Text mining of abstracts from more than 30 years of publication of the *African Journal Marine Science.*

Dawit Yemane

2018-08-20 13:15:41

1 Summary

The *African Journal of Marine Science* (*AJMS*) (formerly the *South African Journal of Marine Science*), as the only dedicated marine-focused journal on the continent, is one route of dissemination of scientific knowledge for local and regional scientists. Although the aims and scope of the journal provide a rough idea of what types of publication to expect to find in it, the range of topics is wide. This work is intended to: (i) demonstrate (albeit briefly) some of the potential uses/applications of quantitative literature review; (ii) identify optimal sets of themes into which all published papers can be grouped; and (iii) summarise their temporal patterns. This was done based on all the abstracts of publications in *AJMS* since its first volume in 1983 up until this year. Topic modelling was applied to classify the abstracts into an optimal number of themes/topics. The result suggests the papers in *AJMS* can optimally be classified into 24 themes/topics. Over the last 35 years the numbers of papers within most of the themes/topics remained relatively stable, but there were some exceptions. For example, the theme on harmful algal bloom was largely patchy and with periodic up-ticks, probably related to the publication of special issues. Another exception was the theme of fish biology, in which there was a gradual increase in the number papers. Although this work is generic and demonstrates only a few aspects of text mining, it provides a glimpse of the potential value of text mining as a method of quantitative literature review.

2 Background

The review of existing literature is a starting point for all scientific studies as it provides a basis from which to identify gaps in knowledge (and hence to indicate what the new study intends to achieve) and to generate new hypotheses. It can even act as the main data source for a study (meta-analysis). Over time the total number of publications, in all scientific fields, has increased substantially, to a point where conducting an exhaustive literature review is becoming difficult. Currently, although not as a replacement but rather as a complement to the traditional literature review, the use of quantitative literature reviews (or text mining) is becoming a common occurrence (especially in the social and medical sciences). Quantitative literature review is broadly the combined use of automated text extraction and a range of machine learning (or data mining) tools to synthesize or extract relevant information from massive collections (ranging from tens of thousands to millions of publications) of abstracts or of the full content of published literature.

Most research is largely based on what is known as structured data, of which there is a range of examples in all scientific disciplines, e.g. catch data from surveys (including species composition and size structure), dose-response data from experiments, . . . etc. But there is a large amount of information/knowledge locked in written documents (including both scientific and non-scientific publications) that is known as unstructured data. In the past, information from unstructured sources was synthesized manually, which was fine when dealing with limited numbers of publications but almost impossible when dealing with thousands or millions of publications. Although it is not common in ecology (nor in many of its subfields), quantitative synthesis of text data to generate insight is relatively common in the social and medical sciences. This has largely been made possible by: (i) the availability of online data (from different sources, e.g. publishers, data- or publication repositories, and others) that can be accessed programmatically using various analytic environments (e.g. R, Python and others); (ii) text processing; and (iii) machine-learning algorithms that are now widely available.

The process of synthesizing/extracting of relevant information from text data is generally referred to as text mining. This usually includes a steps: (i) from programatically accessing of the required sets of text (e.g. lists of abstracts, email collections, tweets,. . . etc); (ii) preprocessing of the data (to exclude un-necessary characters/words); (iii) extracting characteristic sets of words; and (iv) applying range of modelling approach (e.g supervised and unsupervised classification, network modelling, . . . etc.) to address pre-specified sets of questions (Carlos and Thiago 2015). Some of the applications of text mining in ecology have included the following:

- In the biomedical field - Westergaard et al. (2017) analysed 15 million full-text articles published over the period 1823 – 2016 that they extracted from Elsevier, Springer and an open-access component of *PMC* (*Pub Med Central*). One focus of their work was to demonstrate the potential use of text mining to extract protein-protein associations, disease-gene associations and protein subcellular localisations from massive collections of published articles.
- In the field of ecology and evolution - Nunez-Mir et al. (2016) introduced the concept of Automated Content Analysis (ACA) and its potential in ecological and evolutionary research. *ACA* refers to the use of machine-learning tools for the qualitative and quantitative analysis of massive amounts of scientific literature.
- Nunez-Mir et al. (2017), in the study of biotic-resistance, used ACA to conduct a comprehensive literature review of biotic resistance in the context of invasion biology in forest ecosystems. An example of the type of observation to emerge from the use of ACA was that seedling survival and recruitment was a prominent topic.

The main aim of this work was to identify major themes across all papers published in the *(South) African Journal of Marine Science*, and trends in these themes over time. Abstracts published since the first volume of the journal in 1983 were collected from an online source and analyzed.

3 Methods

The data-analysis approach adopted here is commonly used for the analysis of text in various disciplines: e.g. quantitative literature studies; social-media text mining for marketing; biomedical science; social and economic studies; ecology; and others. The process of text mining starts with: (i) accessing the text data (usually done programmatically from providers that allow this); (ii) reading and processing text data (depending on the type of the text data, e.g. pdf, simple text, tweets – various methods of cleaning and processing are required); (iii) exploratory analysis (usually entails summarising the frequency of word usage and can be done in different ways, with the creation of word clouds being the most common), which provides an indication of the content of the collection of text; and (iv) depending on the study objective, the application of different types of modelling. Usually there is an interest in extracting from the collection of text data the underlying structure – when this exists – as sets of topics/themes. There are sets of models/algorithms that are commonly utilized for this purpose and that are generally referred to as *topic models* or *concept mapping models* (Ponweiser 2012; Nunez-Mir et al. 2016).

The most common topic model is Latent Dirichlet Allocation (LDA), which is widely used across a range of disciplines (Silge and Robinson 2016; Ponweiser 2012). Topic modelling is part of a class of classification models known as unsupervised classification methods. In principle, topic modelling is the same as the clustering methods applied to numerical data. LDA treats each set of documents (in this case the collection of abstracts of papers published in *AJMS*) as a mixture of topics and each topic as a mixture of words. This in turn allows documents to overlap in terms of their content. These aspects/principles of LDA are expanded on below (after Silge and Robinson 2016):

- Every topic is a mixture of words: For example, in the context of the Benguela system, one could think of a two-topic model in a paper that deals with different aspects of the system, with Topic1 being physical oceanography and Topic2, small-pelagic fish dynamics. Topic1 one could be described by words such as *upwelling*, *current*, *temperature*, *Agulhas Bank* and *wind*. Topic2, on the other hand, could be described by *sardine*, *anchovy*, *spawning*, *recruitment*, *temperature* and *Agulhas Bank*. The two topics

would then share words such as *temperature* and *Agulhas Bank*.
- Every document is a mixture of topics: One can think of each document as consisting of different topics in varying proportions. If we consider the above example of a two-topic model, we might note that Document1 consists of, say, 20% Topic1 and 80% Topic2. Similarly, all the remaining documents could be split compositionally into the different topics.

To use *LDA* one needs to specify the desired numbers of topics/themes into which to split the text collection. Thus one needs to first determine the optimal number of topics in the collection. There is a range of algorithms/metrics that one can use for this purpose: *Griffiths2004* (Griffiths and Steyvers 2004), *CaoJuan2009* (Cao et al. 2009),*Arun2010* (Arun et al. 2010),*Deveaud2014* (Deveaud, SanJuan, and Bellot 2014)). The automated method of topic selection employed here used the CaoJuan2009 algorithm in the *ldatuning* package (Nikita 2016) in R (R Core Team 2018). In addition a number R packages were used to read, process, analyze and visualize the results (Robinson and Silge 2018; Grün and Hornik 2017; Fellows 2014; Feinerer and Hornik 2018; Dahl 2016; Xie 2018; Ottolinger 2018; Wickham 2018).

3.1 Data analysis

The whole process from data collection to analysis can be summarised as follows:

- 1) The abstracts used in this analysis were extracted using a Mendeley desktop [http://www.mendeley.com/]. Once all the abstracts of papers published since the first volume of AJMS were collected, they were exported as a BibTeX file (which can be read in R).
- 2) The content of the .bib file was read and converted to a data frame.
- 3) Standard text-mining approaches were applied to first split each abstract into words, then to remove 'stop' words (in the jargon of text mining these are words that are unnecessary e.g. *is*, *was*, *are*, *to*, *and*, as well as numbers, punctuation, … etc.).
- 4) An exploratory analysis of the collection of words was conducted (e.g. single-word word clouds and pairs-of-words word clouds or bigram clouds)
- 5) Topic modelling was performed (this included selection of the optimal number of topics, fitting LDA, and extracting and summarising the results)

4 Result and discussions

Figure 1 shows the word clouds based on all the abstacts. It highlighs that in most of the publications the most common words are *species, south, coast, cape, data, distribution, management*. This suggests that most of the publications are from, or focused on, the marine environment of South(ern) Africa. This becomes even clearer when looking at the bigram word-cloud plot Figure 2. The bigram plot shows the frequency of occurrence of word pairs – words that occur next to each other. It can clearly be seen that most of the publications are South(ern) African-focused marine studies. But given that *AJMS* was initially the only local marine-focused research outlet it is not surprising to see this.

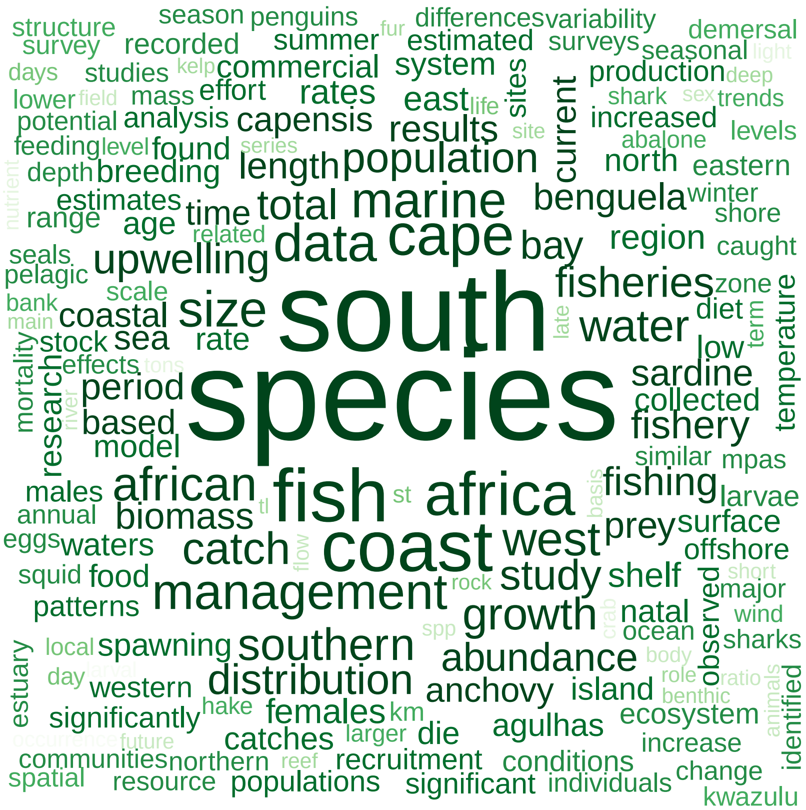

Figure 1: Word clouds based on all the papers published in *AJMS* since inception in 1983

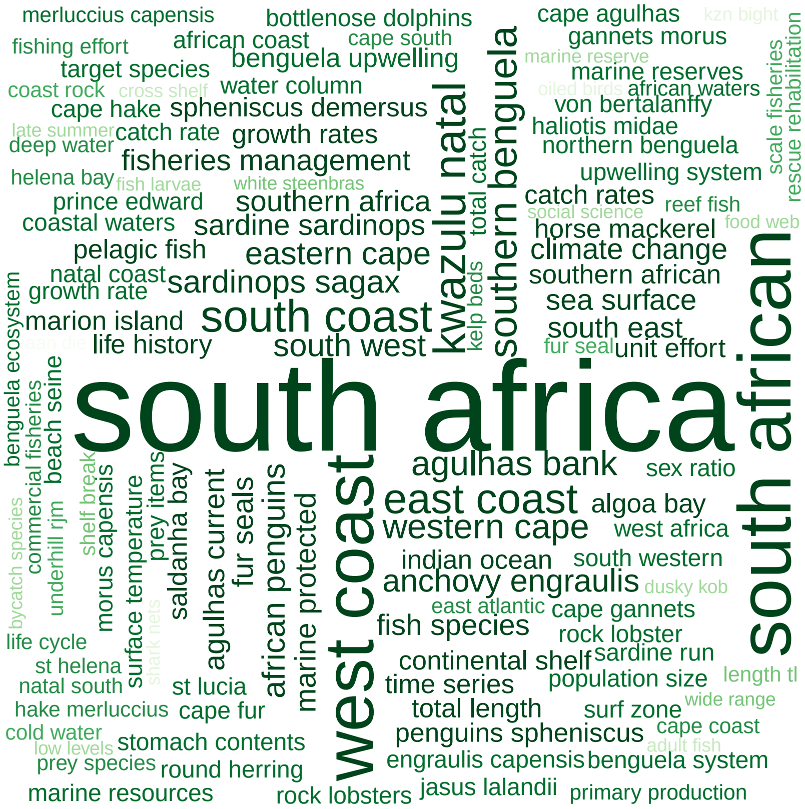

Figure 2: Word clouds of bigrams (word-pairs), based on the same set of abstracts as in Figure 1

In total 24 themes/topics were identified using the *CaoJuan2009* metric, a minimization metrics, Figure 3.

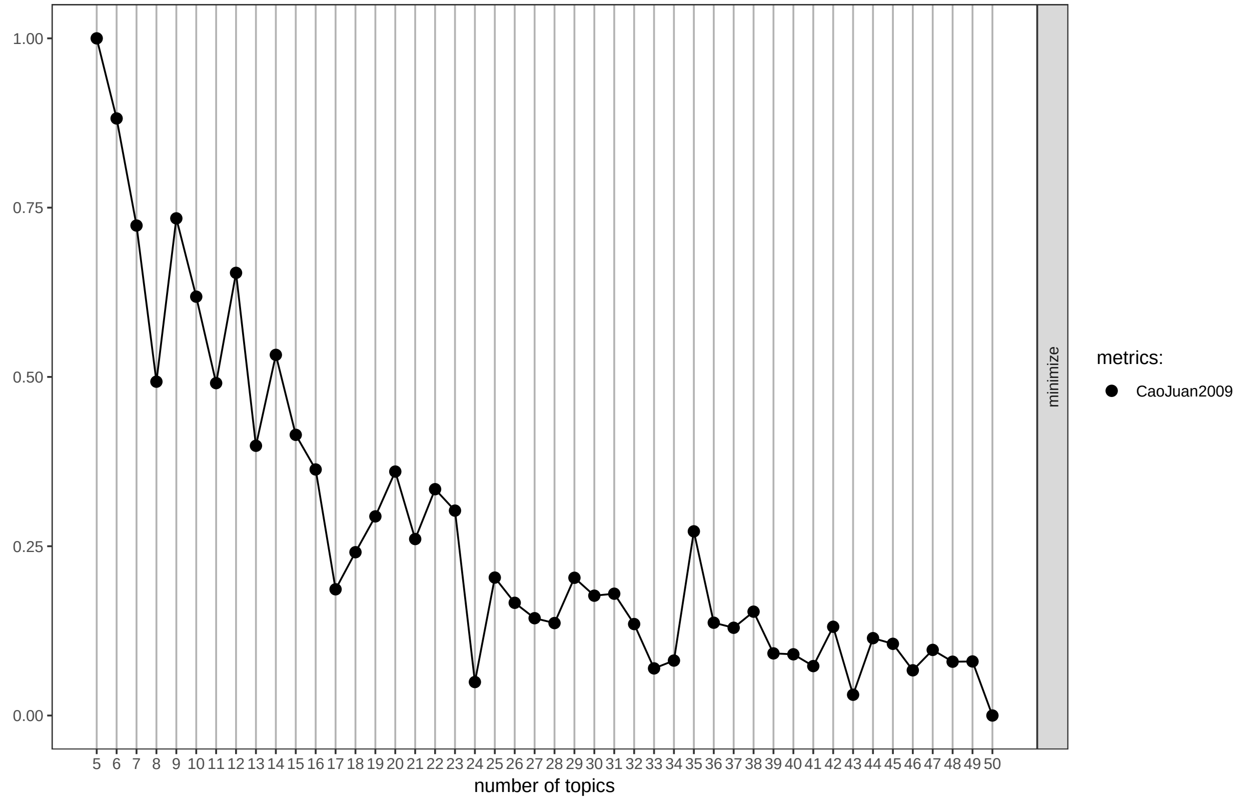

Figure 3: Profiles of the metrics/algorithm, CaoJuan2009, used for the identification of optimal number of topics

The word clouds (sets of words) that characterise each of the topics/themes are shown in Figure 4 for first 12 topics and in Figure 5 for the remaining 12 topics. Time-series of the numbers of papers on each of the topics are given in Figure 6. As can be seen from Figure 6,for most of the themes/topics identified the numbers of papers were variable but largely stable. There were some exceptions, however. For example, the numbers of papers related to harmful algal blooms were largely patchy with periodic up-ticks, some of which might potentially be linked to special issues on the topic. On the other hand, there was a gradual increase in the number of papers related to the theme/topic of generic fish biology.

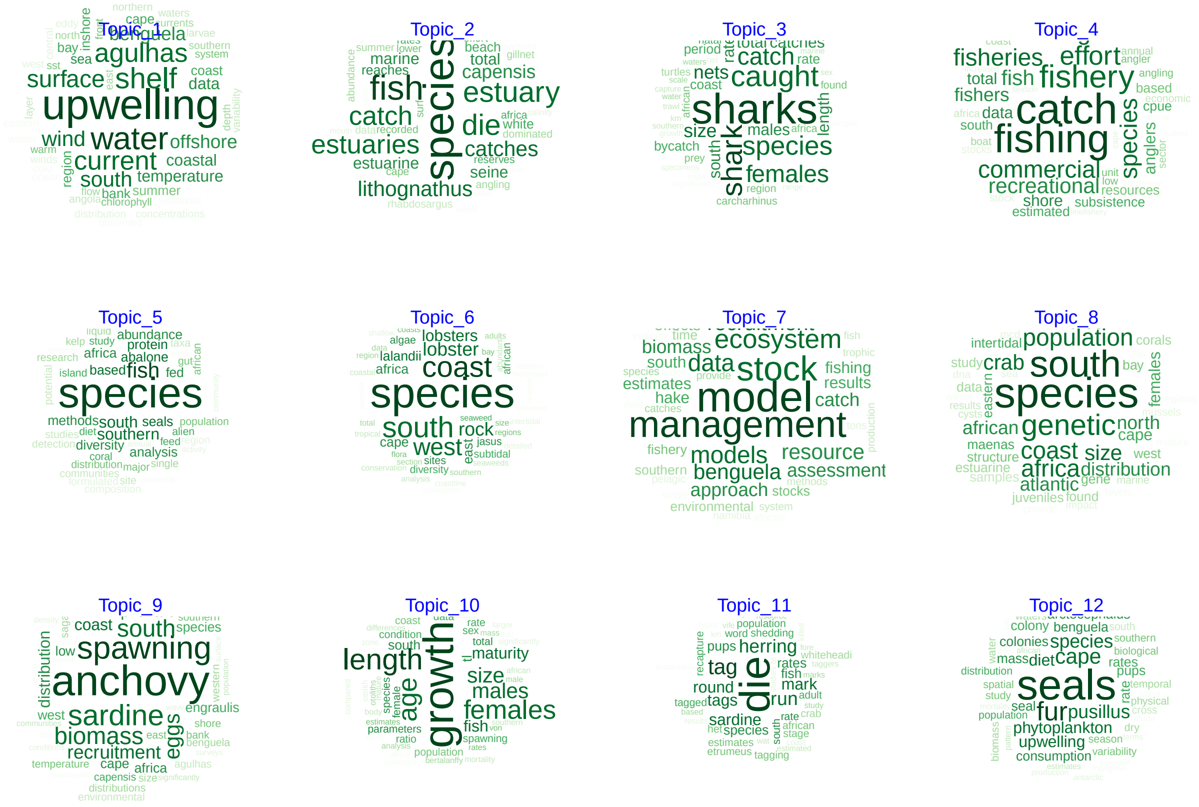

Figure 4: Word clouds that characterise each topic identified; the first 12 topics

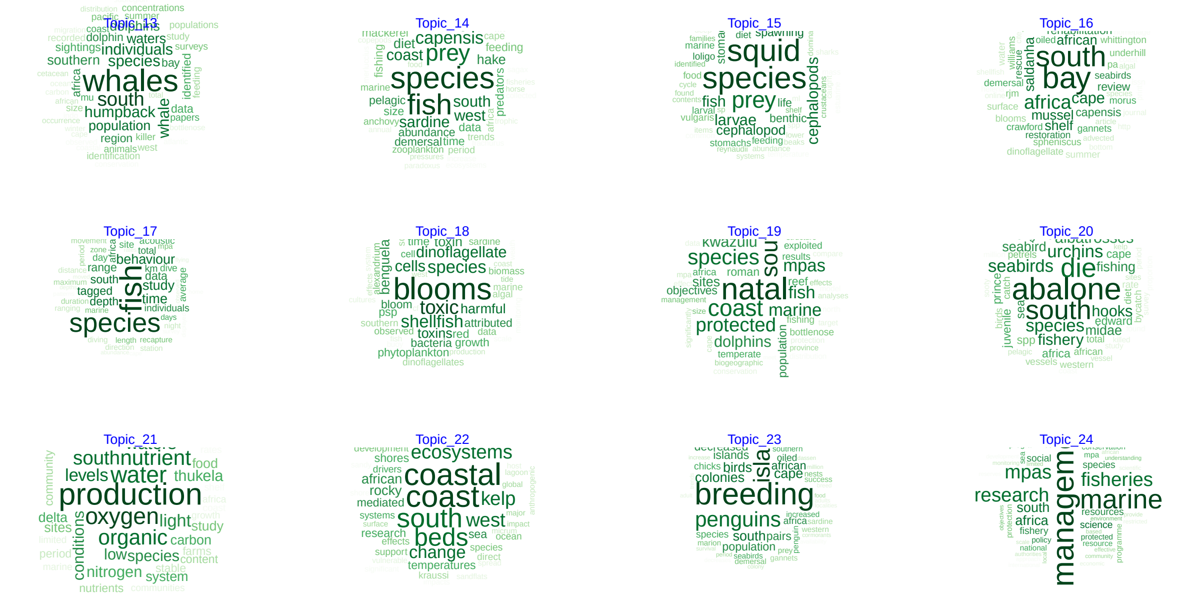

Figure 5: Word clouds that characterise each topic identified; the remaining 12 topics

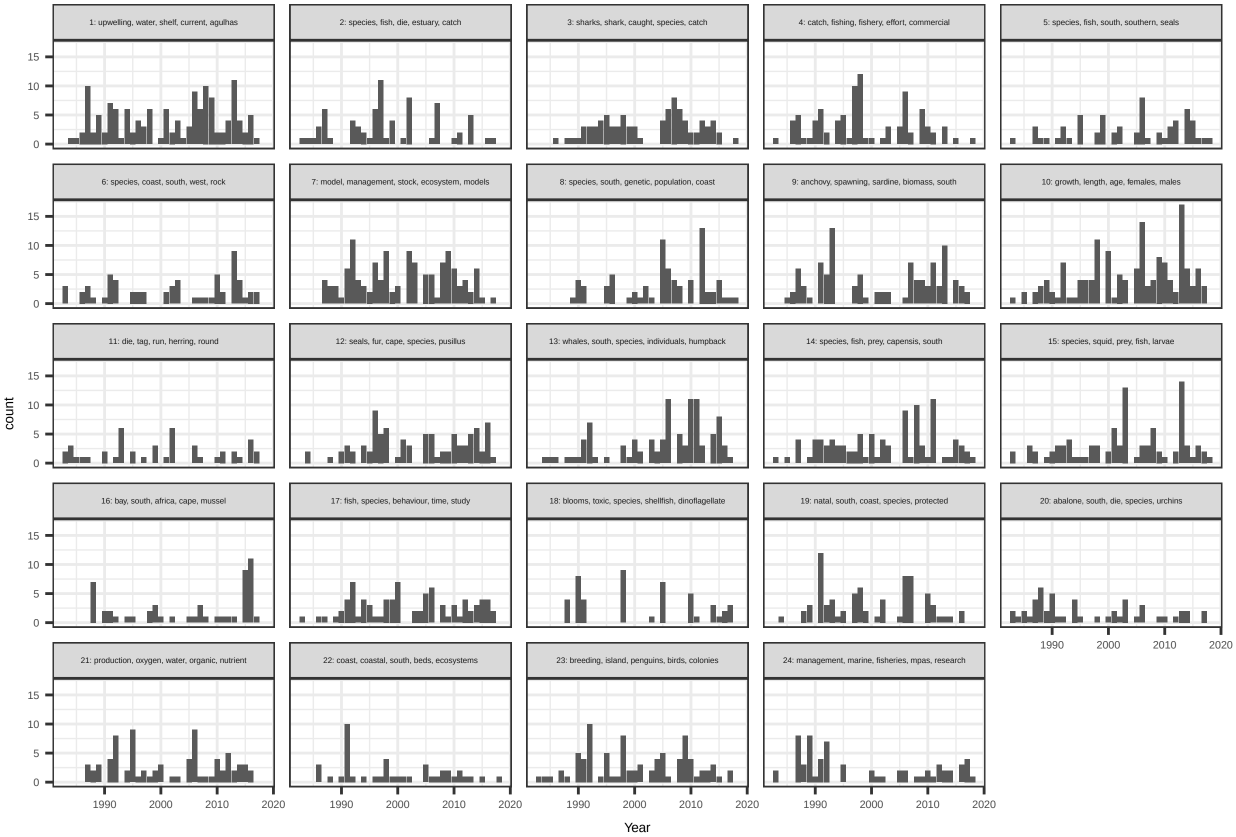

Figure 6: Time-series of the numbers of publications that fall in each of the topics identified

The final sets of topics/themes identified can potentially be interpreted as in Table 1.

Table 1: List of the topics identified and potential interpretation

Topic	Interpretation
Topic_1	Benguela and Agulhas current with empahsi on upwelling related physical processes
Topic_2	Estaurie/estuarine related catches, using e.g. seine nets, of estuarie associated fish species
Topic_3	Bilogy and various fisherie aspects of sharks
Topic_4	Commercial fisheries and their management
Topic_5	Abundance, diversity various aspects of different taxons (species)
Topic_6	South and west coast lobster: their abundance, distribution, diet . . . etc
Topic_7	Stock assessment models, approaches and managment of marine resources
Topic_8	Species/population distriribution and genetics
Topic_9	Small pelagic focused: on the spawning, biomass,and recruitment of sardine and anchovy
Topic_10	Generic fish biology: on fitting growth curves, comparing growth rate by sex,. . . etc

Topic	Interpretation
Topic_11	To tag and re-capture studies and also in relation to the sardine run
Topic_12	Fur seal dynamic in relation to physical processes (upwelling) and productivity
Topic_13	On marine mammals (humpback whales dophins) in southern africa: sighting, feeding, suvreys
Topic_14	Trophic ecology off the west and southrn africa
Topic_15	Squid/cephlopod focused: Diet, spwawning
Topic_16	On Seabird/penguin rehabilitation
Topic_17	fish behavior studies based on tagging
Topic_18	On the dynamics of harmful algal blooms, potential consequences
Topic_19	On marine protected areas with special emphasis on the KZN coast
Topic_20	The abalone fishery
Topic_21	Biological oceanography: production in relation to nutrient (nitrogen), light and oxygen dynamics
Topic_22	Coatals ecosystem focused: kelp bed ecosystems, rocky shores, eelgrass . . . etc
Topic_23	On Penguin colonies: work on the dynamics of breeding colonies
Topic_24	Management of marine resources: issue of mpas and fisheries

References

Arun, Rajkumar, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. 2010. “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations.” In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 391–402. Springer.

Cao, Juan, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2009. “A Density-Based Method for Adaptive Lda Model Selection.” *Neurocomputing* 72 (7-9). Elsevier: 1775–81.

Carlos, ASJG, and RPMR Thiago. 2015. “Text Mining Scientific Articles Using the R Language.” In *10th Doctoral Symposium in Informatics Engineering, Porto, Portugal*, 29–30.

Dahl, David B. 2016. *Xtable: Export Tables to Latex or Html*. https://CRAN.R-project.org/package=xtable.

Deveaud, Romain, Eric SanJuan, and Patrice Bellot. 2014. “Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval.” *Document Numérique* 17 (1). Lavoisier: 61–84.

Feinerer, Ingo, and Kurt Hornik. 2018. *Tm: Text Mining Package*. https://CRAN.R-project.org/package=tm.

Fellows, Ian. 2014. *Wordcloud: Word Clouds*. https://CRAN.R-project.org/package=wordcloud.

Griffiths, Thomas L, and Mark Steyvers. 2004. “Finding Scientific Topics.” *Proceedings of the National Academy of Sciences* 101 (suppl 1). National Acad Sciences: 5228–35.

Grün, Bettina, and Kurt Hornik. 2017. *Topicmodels: Topic Models*. https://CRAN.R-project.org/package=topicmodels.

Nikita, Murzintcev. 2016. *Ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. https:

//CRAN.R-project.org/package=ldatuning.

Nunez-Mir, Gabriela C, Basil V Iannone, Bryan C Pijanowski, Ningning Kong, and Songlin Fei. 2016. "Automated Content Analysis: Addressing the Big Literature Challenge in Ecology and Evolution." *Methods in Ecology and Evolution* 7 (11). Wiley Online Library: 1262–72.

Nunez-Mir, Gabriela C, Andrew M Liebhold, Qinfeng Guo, Eckehard G Brockerhoff, Insu Jo, Kimberly Ordonez, and Songlin Fei. 2017. "Biotic Resistance to Exotic Invasions: Its Role in Forest Ecosystems, Confounding Artifacts, and Future Directions." *Biological Invasions* 19 (11). Springer: 3287–99.

Ottolinger, Philipp. 2018. *Bib2df: Parse a Bibtex File to a Data.frame*. https://CRAN.R-project.org/package=bib2df.

Ponweiser, Martin. 2012. "Latent Dirichlet Allocation in R." WU Vienna University of Economics; Business.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, and Julia Silge. 2018. *Tidytext: Text Mining Using 'Dplyr', 'Ggplot2', and Other Tidy Tools*. https://CRAN.R-project.org/package=tidytext.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." JOSS.

Westergaard, David, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. 2017. "Text Mining of 15 Million Full-Text Scientific Articles." *bioRxiv*. Cold Spring Harbor Laboratory, 162099.

Wickham, Hadley. 2018. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Xie, Yihui. 2018. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.