Supplementary Information

Bibliographic data and bibliometric analysis of the literature about "Seaweed Biorefinery" from WoS

Xueqian Zhang[a], Marianne Thomsen[a*]

[a]*Research Group on EcoIndustrial System Analysis, Department of Environmental Science, Aarhus University, Frederiksborgvej 399, Postboks 358, DK-4000 Roskilde, Denmark*

*Corresponding author. Tel.: +45 8715 8602; fax: +45 8715 5021. *E-mail address:* mth@envs.au.dk

Summary

This document presents the details of the bibliometric analysis (incl. descriptive analysis and co-word analysis) conducted in R based on the bibliographic collection obtained from Web of Science (WoS) database. The R scripts and bibliographic collection are provided in the file "R Scripts and Data of the Bibliometric Analysis.zip."

Contents

I. Retrieval of the bibliographic data

Bibliographic data were retrieved from WoS database by searching with the keyword "(macroalga* OR seaweed) AND biorefinery." Other settings of this query are summarized in the table below. This query gave a result of 132 hits, including article, proceedings paper, and early access. The bibliographic data of these 132 records were saved to plain text files as inputs for further analysis in R.

Table 1. Details of the query for obtaining literature about "seaweed biorefinery" from WoS databse.

Setting	Description
Data source	Clarivate Analytics Web of Science (WoS)
	Web of Science Core Collection
Database	- Science Citation Index Expanded (SCI-EXPANDED) --1900-present
	- Social Sciences Citation Index (SSCI) --1956-present
Query (Topic)	(macroalga* OR seaweed) AND biorefinery
Timespan	All years (1900 – 2019)
Document Type	Article, proceedings paper, and early access
Query date	28 January 2020

II. General introduction to bibliometric analysis

Bibliometric analysis is an objective quantitative approach to reviewing the literature. It provides evidence-based insights into the intellectual, social, and conceptual structure of a given research field based on a large body of bibliographic data. Moreover, this approach incorporates diverse techniques that can be employed to answer different research questions based on various statistical measurements, e.g., of science, scientists, or science activities in the field.

The uploaded R codes were based on R-package *Bibliometrix* (Aria & Cuccurullo 2017). Table below summarized all functions employed in the uploaded R codes. The first part of the R codes is intended to carry out a descriptive analysis of the collected literature, and the second part is to carry out a co-word analysis using keyword co-occurrence networks. The methodology and corresponding functions in R are introduced in the following sections.

Table 2. Summary of functions in the Bibliometrix R-package employed in this study.

Step	Function	Description	Output
Data loading and converting	readFiles()	Load a sequence of files exported from WoS into R	Bibliographic data frame
	convert2df()		
Descriptive bibliometric analysis	biblioAnalysis()	Create a bibliographic data frame	Tables of results
	summary () and plot()		
Network matrix creation	biblioNetwork()	Calculate co-occurrence/co-citation/collaboration networks	Network matrix
	networkStat() and summary()	Summarize the characteristics of the network	
Mapping	networkPlot()	Plot a bibliographic network	Plots of networks
Mapping and Data reduction	conceptualStructure()	Create a conceptual structure map of a scientific field using MCA and clustering	Network graphs

III. Descriptive Analysis by R-package *Bibliometrix*

The descriptive analysis provides an overview of the records and rankings of most productive authors and countries, top-cited manuscripts, and most relevant sources, based on attributes of the bibliographic data frame as a whole and individual documents.

IV. Co-word Analysis by R-package *Bibliometrix*

The co-word analysis uses keyword co-occurrence networks to identify prominent research themes and changes in the conceptual structure of the research field in time. The co-word analysis is based on the assumption that keywords act as signal-words reflecting the core content of the collected literature, and the co-occurrences of a pair of keywords indicate the similarity of the scientific themes to which these two keywords refer.

To explore the trends in research and innovations over time, we divided the whole timespan of our bibliographic collection (2011-2019) into two consecutive periods and grouped the bibliographic records of the publications accordingly. In order to get identical amounts of publications for two periods and given the fact that the number of publications found on WoS by using the keyword "Seaweed Biorefinery" was fewer years back, the first period was set to be seven years (2011-2017; 79 publications) and the second period was set to be two years (2018-2019; 53 publications).

For the bibliographic collection of each period, we extracted keywords from textual fields (i.e., Author Keywords, KeyWords Plus, titles, and abstracts; keywords include words or terms) and stored the extracted keywords as variables in a rectangular binary matrix A (Document × Keyword). Each row of matrix A was a document, and each column concerned a keyword from the bibliographic collection. Then we constructed a coupling-network based on the indicator matrix A by using the formulation (B=A×A^T) in R by applying the function biblioNetwork. The resulting matrix B is a symmetric co-occurrence matrix where the diagonal cells indicate the occurrences of the extracted keywords in the entire bibliographic collection, and the non-diagonal cells indicate the frequencies of two keywords occurring in the same publication.

Next, we applied cluster network analysis based on the co-occurrence matrix B to provide an intuitive picture of the conceptual structure of the research field, i.e., presenting in the form of a plot of clusters of nodes and edges; the nodes represent the extracted keywords, and the edges signify the co-occurrence of pairs of keywords. In R, this visualization step was realized by function networkPlot (settings are summarized in Table below). The clustering of keywords in the networks was conducted based on Girvan–Newman (GN) edge betweenness algorithm (Newman & Girvan 2004). In brief, edge betweenness of one edge measures the number of the shortest paths between pairs of nodes that run along with it. In a network containing different clusters that are loosely connected by a few inter-cluster edges, the edges connecting separate clusters are likely to have high edge betweenness as all the shortest paths from one cluster to another must travel through them. GN algorithm separates clusters from one another based on an iterative process of calculating the edge betweenness scores of each edge and removing the edges of the highest betweenness score to reveal the underlying community structure of one network. Ultimately, the connected nodes of the remaining network are the communities (Newman & Girvan 2004).

Table 3. Settings of the network plotting function networkPlot in R.

Settings	Description
normalize="association"	The vertex similarities are normalized using association strength
n = 30	The function *networkPlot* plots the main 30 cited references
type = "fruchterman"	The network layout is generated using the Fruchterman-Reingold Algorithm

weighted=NULL	With this setting, an unweighted graph is created and the elements of the adjacency matrix gives the number of edges between the vertices. If it is a character constant then for every non-zero matrix entry an edge is created and the value of the entry is added as an edge attribute named by the weighted argument.
cluster="edge_betweenness"	The type of cluster to perform is based on edge betweeness
remove.isolates=TRUE	Isolated vertices are not plotted.
size.cex=TRUE	The size of the vertices is proportional to their degree
size=10	The max size of the vertices is 10
curved=T	*networkPlot* function plots the edges with an optimal curvature
edgesize = 1	The thickness of the edges is proportional to their strength. Edgesize defines the max value of the thickness to be 1
labelsize=1	The size of vertex labels is fixed and defined to be 1
label.cex=F	The vertex label sizes are NOT proportional to their degree
label.n=20	Labels are plotted only for the main 20 vertices

Furthermore, to understand the structure of the research field of Seaweed Biorefinery and research themes, we evaluated the correlation between keywords and the diversification of the field by measuring the centrality of each keyword and the entire bibliographic collection. The degree of centrality of each keyword is represented by the size of the node in the network plots while other characteristics, including the degree of centrality of the network together with other network characteristics such as density, size, transitivity, and average path length, are summarized by employing function networkStat in a table. A node's centrality measures the degree of interaction of it with other nodes in the network. For the network, centrality is a measure of the strength of clusters' external ties of clusters to the others.

As a side note, the networks were presented in a Fruchterman-Reingold (FR) layout based on force-directed algorithms to facilitate easy viewing of the network edges and the clustering structures (Fruchterman & Reingold, 1991). Aiming for clarity and aesthetics of network edges, the FR algorithm minimizes the number of crossing edges and positions nodes in a manner that edges have approximately equal length.

Reference

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959-975.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129-1164.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Appendix I - Results of the Descriptive Analysis

This section presents the results of the descriptive analysis (see Section II and III) based on the bibliographic collection obtained from WoS (see Section I).

Table A1. An overview of principal attributes of the bibliographic collection.

Attributes	Amount
Documents	132
Sources (Journals, Books, etc.)	51
Keywords Plus (ID)	592
Author's Keywords (DE)	410
Period	2011 - 2019
Average citations per documents	21.08
Authors	488
Author Appearances	675
Authors of single-authored docu	1
Authors of multi-authored docum	487
Single-authored documents	1
Documents per Author	0.27
Authors per Document	3.7
Co-Authors per Documents	5.11
Collaboration Index	3.72

Table A2. Annual scientific production and average percentage growth rate.

Year	Articles
2011	2
2013	9
2014	7
2015	13
2016	19
2017	29
2018	20
2019	32
Annual Percentage Growth Rate	41.42

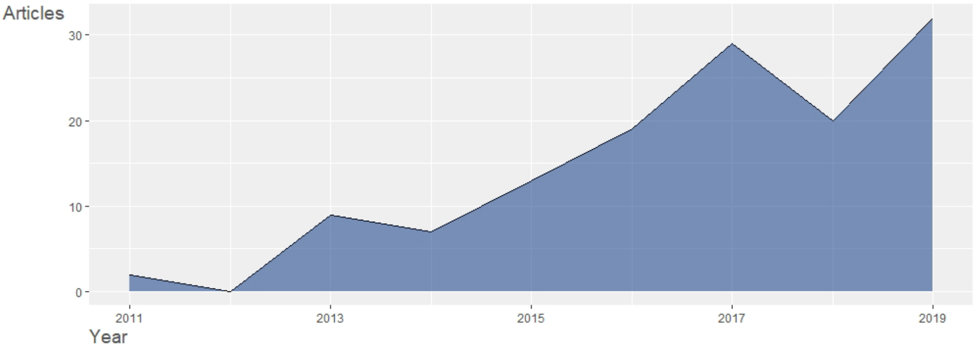

Figure A1. Annual scientific production over the years (2011 – 2019).

Table A3. Ranking of the most productive countries (of corresponding authors).

Ranking	Country	Articles	Frequency	Single Country Publications (SCP)	Multiple Country Publications (MCP)	MCP/SCP Ratio
1	UNITED KINGDOM	18	0.1374	12	6	0.3333
2	ISRAEL	15	0.1145	14	1	0.0667
3	DENMARK	11	0.0840	4	7	0.6364
4	INDIA	10	0.0763	7	3	0.3000
5	KOREA	10	0.0763	9	1	0.1000
6	AUSTRALIA	9	0.0687	7	2	0.2222
7	SPAIN	6	0.0458	4	2	0.3333
8	NETHERLANDS	5	0.0382	4	1	0.2000
9	SWEDEN	5	0.0382	5	0	0.0000
10	CHILE	4	0.0305	3	1	0.2500

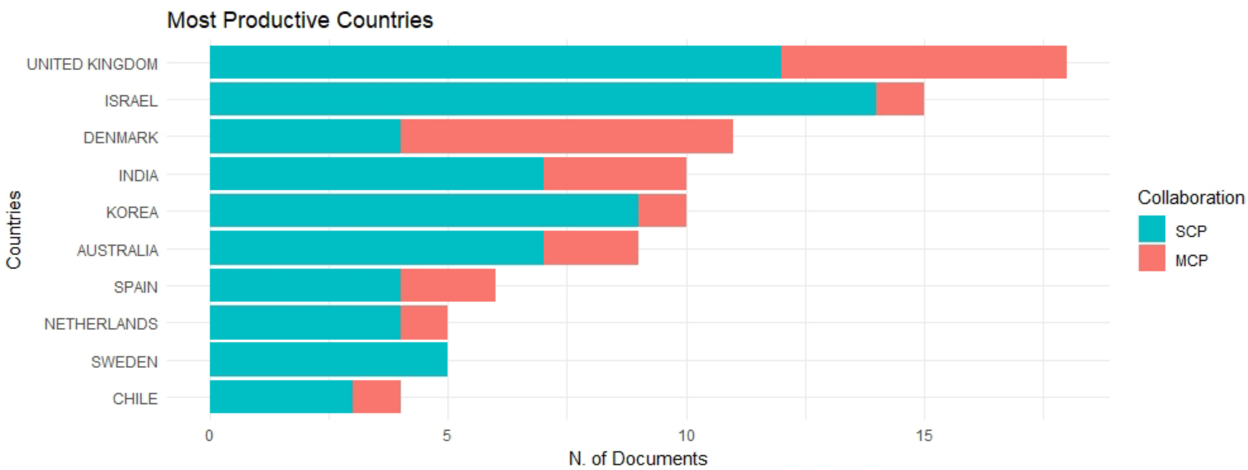

Figure A2. Most productive countries. (SCP: single country publications; MCP: multiple country publications).

Table A4. Ranking of the most productive authors.

Ranking	Authors	Articles	Authors	Articles Fracionalized
1	GOLBERG A	17	GOLBERG A	3.55
2	ISRAEL A	7	TEDESCO S	1.50
3	CHEMODANOV A	6	ISRAEL A	1.19
4	ALLEN MJ	5	BAGHEL RS	1.10
5	BAGHEL RS	5	REDDY CRK	1.10
6	CHUCK CJ	5	CHEMODANOV A	1.08
7	DE NYS R	5	LIBERZON A	1.04
8	MAGNUSSON M	5	ZHANG W	1.04
9	REDDY CRK	5	DANIELS S	1.00
10	ZHANG W	5	GIWA A	1.00

Table A5. Top manuscripts per citations.

Ranking	Manuscripts	Total Citation (TC)	TC/Year
1	HOLDT SL, 2011, J APPL PHYCOL	595	59.50
2	JUNG KA, 2013, BIORESOURCE TECHNOL	188	23.50
3	ADAMS JMM, 2011, BIORESOURCE TECHNOL	123	12.30
4	KUMAR S, 2013, BIORESOURCE TECHNOL	111	13.88
5	KERTON FM, 2013, GREEN CHEM	104	13.00
6	VAN DER WAL H, 2013, BIORESOURCE TECHNOL	103	12.88
7	FRANCAVILLA M, 2013, MAR DRUGS	63	7.88
8	BIKKER P, 2016, J APPL PHYCOL	57	11.40
9	DUMAN G, 2014, BIORESOURCE TECHNOL	55	7.86
10	BUSCHMANN AH, 2017, EUR J PHYCOL	55	13.75

Table A6. Most relevant sources.

Ranking	Sources	Articles
1	BIORESOURCE TECHNOLOGY	24
2	ALGAL RESEARCH-BIOMASS BIOFUELS AND BIOPRODUCTS	20
3	JOURNAL OF APPLIED PHYCOLOGY	11
4	MARINE DRUGS	5
5	BIOENERGY RESEARCH	4
6	BIOMASS & BIOENERGY	4
7	SCIENTIFIC REPORTS	4
8	APPLIED BIOCHEMISTRY AND BIOTECHNOLOGY	3
9	GREEN CHEMISTRY	3
10	JOURNAL OF CLEANER PRODUCTION	3

Appendix II - Results of the Co-word Analysis

This section presents the results of the descriptive analysis (see Section II and IV) based on the bibliographic collection obtained from WoS (see Section I).

Table A7. Characteristics of the word co-occurrence networks shown in Figure A3.

Attribute	Period 1 (2011 – 2017)	Period 2 (2018 – 2019)
Size	399	301
Density	0.034	0.043
Transitivity	0.348	0.385
Diameter	5	4
Degree of Centralization	0.263	0.267
Average path length	2.642	2.555

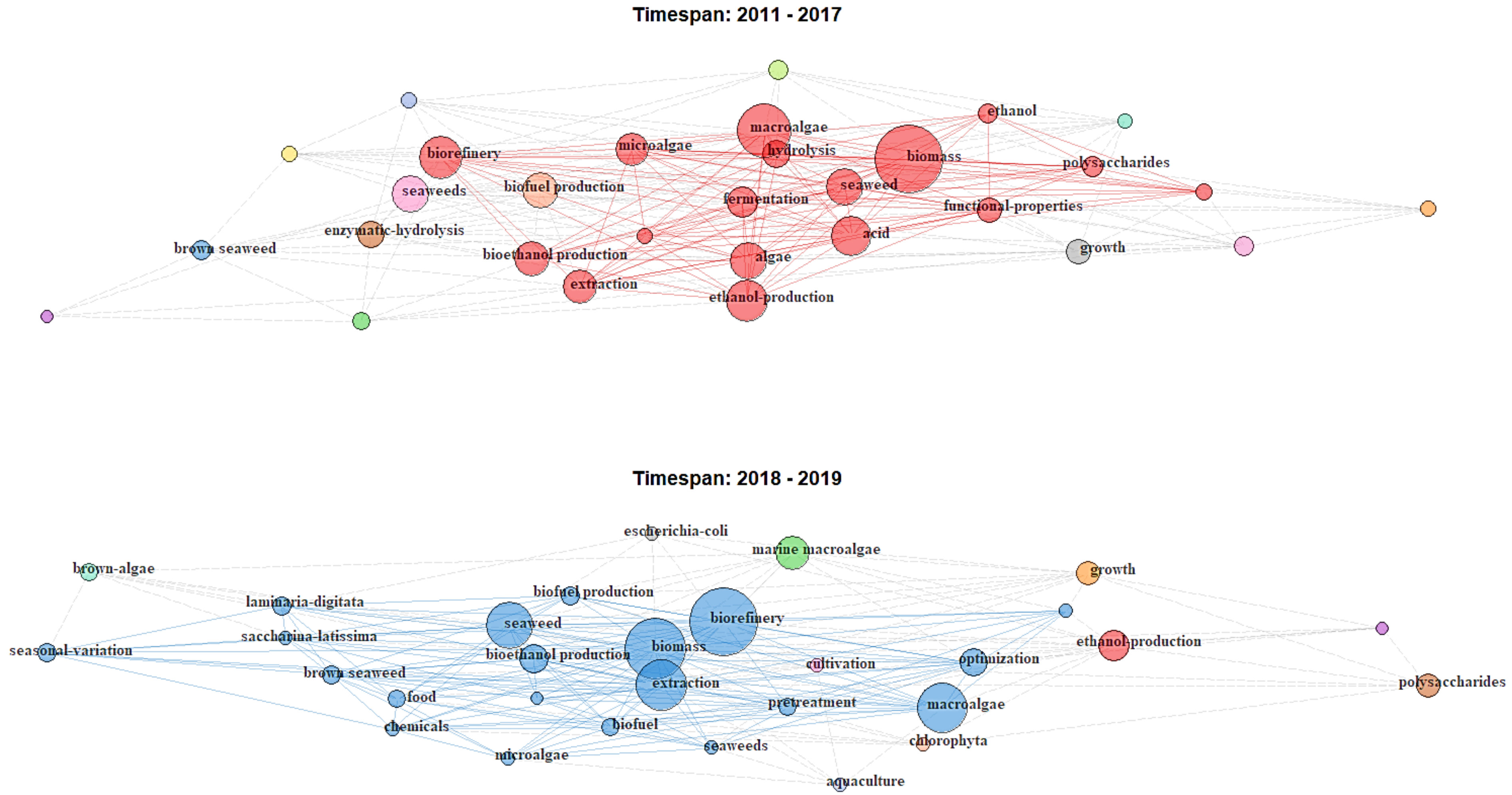

Figure A3. Keyword co-occurrence network based on publications in the period from 2009 to 2016 (upper) and keyword co-occurrence network based on publications in the period from 2018 to 2019 (bottom).