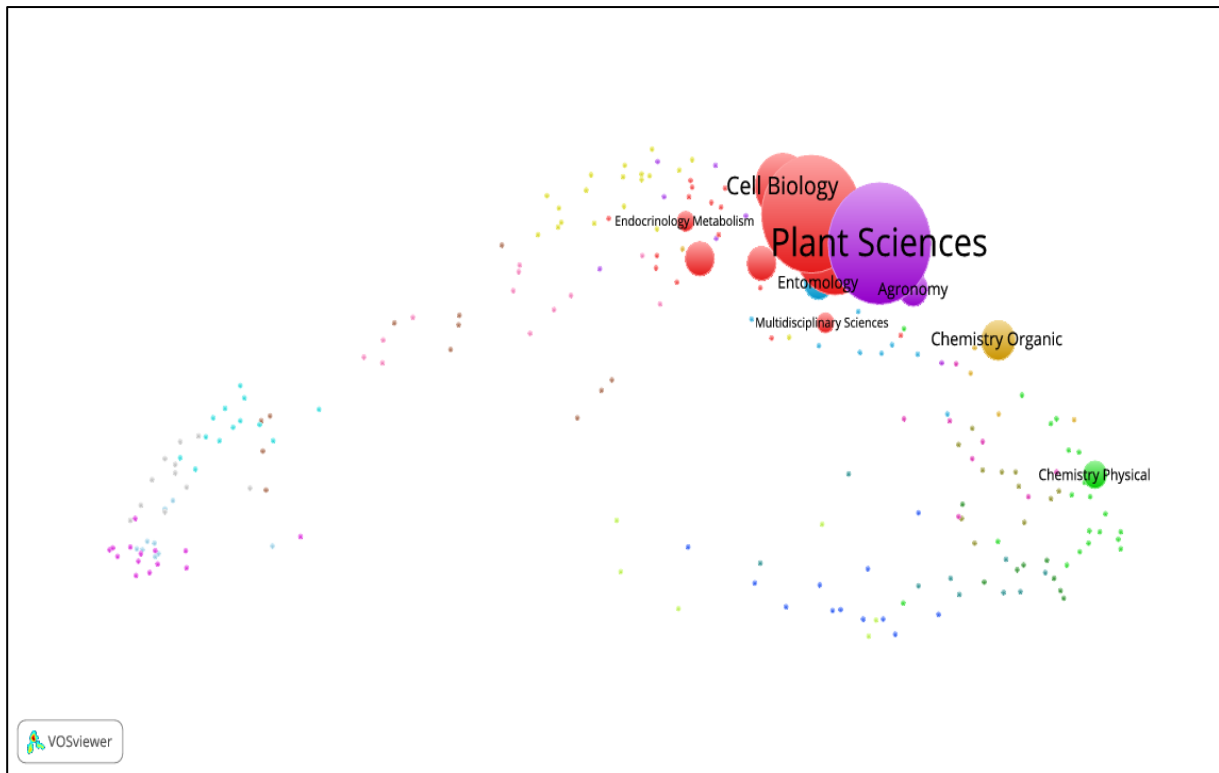
**Figure 51. WoS SCs overlay map of Biology research groups' publications**



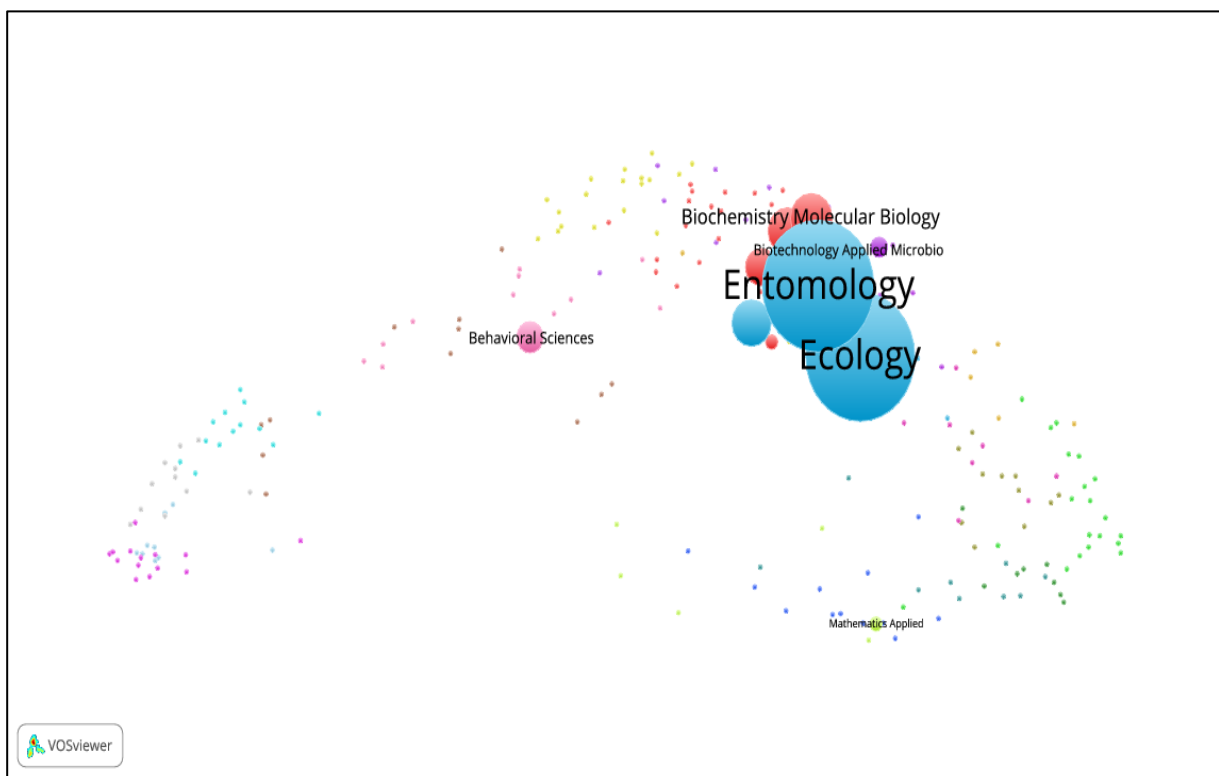
**Figure 52. WoS SCs overlay map of PM1's publications**



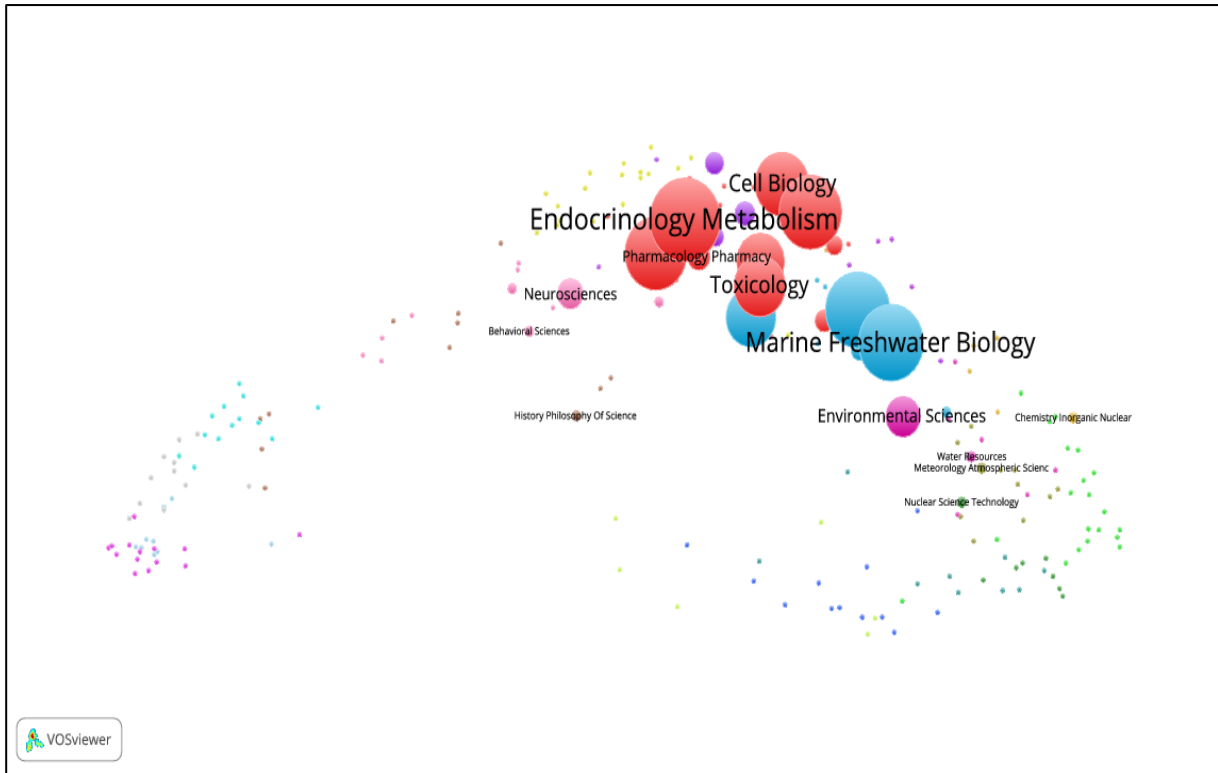
**Figure 53. WoS SCs overlay map of PM2's publications**



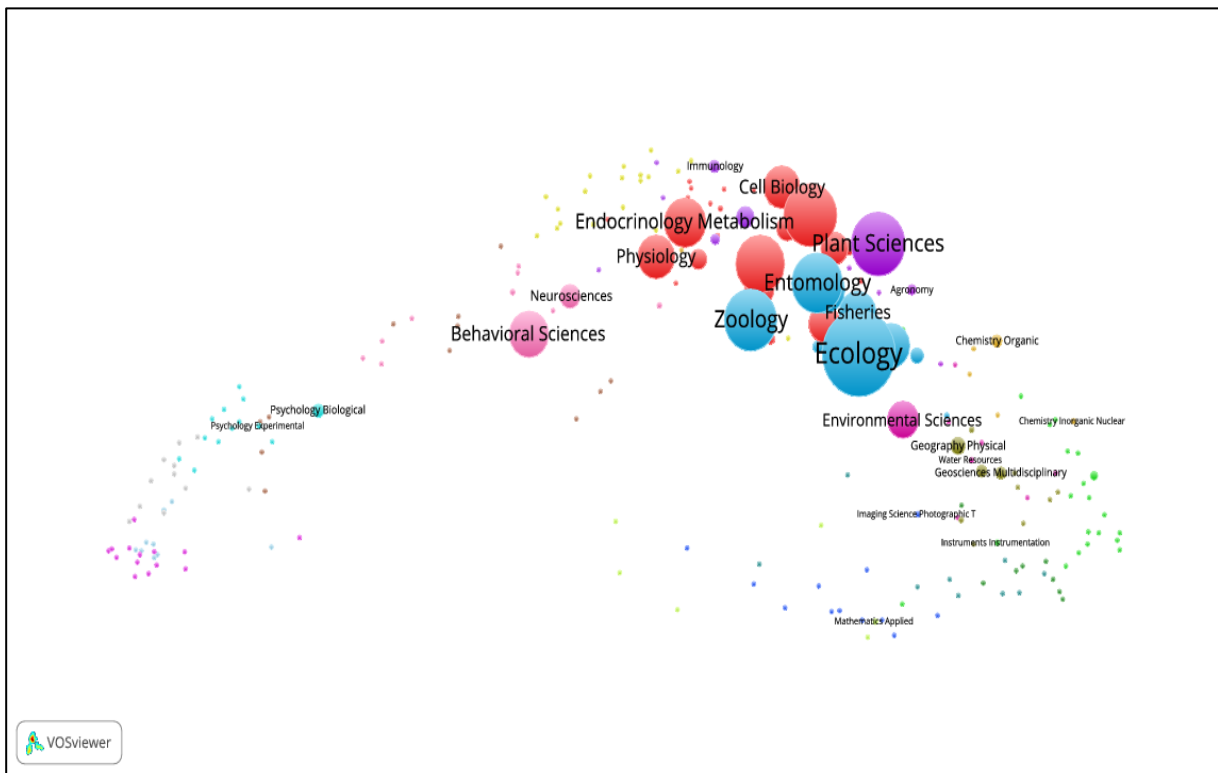
**Figure 54. WoS SCs overlay map of PM3's publications**



**Figure 55. WoS SCs overlay map of PM4's publications**



**Figure 56. WoS SCs overlay map of PM5's publications**



**Figure 57. WoS SCs overlay map of panel's publications**

## Appendix B

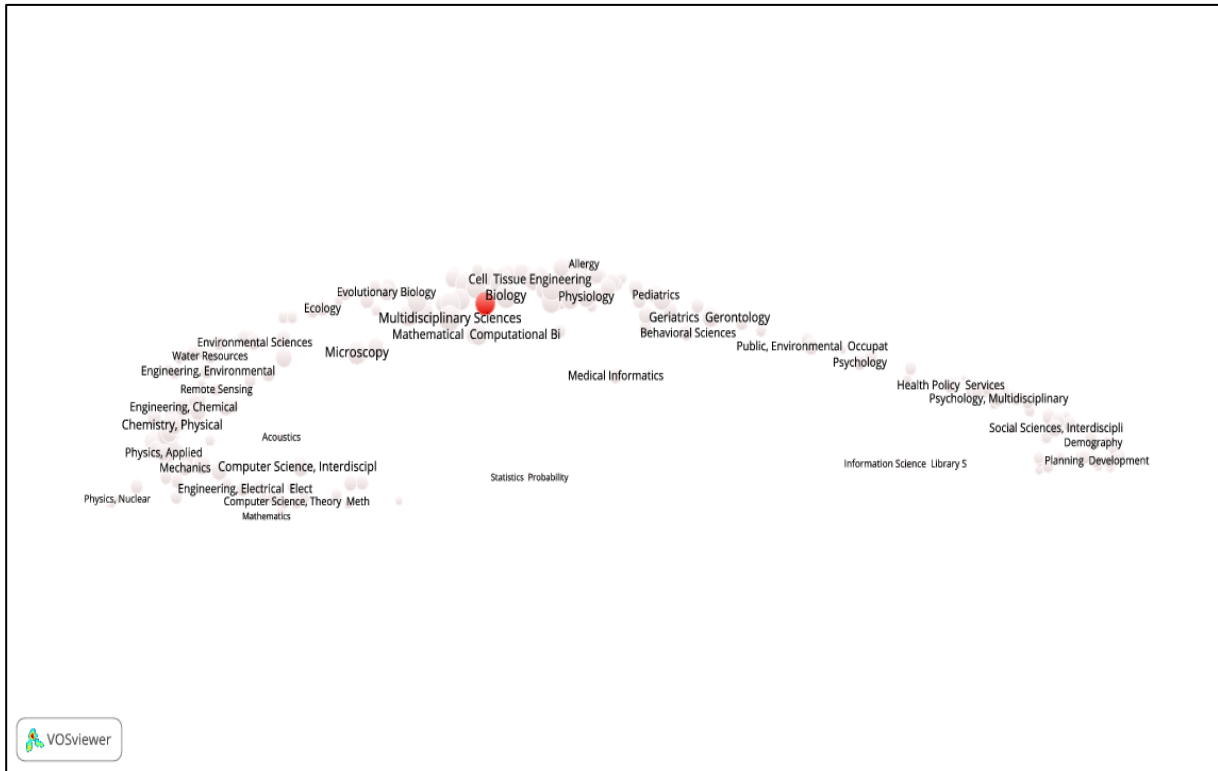


Figure 58. SAPV of the BIOL-A research group's publications in WoS SCs similarity matrix

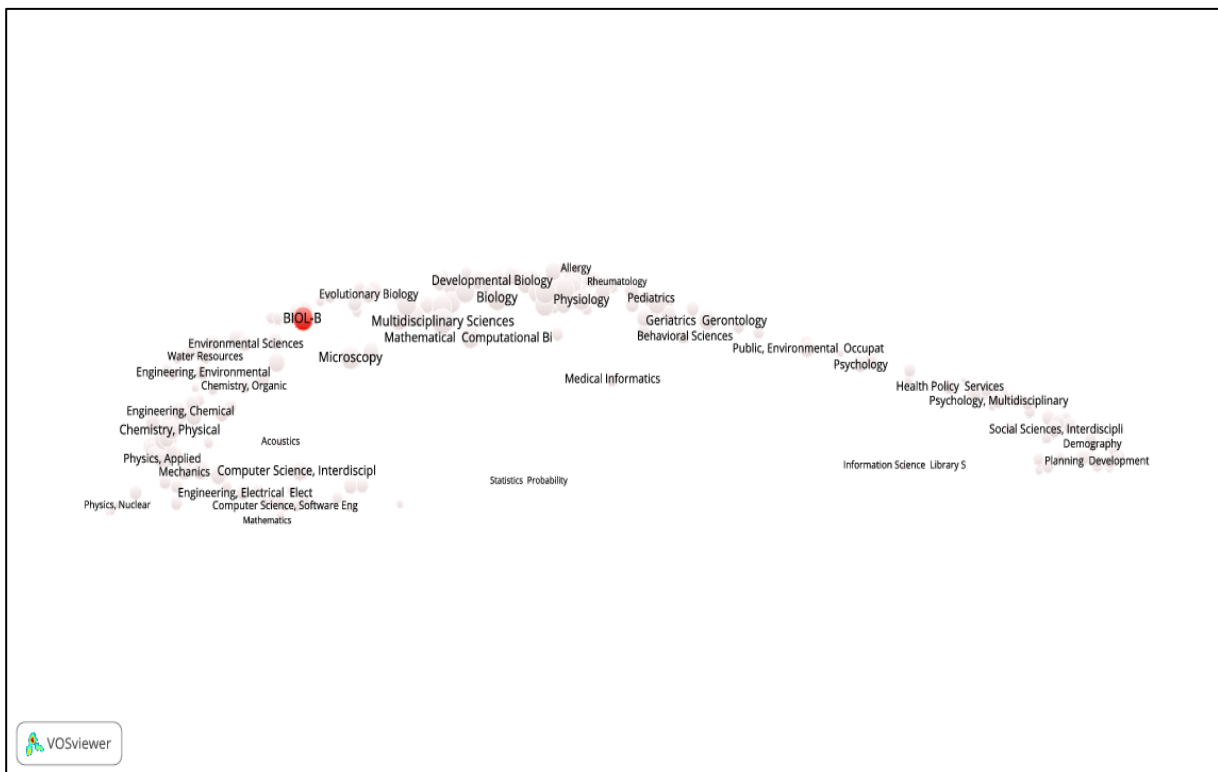


Figure 59. SAPV of the BIOL-B research group's publications in WoS SCs similarity matrix



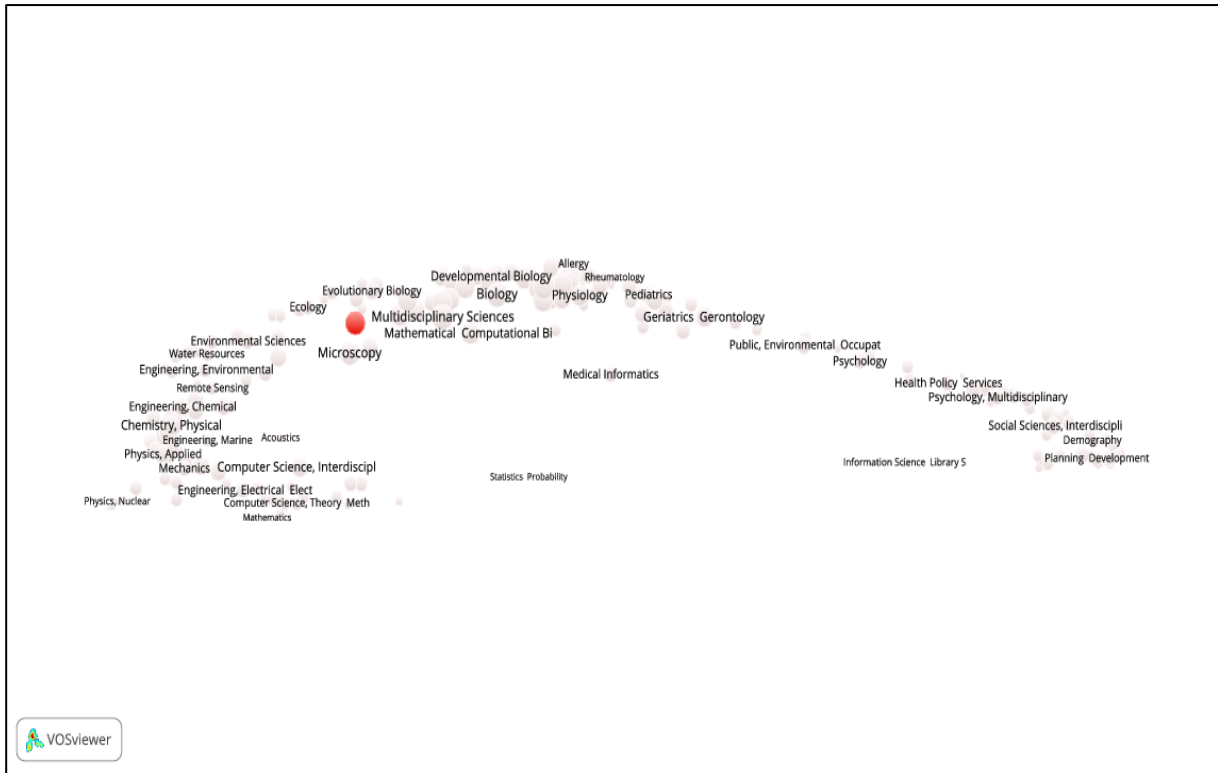


Figure 60. SAPV of the BIOL-C research group's publications in WoS SCs similarity matrix

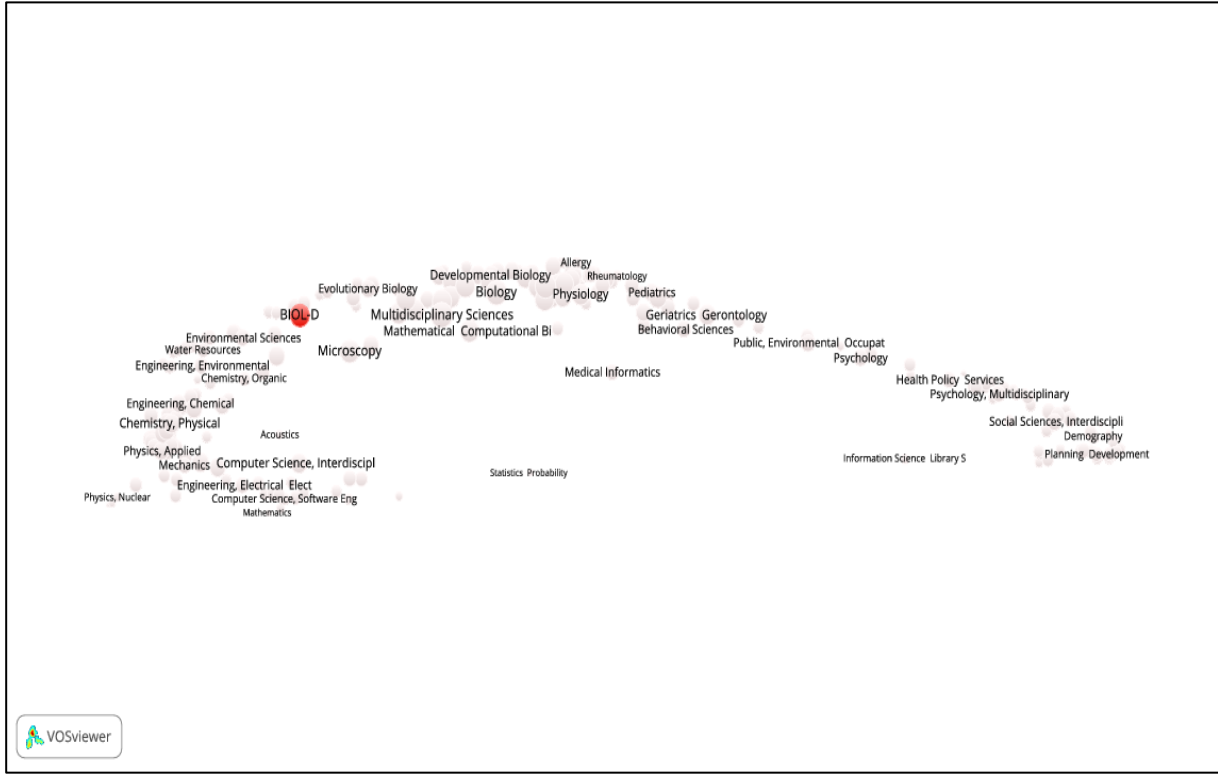


Figure 61. SAPV of the BIOL-D research group's publications in WoS SCs similarity matrix

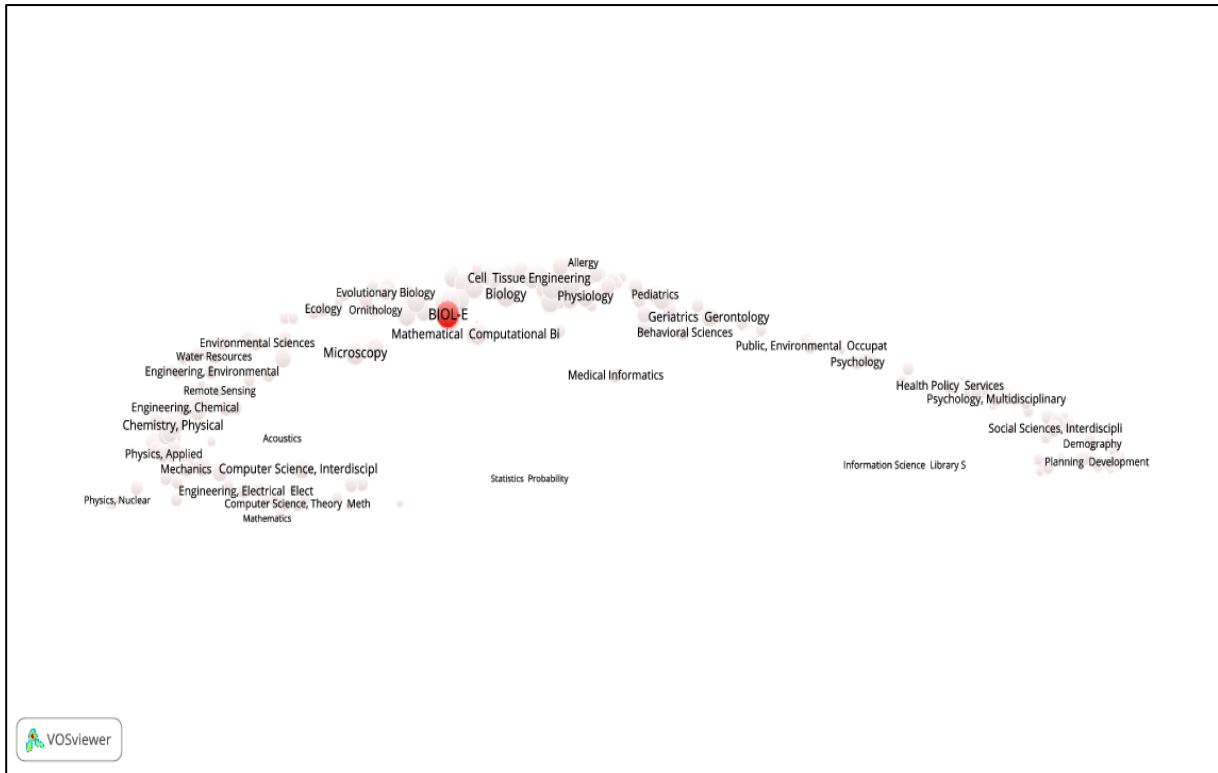


Figure 62. SAPV of the BIOL-E research group's publications in WoS SCs similarity matrix

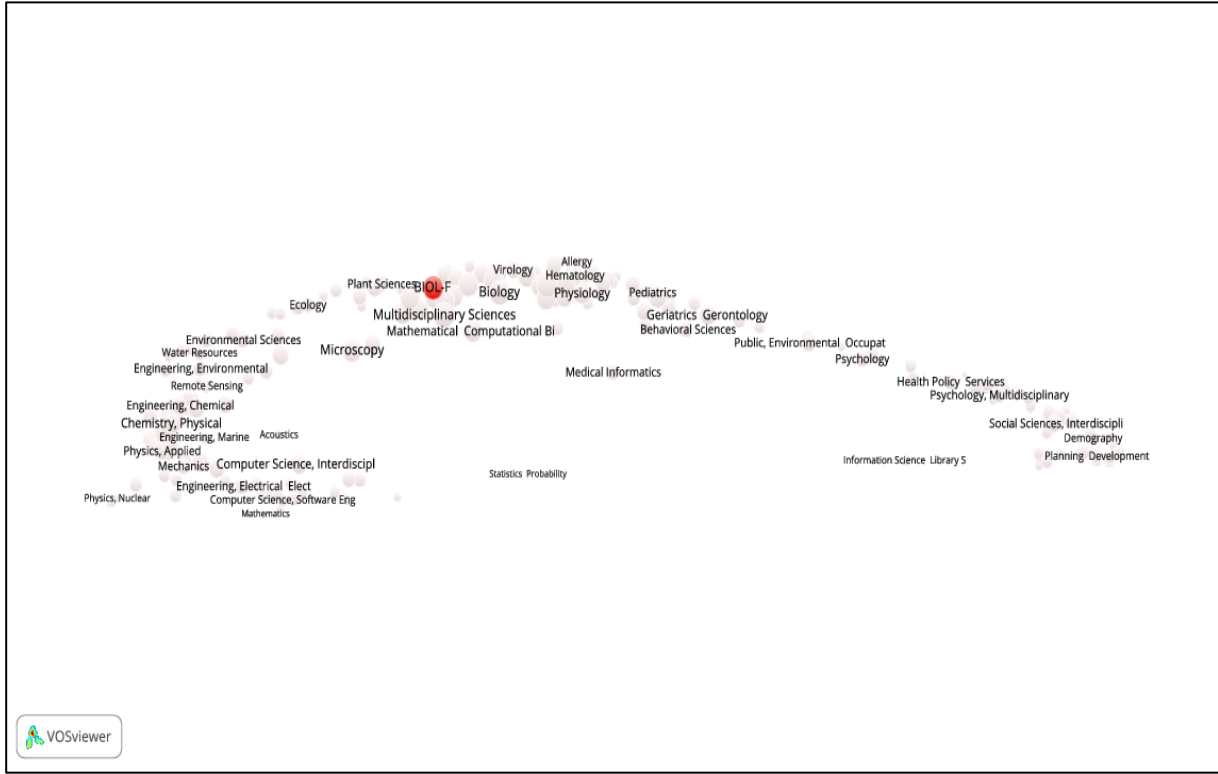
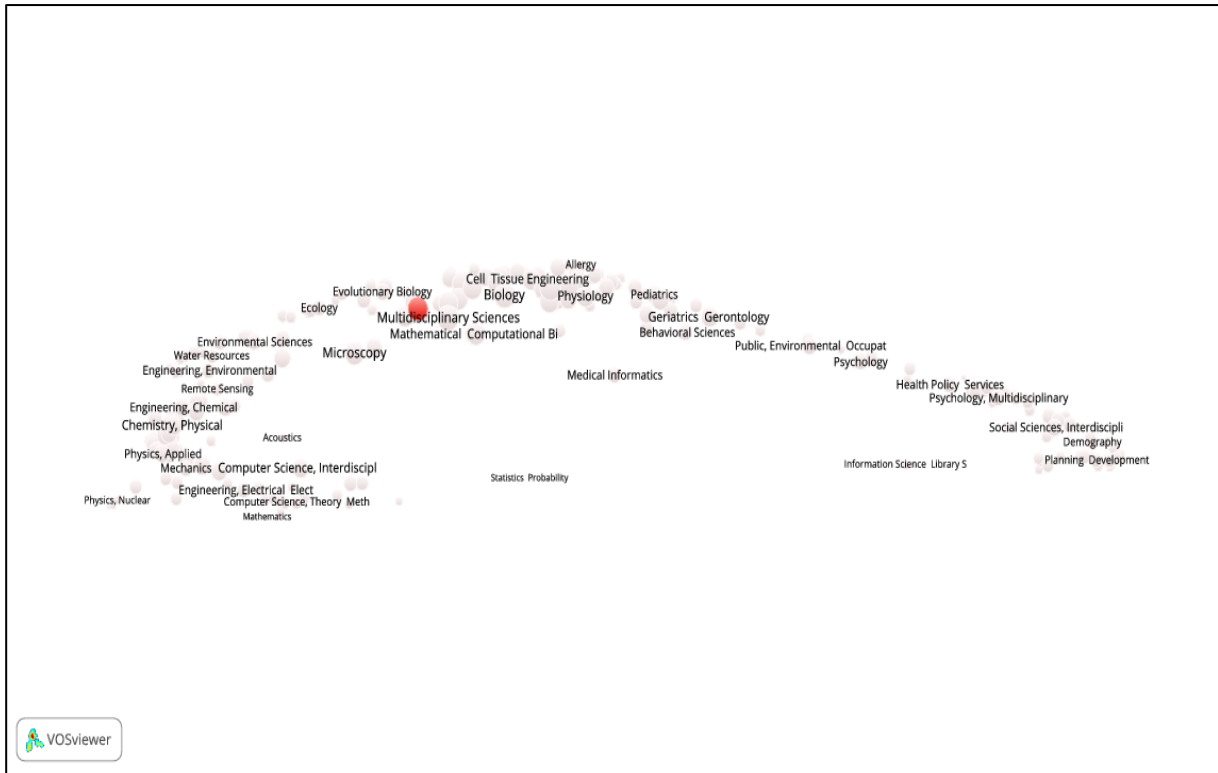
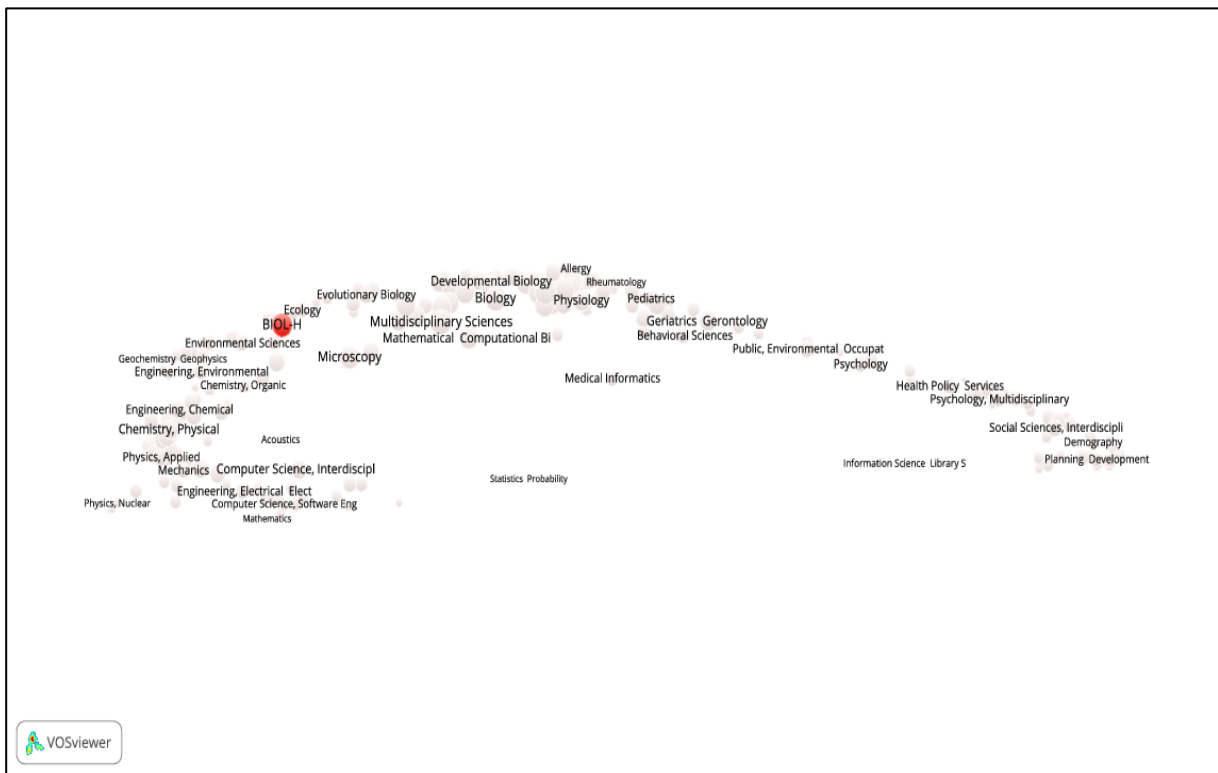


Figure 63. SAPV of the BIOL-F research group's publications in WoS SCs similarity matrix



**Figure 64. SAPV of the BIOL-G research group's publications in WoS SCs similarity matrix**



**Figure 65. SAPV of the BIOL-H research group's publications in WoS SCs similarity matrix**

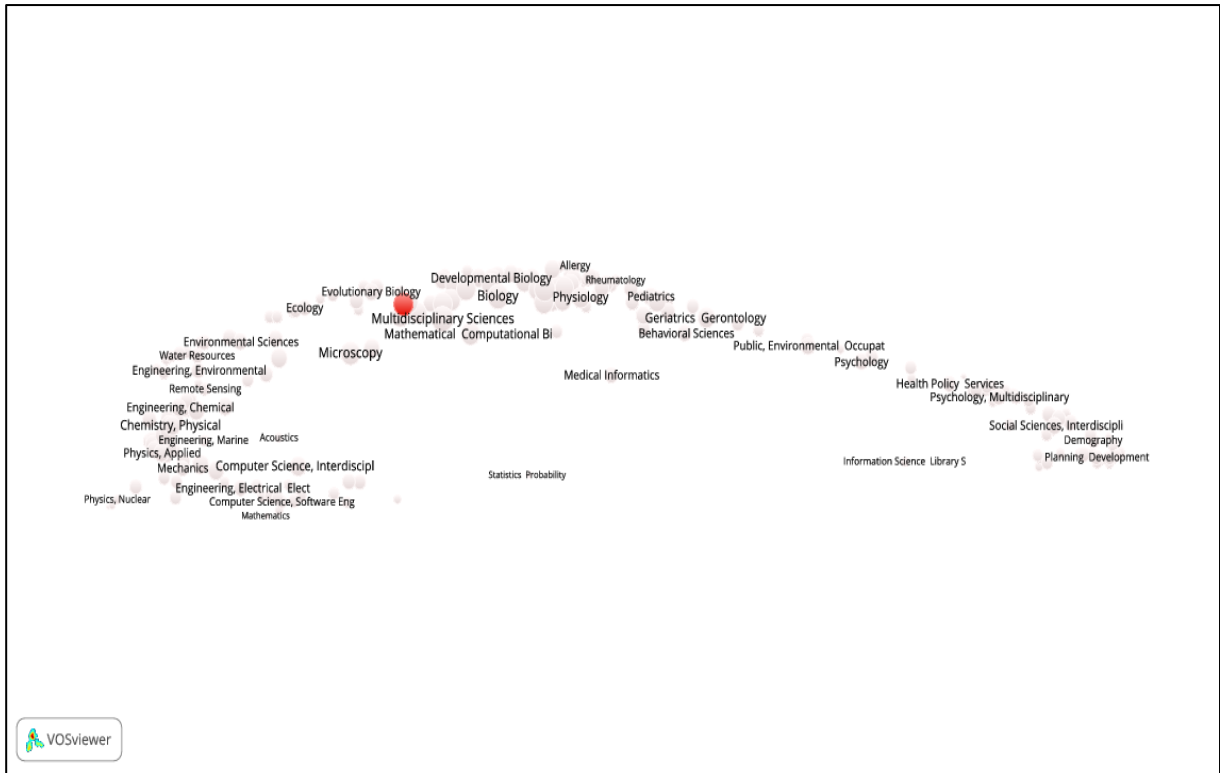
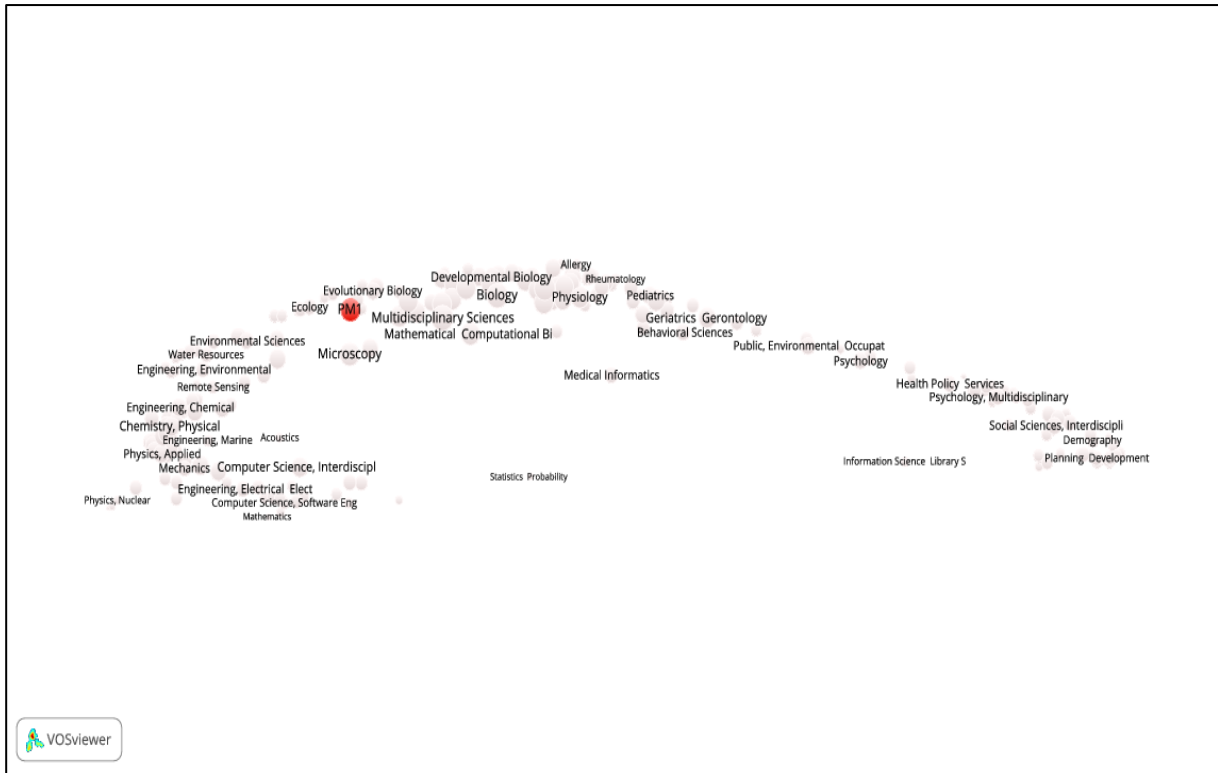


Figure 66. SAPV of the BIOL-I research group's publications in WoS SCs similarity matrix



Figure 67. SAPV of the Biology research group's publications in WoS SCs similarity matrix



**Figure 68. SAPV of the PM1's publications in WoS SCs similarity matrix**



**Figure 69. SAPV of the PM2's publications in WoS SCs similarity matrix**

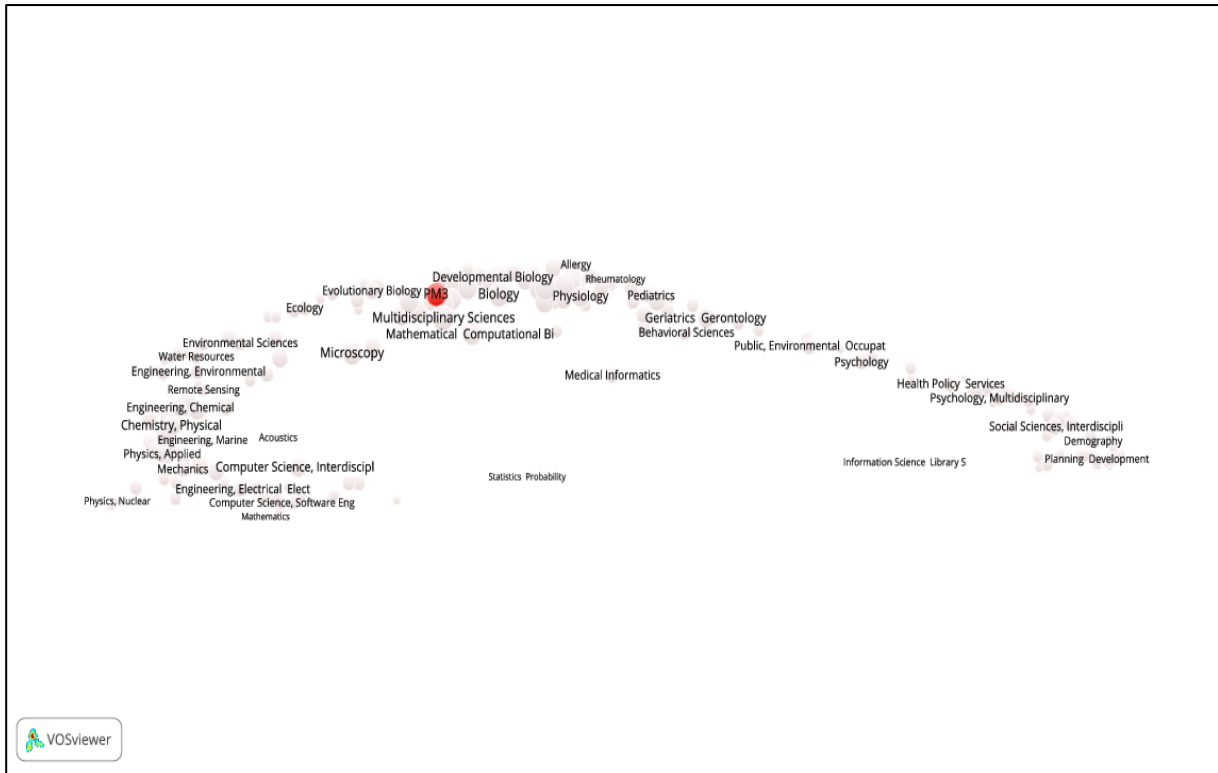


Figure 70. SAPV of the PM3's publications in WoS SCs similarity matrix

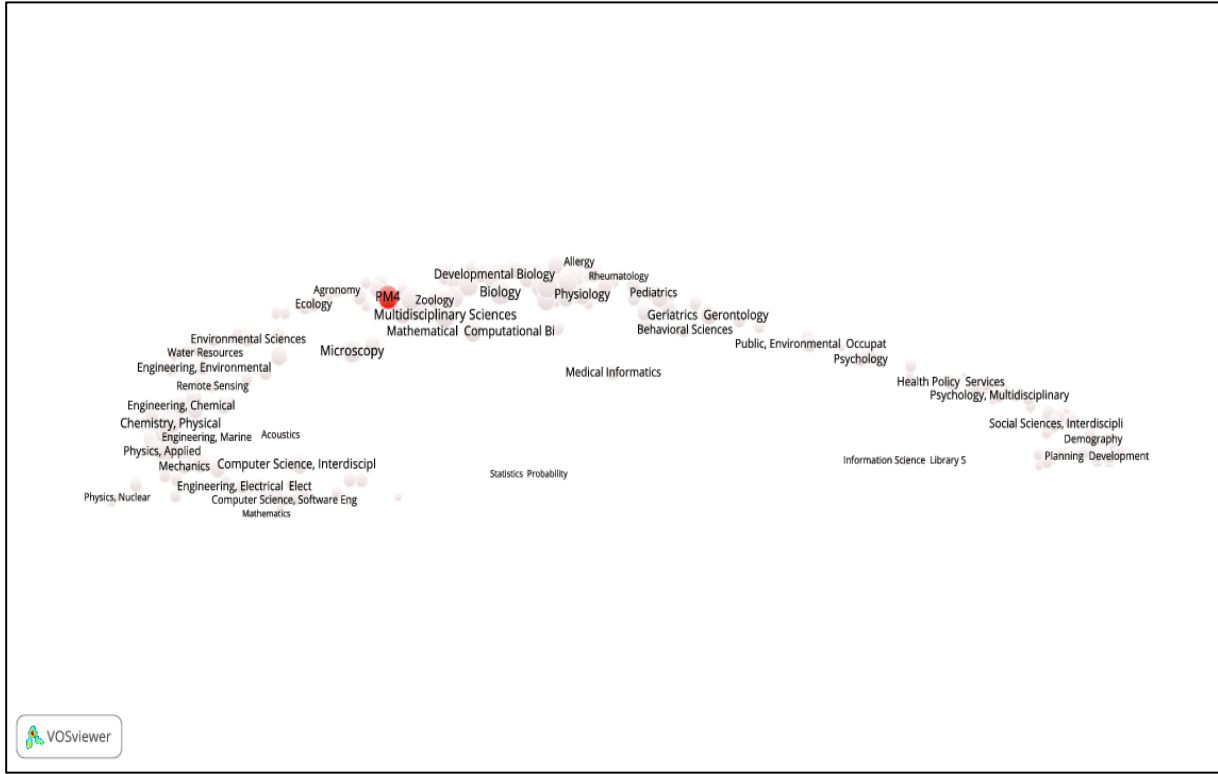
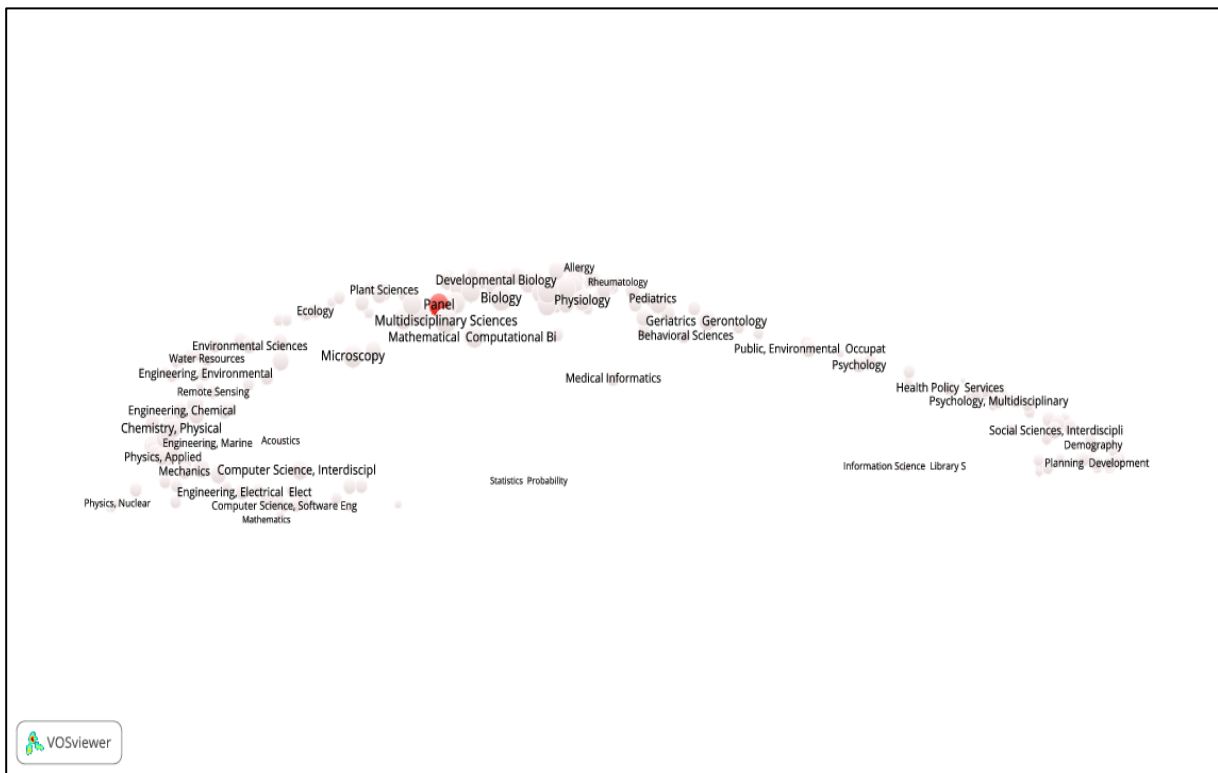


Figure 71. SAPV of the PM4's publications in WoS SCs similarity matrix

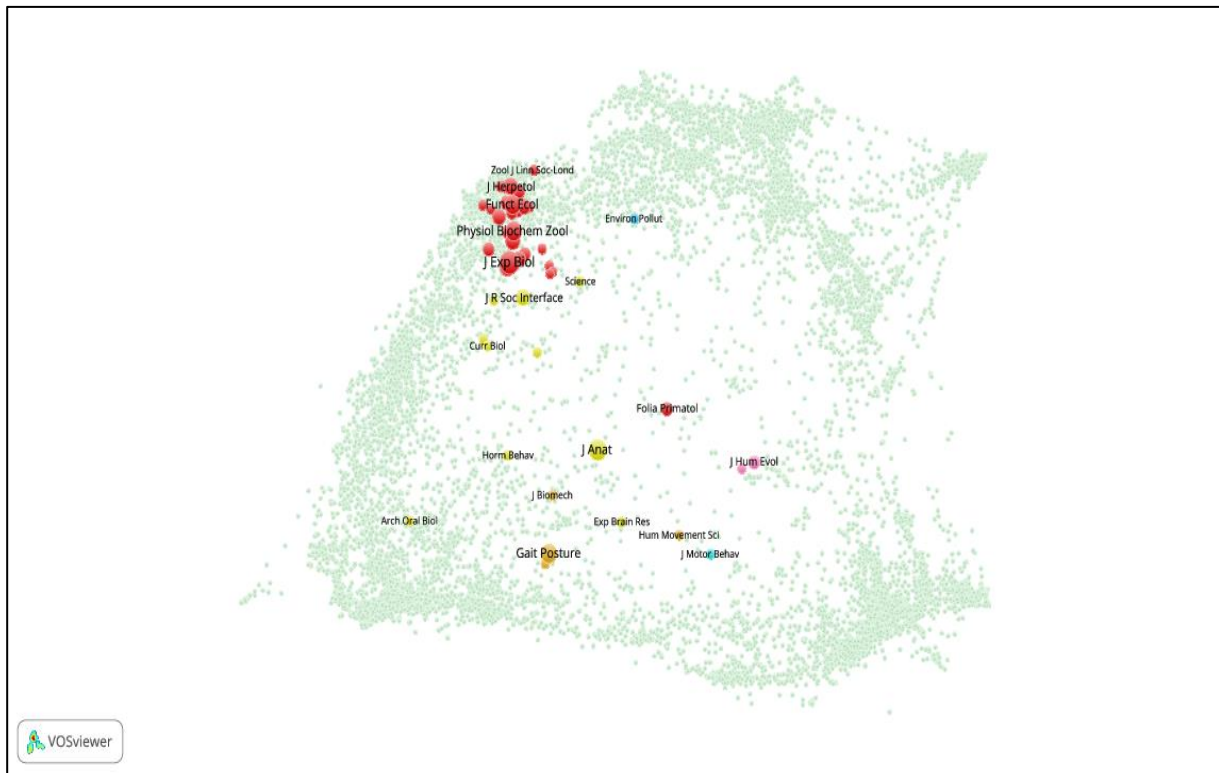


**Figure 72. SAPV of the PM5's publications in WoS SCs similarity matrix**

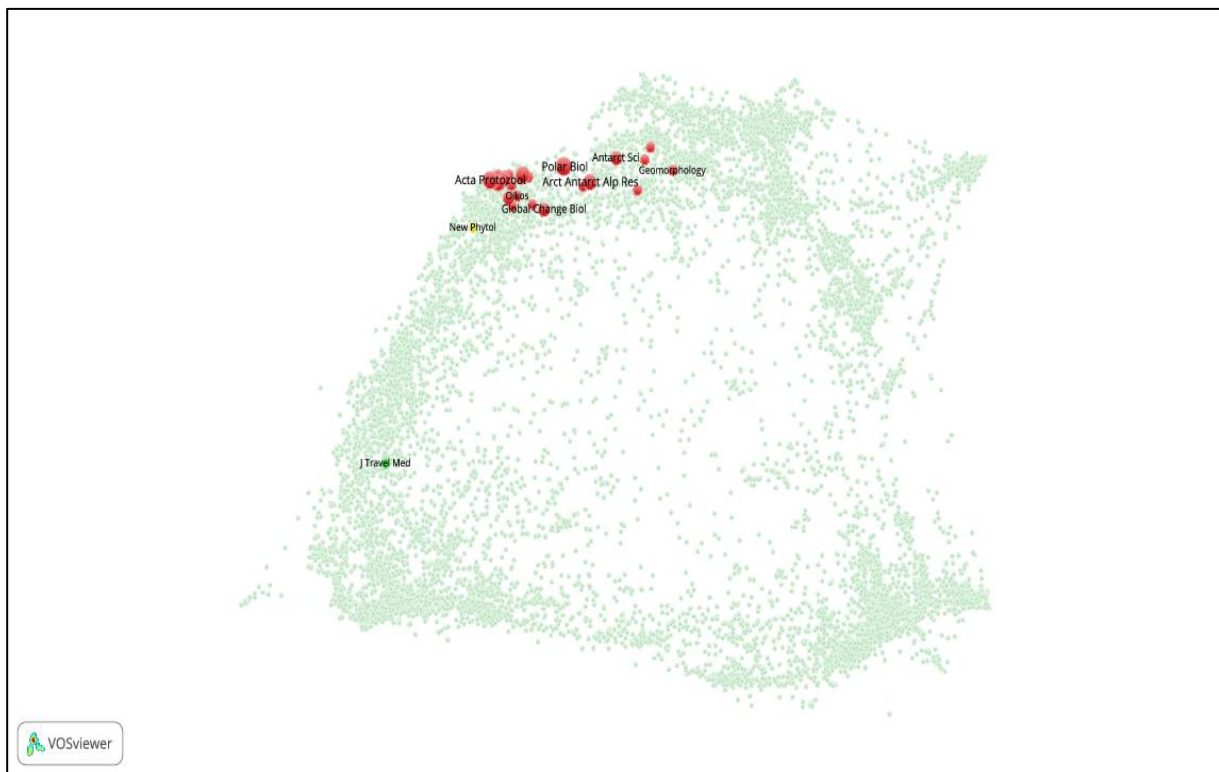


**Figure 73. SAPV of the panel publications in WoS SCs similarity matrix**

## Appendix C

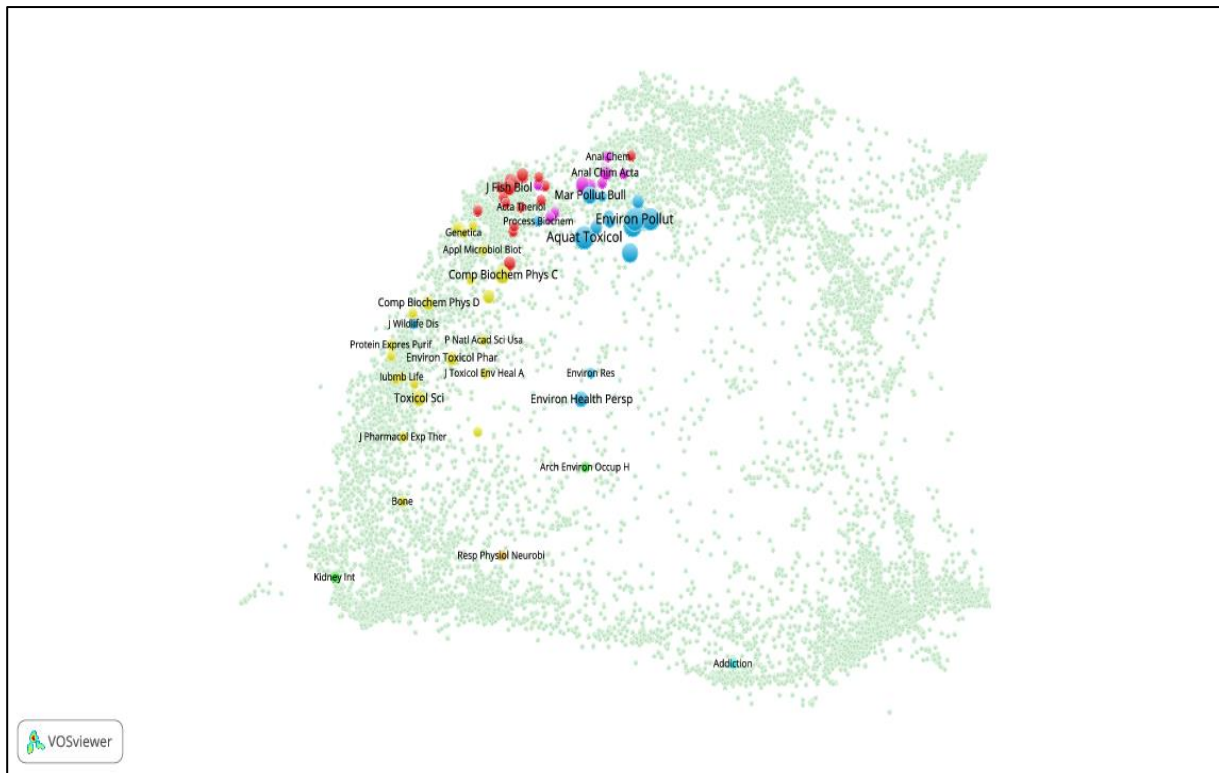


**Figure 74. Journal overlay map of BIOL-A research group's publications**

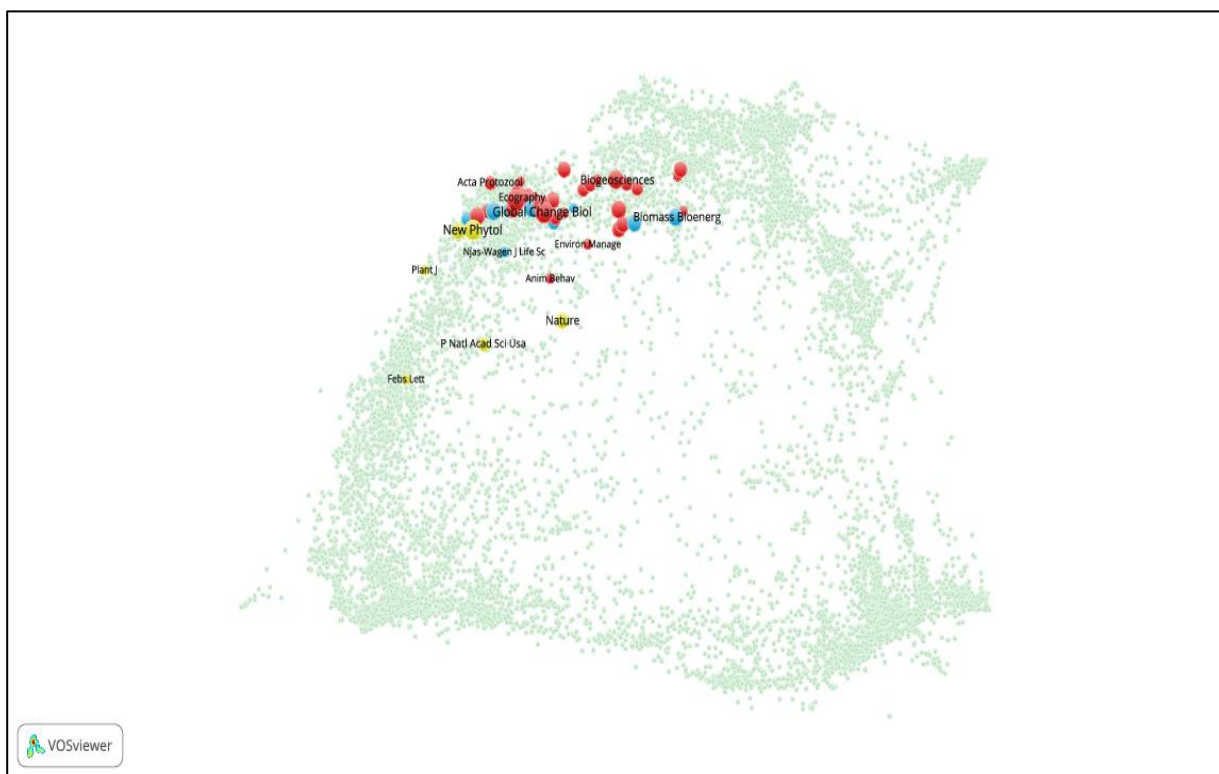


**Figure 75. Journal overlay map of BIOL-B research group's publications**

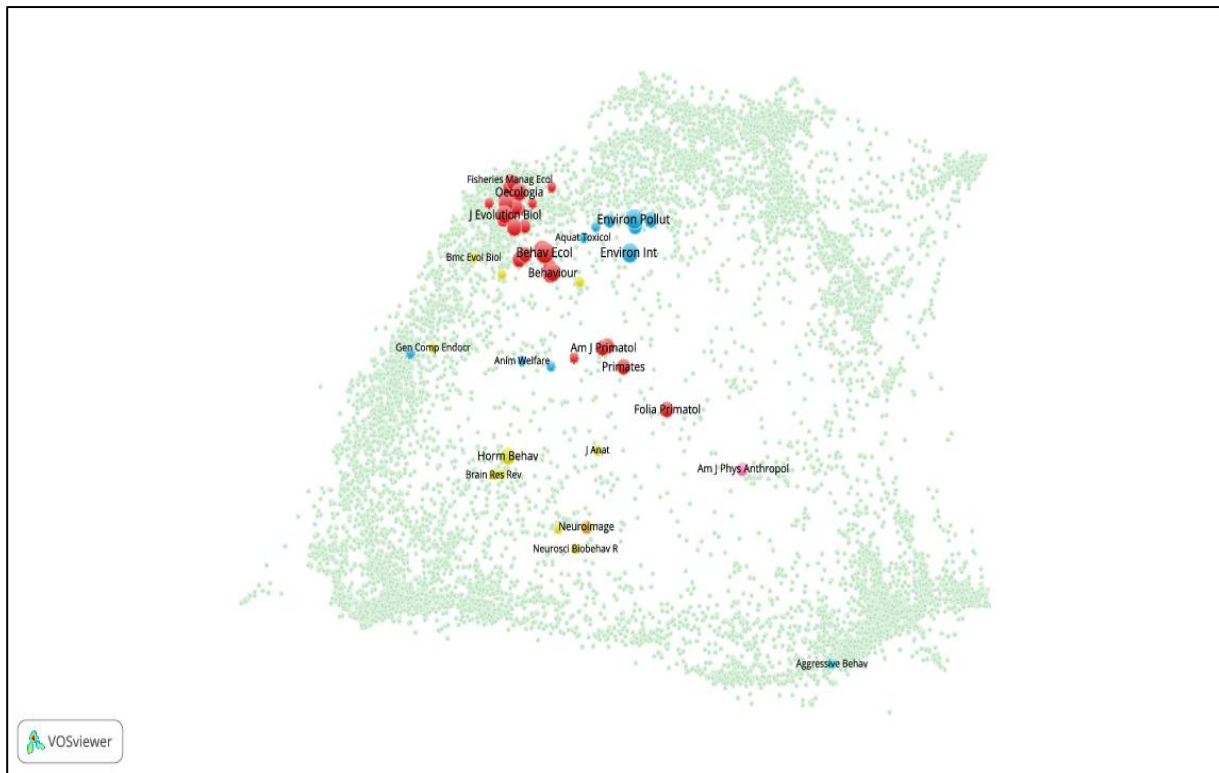




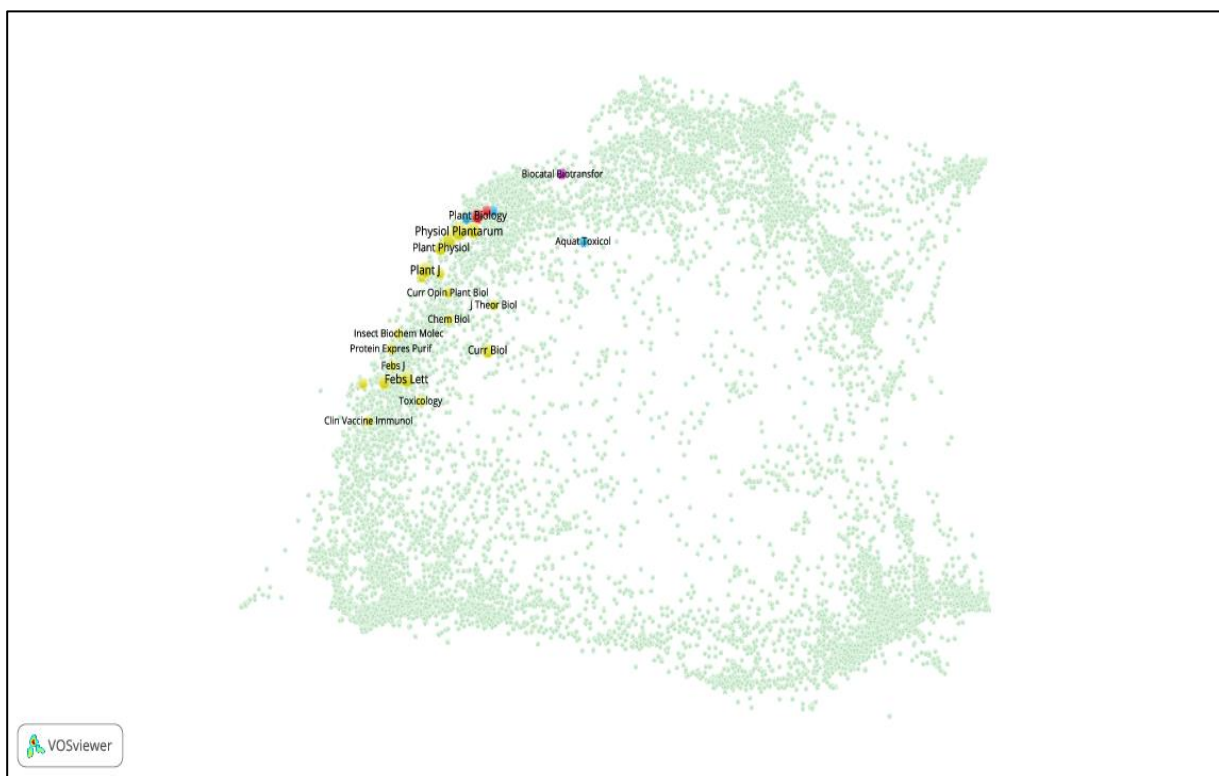
**Figure 76. Journal overlay map of BIOL-C research group's publications**



**Figure 77. Journal overlay map of BIOL-D research group's publications**



**Figure 78. Journal overlay map of BIOL-E research group's publications**



**Figure 79. Journal overlay map of BIOL-F research group's publications**

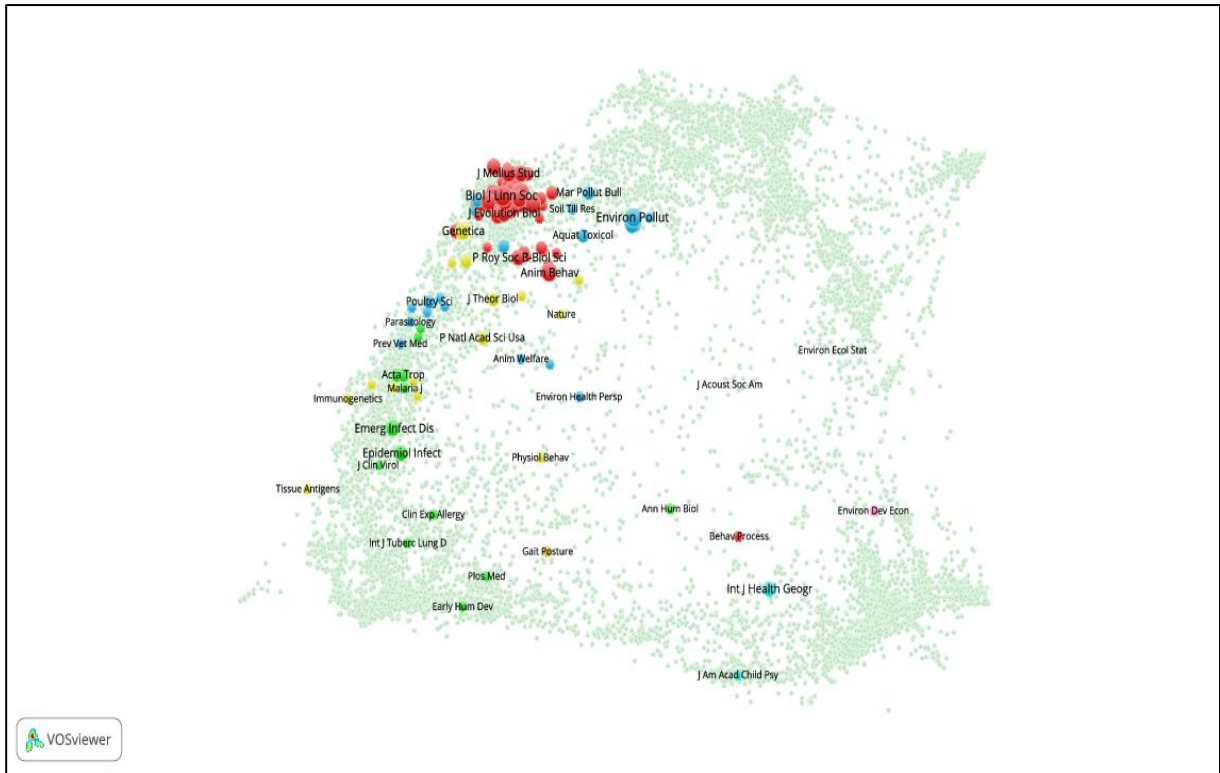


Figure 80. Journal overlay map of BIOL-G research group's publications

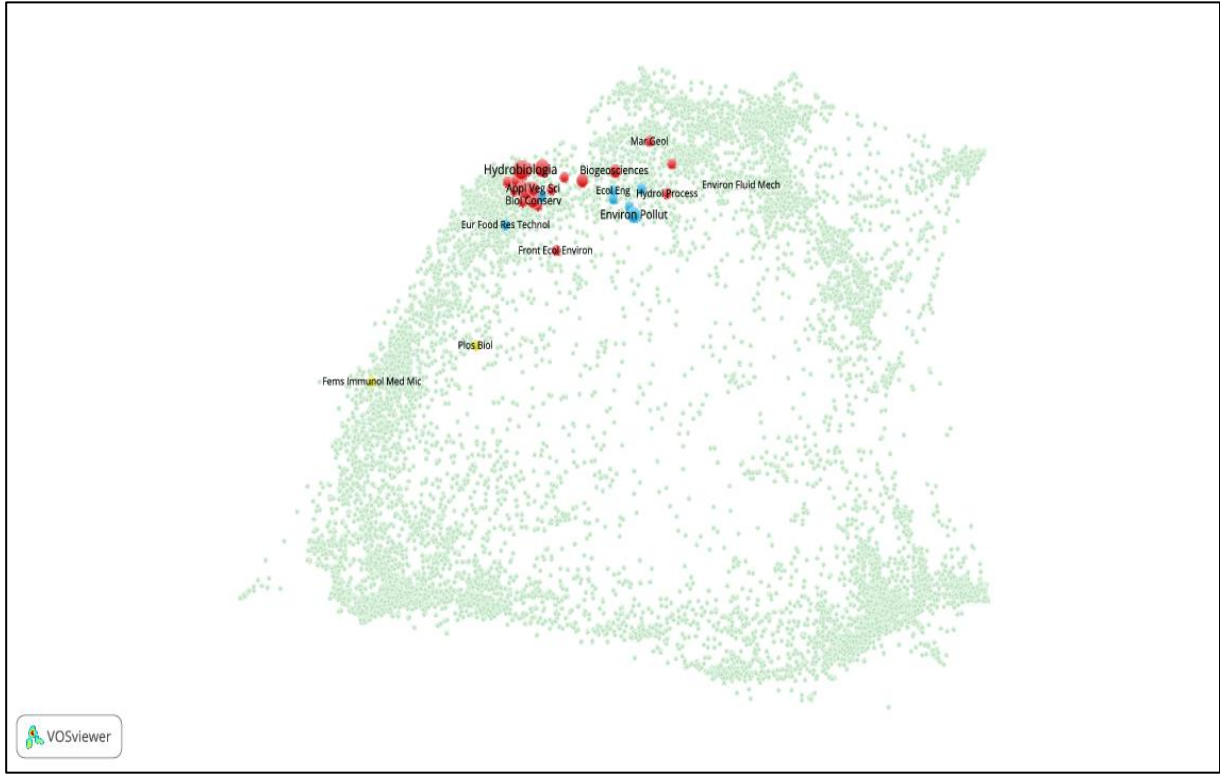


Figure 81. Journal overlay map of BIOL-H research group's publications

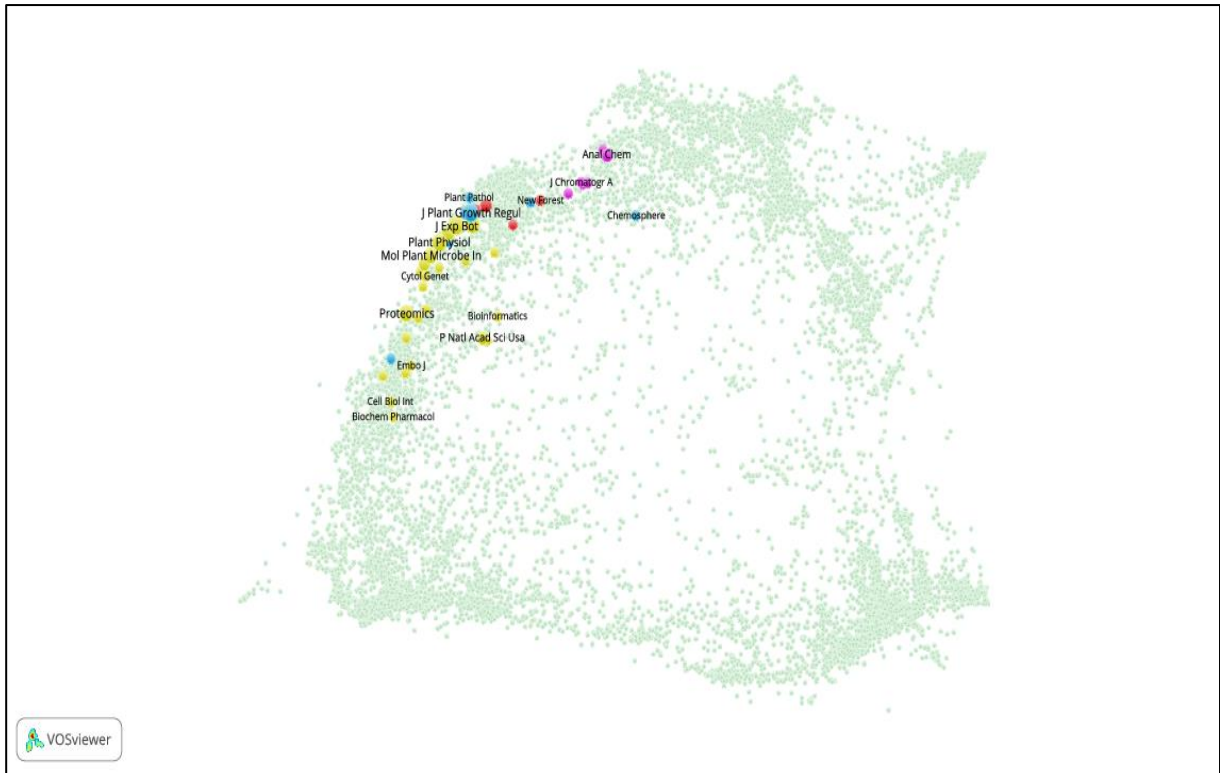


Figure 82. Journal overlay map of BIOL-I research group's publications

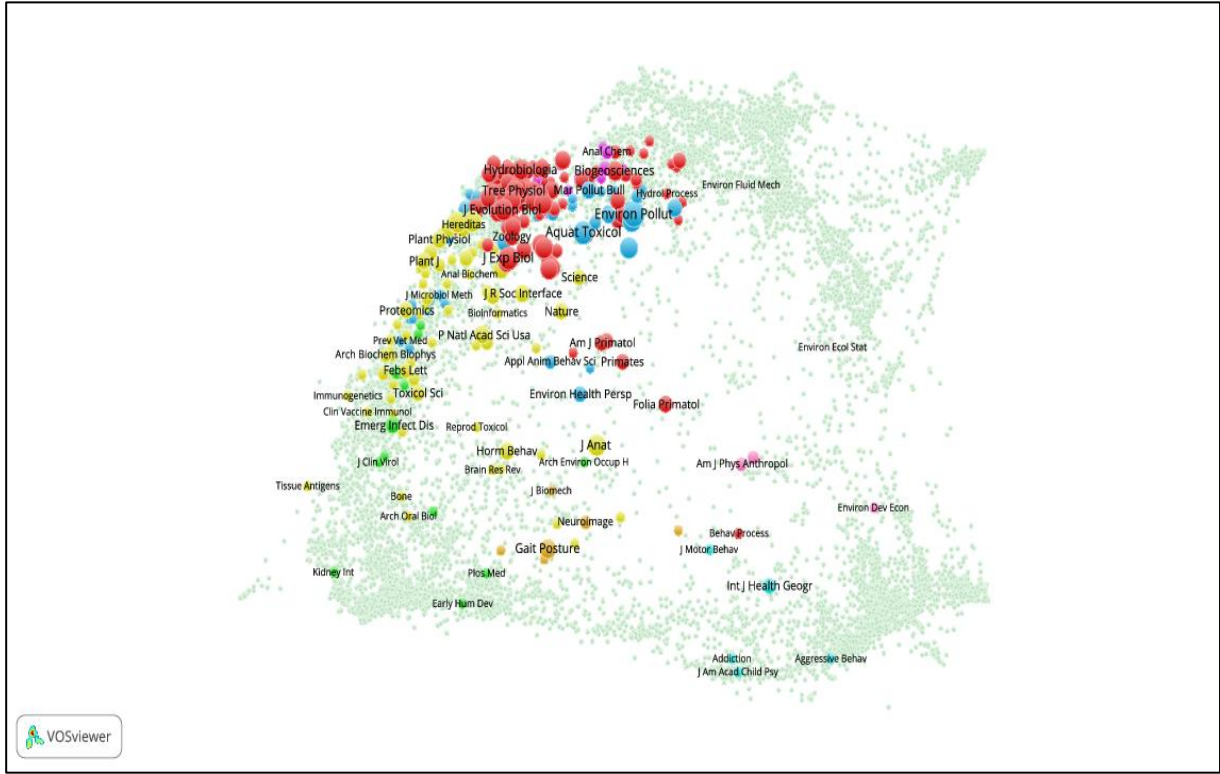
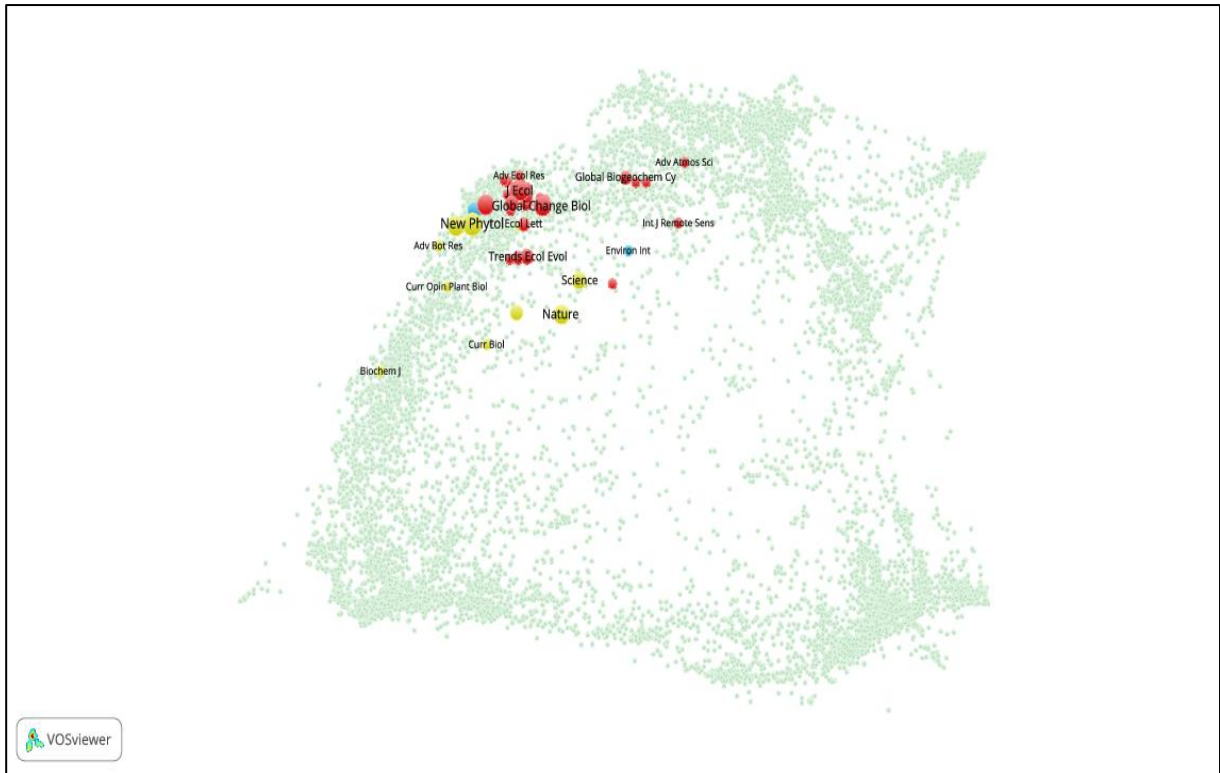
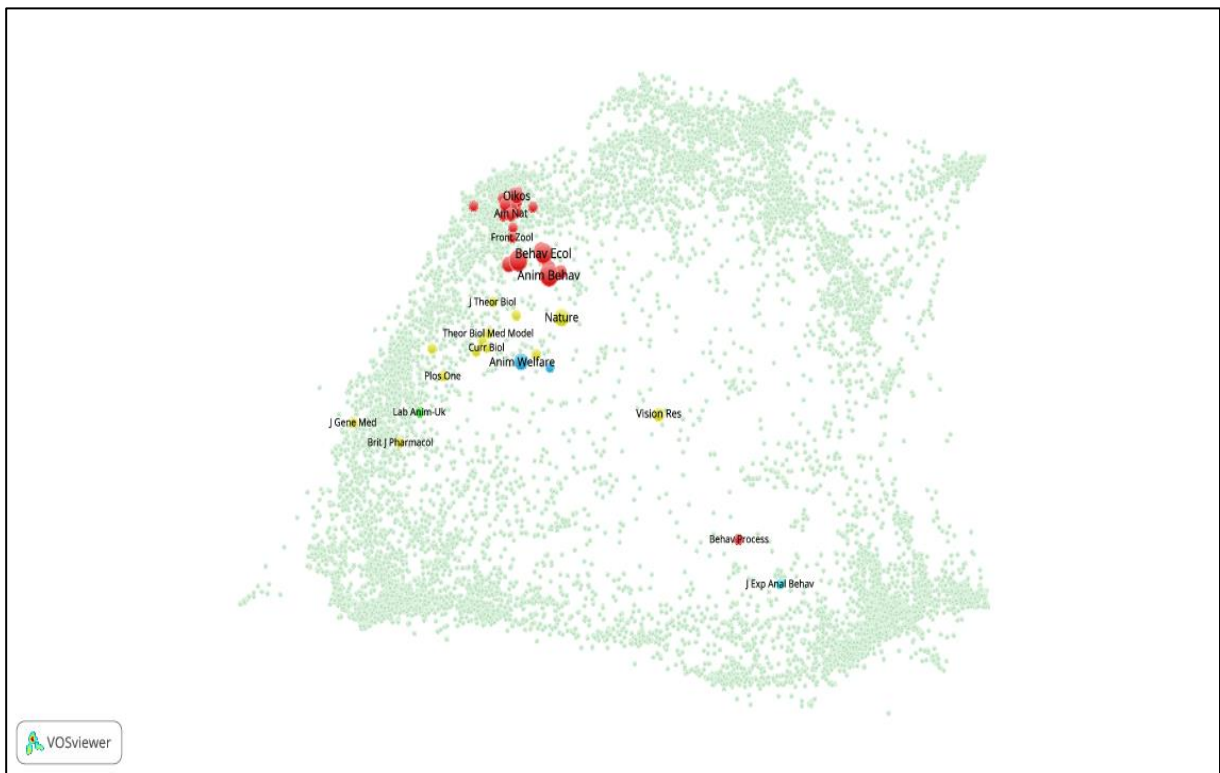


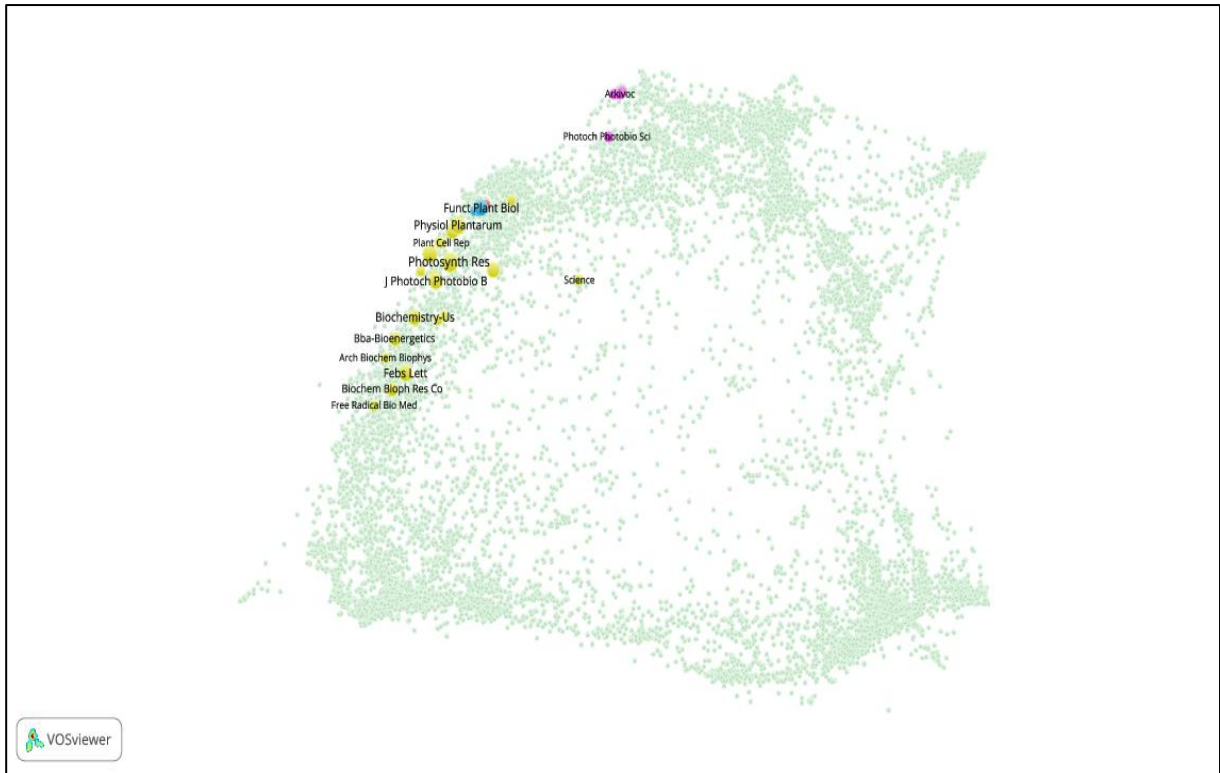
Figure 83. Journal overlay map of Biology research groups' publications



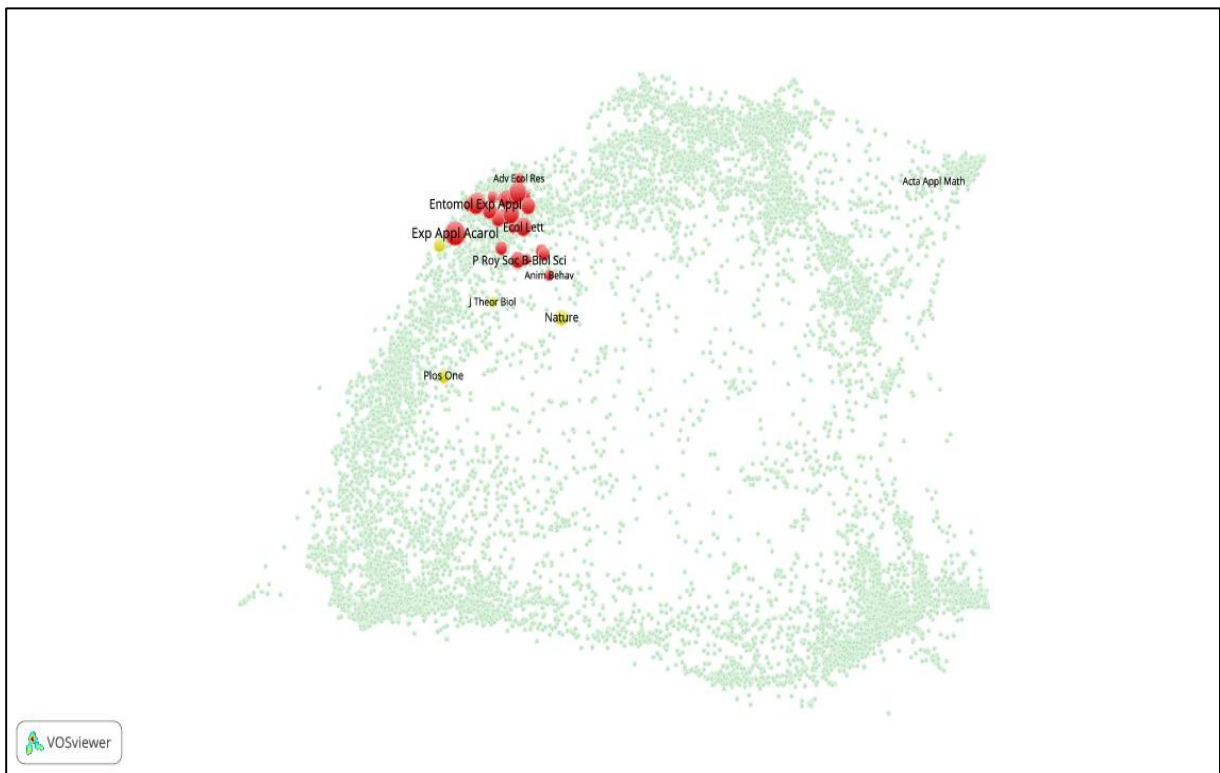
**Figure 84. Journal overlay map of PM1's publications**



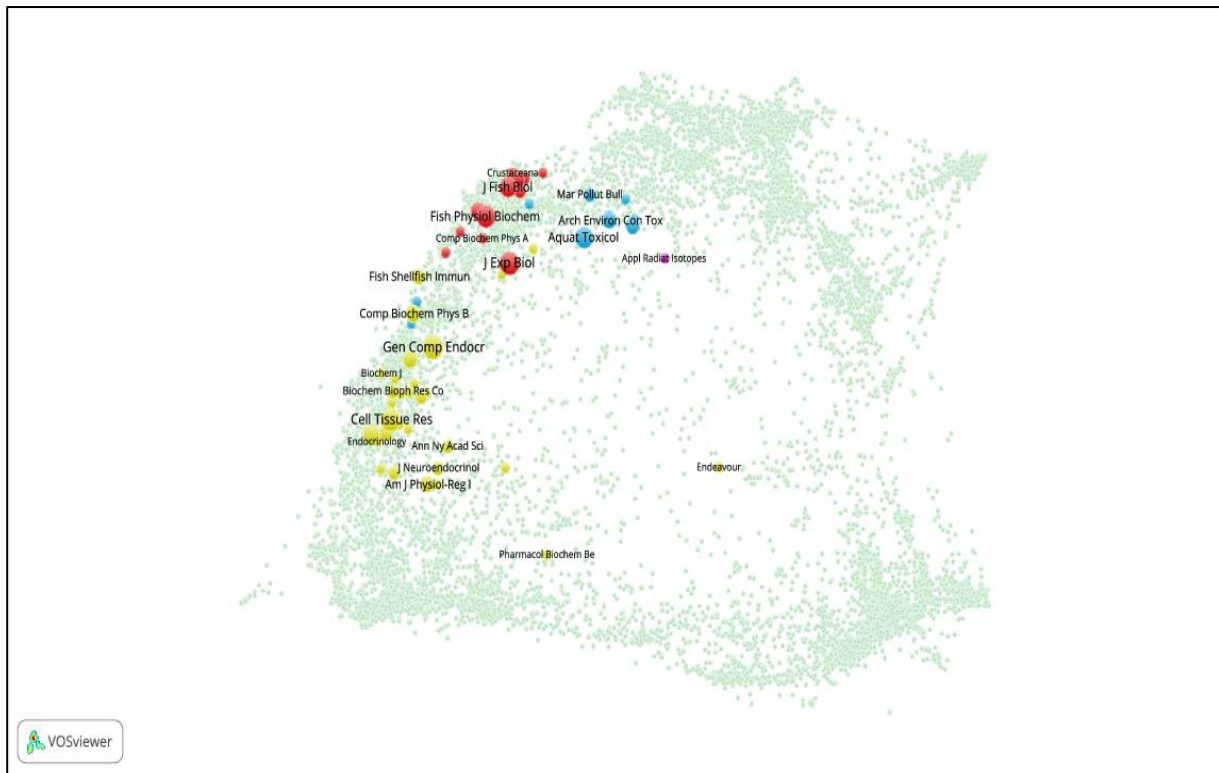
**Figure 85. Journal overlay map of PM2's publications**



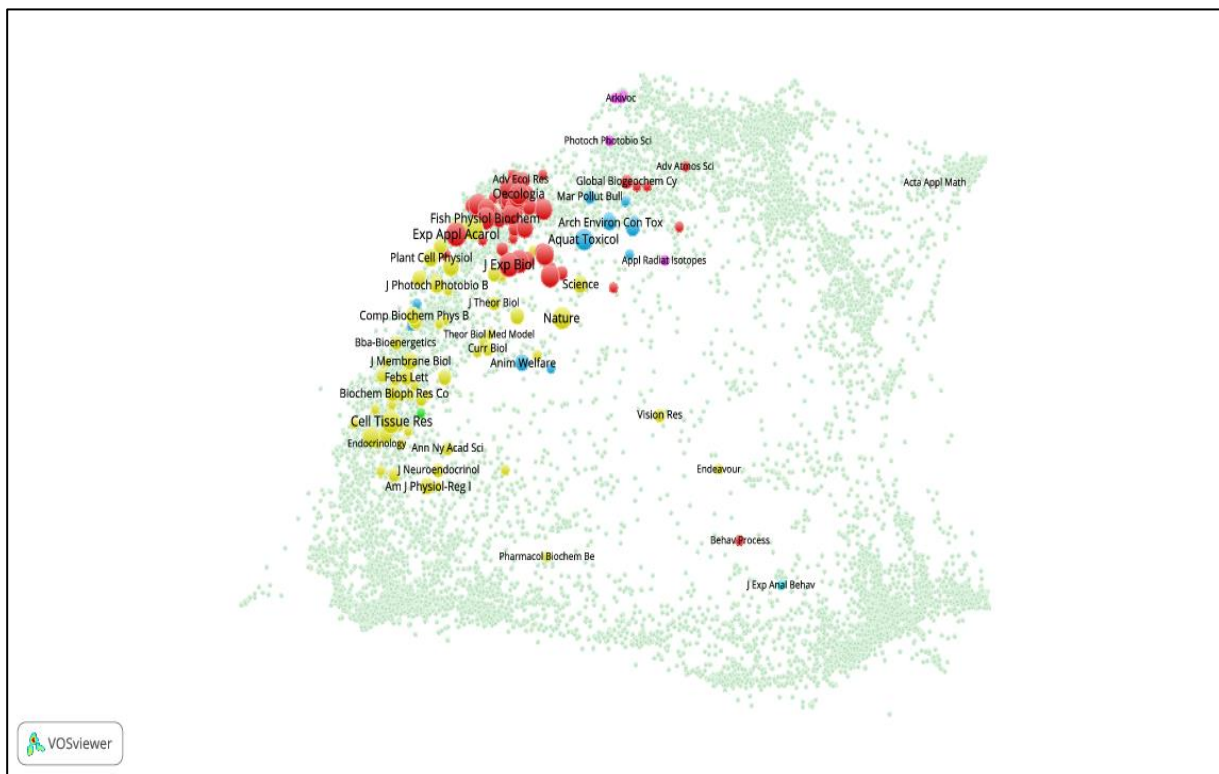
**Figure 86. Journal overlay map of PM3's publications**



**Figure 87. Journal overlay map of PM4's publications**



**Figure 88. Journal overlay map of PM5's publications**



**Figure 89. Journal overlay map of the panel's publications**