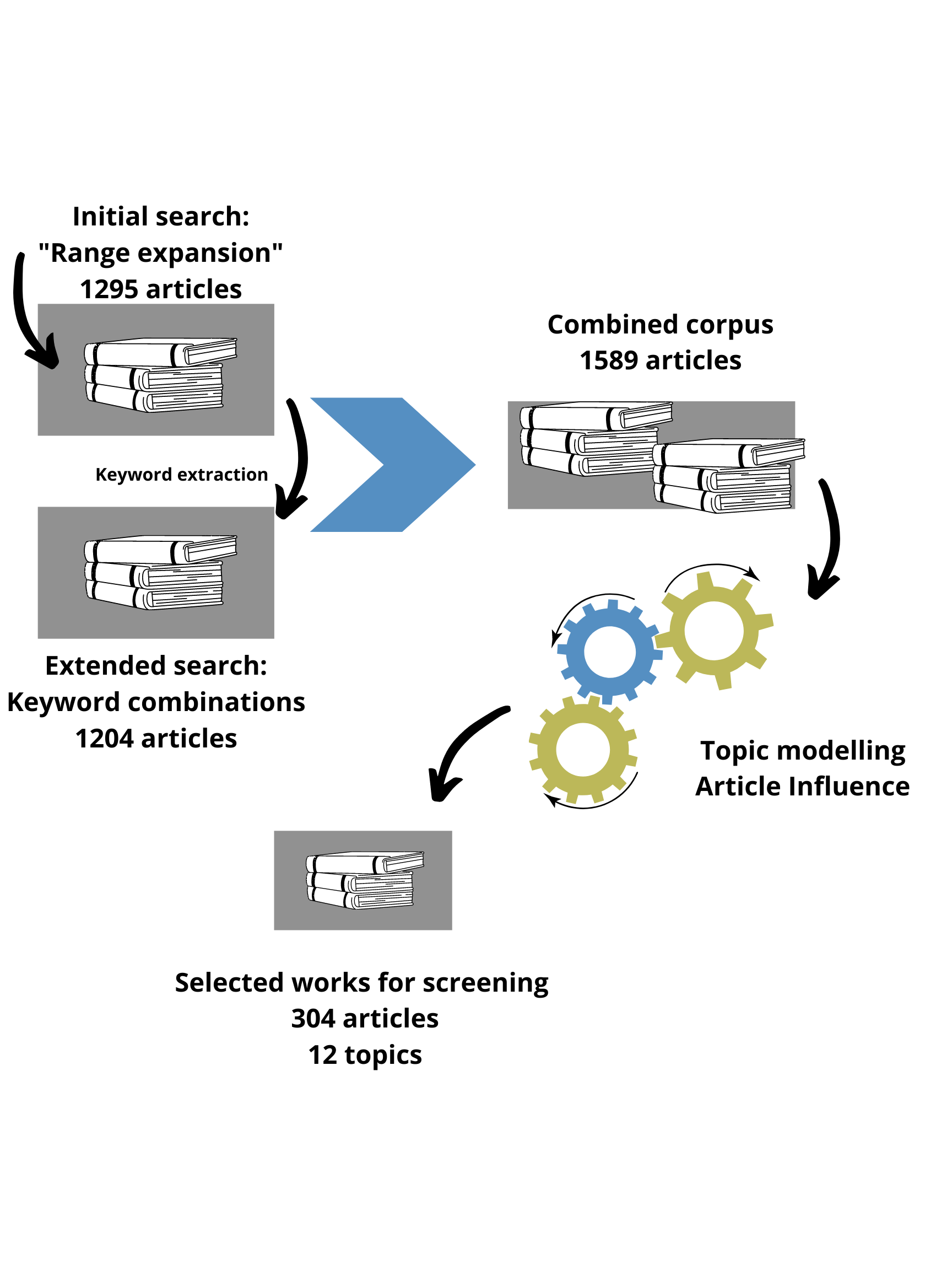
Automated Content Analysis:Literature Review on Range Expansions

Gabriel Muñoz

6/18/2021

Table of Contents

# Automated content analysis of literature on range dynamics



Workflow

## Initial literature search

To get a first impression of the available literature on range dynamics/expansions, we first did a broad search in Web of Science (WoS). We used the text string **“Range expansion”** as the keyword for the first query. The keyword was matched agains the collection of titles, abstracts, and Keywords Plus\* of the full collection of WoS. That is documents from 1979 to the present date (2021). We filter the search results to include documents only those categories that relate with the Natural Sciences. This resulted in 3465 documents We exported the search results from the WoS webpage into a .csv file. The initial search procedure in WoS was done during the first week of June 2021.

## Bibliometric overview of initial search results

We created an initial literature database which contained 3465 rows and 68 columns. Every row contains information for a separate document, every column contains variable fields such as title, DOI, Abstract, which are used to categorize documents. First, we were interested to observe the temporal distribution of articles extracted from the WoS database.

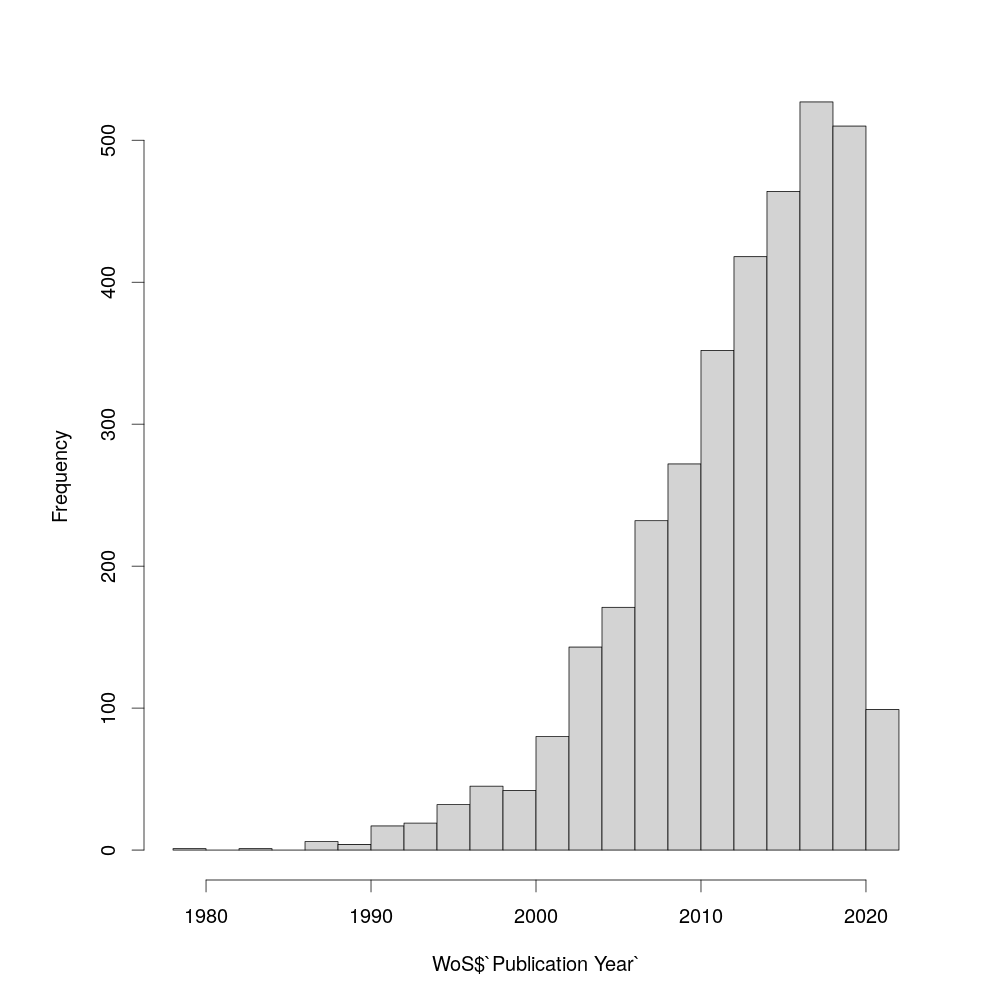


Figure 1: Year frequency distribution of published articles on range expansions

Then, we explored the distinct sources where the queried documents are published. Our search results included articles, proceedings papers, article reviews, and books. We plotted the distribution of the journals with the most number of articles matched with the keyword.

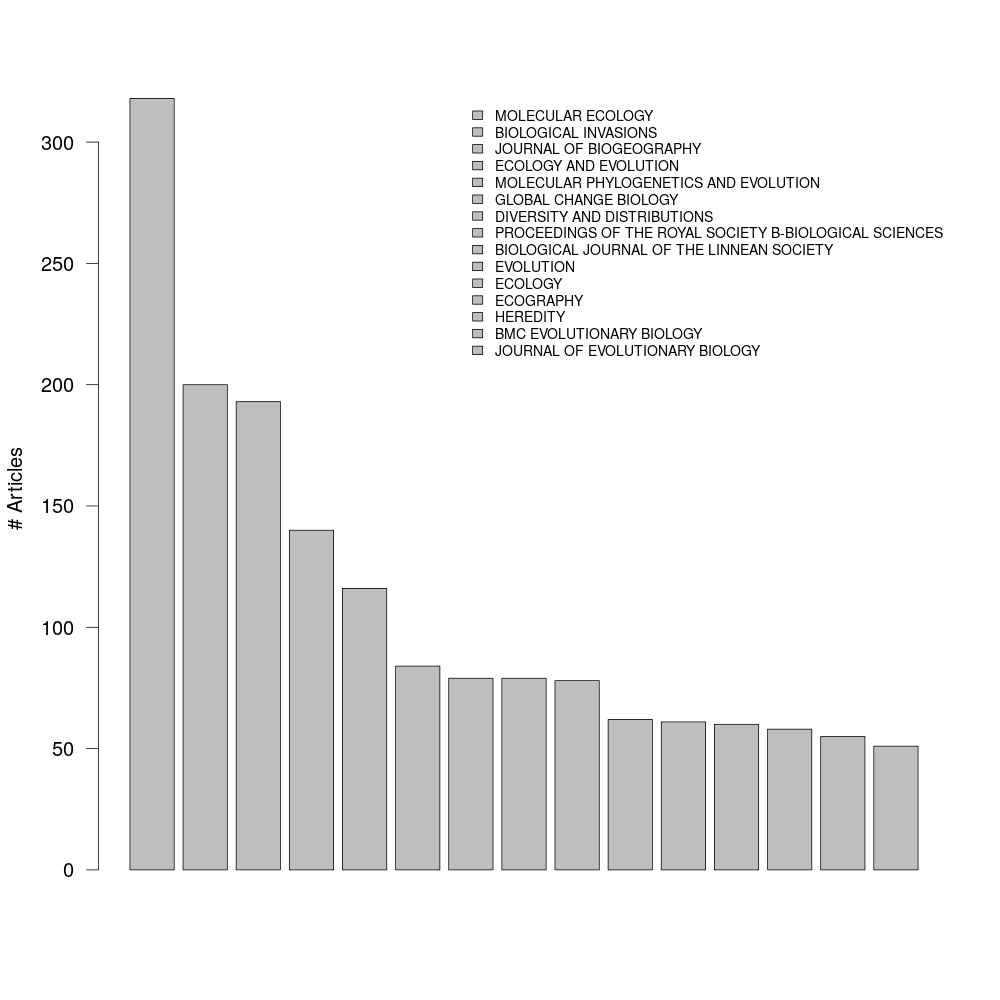


Figure 2: 100 sources with the most information richness on range expansions, counted as unique DOIs

We were also interested to observe the frequency distribution of publications across the time spanned in the database, but this time refined by article source. We did this to observe if there are journals that are more popular to publish science related to range dynamics than others across distinct time periods. Also, how fast have distinct journals have been accumulating information on range dynamics/expansions. To this end, We created a matrix with article sources as rows and year as columns. Entries in this matrix corresponded to the frequency of articles of a given source per a given year. Then we visualized this matrix as a heatmap (Figure 3).

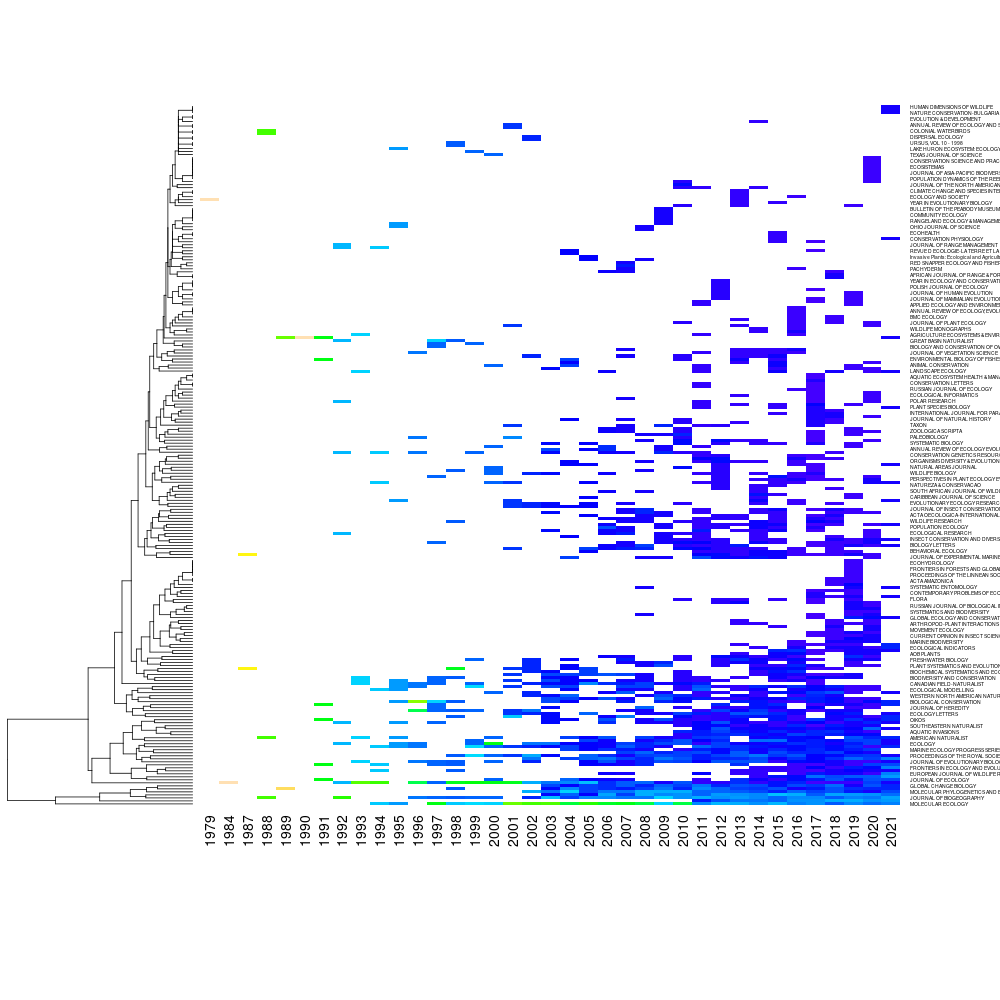


Figure 3: Temporal distribution of publication frequencies for every article source mined, lighter colors simbolize larger frequencies, palette scheme follows column frequencies. White shows areas with no articles

Visual interpretation of the heatmap suggest that there is a temporal aggregation of information, which is increasingly larger over the years but gets published into a increasingly smaller set of journals. However, this may reflect a general trend of scientific publication rather than some particularity of range-dynamics/expansion research. Finally, we created a subset of the initial database to contain only journals that are general and with a broader scope. This sub-setting step reduced the initial database from +3400 articles/233 article sources to 1295 articles/48 article sources.

## Keyword extraction

### Annotating text from titles and abstracts.

We used the package udpipe to perform text annotation. Text annotation is a text mining technique that identifies and separate paragraphs, sentences, and words from text to create variables that can be analyzed later quantitatively and qualitatively. Using a pre-trained model in English grammar, We identify word categories for all words in the set of titles and abstracts of the literature search database. Words were tagged and identified with unique IDs that show whether the word is a noun, verb, adverb, etc. Separate word IDs locate the position of the word within the concatenated string of texts that represent titles or abstracts. The annotation was done separately for the set of titles and for the set of abstracts.

Since words and word associations convey patterns of meaning, word term vectors can be used to quantitatively compute the degree of concept similarity between two pieces of text. As such, articles can be embedded into a hyperspace of term vectors, which is typically extremely sparse. The distance between articles (or words) in this sparse hyperspace reflects the similarity of concepts conveyed in a set of articles, or a set of words. The embedding of documents in the sparce wordterm hyperspace of is usually called as **Latent Semantic Analysis (LSA)**. We performed a LSA for the full set of abstracts from the inital subset of 1295 documents using the packages tm, SentimentAnalysis, and RWeka for R.

The general procedure for a LSA goes went as follows:

1. I removed all numbers from abstracts (numbers in strings are coded as characters, so they can be interpreted by the machine as words and lead to non-meaninfull word co-occurrences).
2. I created a virtual corpus with the abstract text strings. Each entry of the corpus corresponds to an individual abstract.
3. I tokenize terms, remove puntuation marks, english stopwords (e.g. the, a, an, etc.), stemmed tokenized terms (leave only the word roots, e.g “cascading” and “cascade” are both equal to “cascad”)
4. I applied a n-gram tokenizer function to the corpus with n=3. This results in term variables that correspond to 3-wise combination of subsequent words in a sentence. I did not discriminate words based on their type for this analysis.
5. I created a document term weighted matrix, documents correspond to the abstracts and terms are the 3-gram word collocations. Weights of this matrix corresponded to the frequency of terms at a given document.
6. Computed the LSA hyperspace from the document term matrix using the lsa model from the lsa packge (Microsoft).
7. A computed a pairwise distance matrix between documents in the lsa hyperspace.
8. Cluster the distance matrix with k-means. I tested 3 to 10 groupings and tested the clustering of variance within groups with the Calinsky criterion. I selected as significant grouping the grouping with the lowest Calinksly. The smaller the number reflects that more variance is contained within groups than among groups.
9. Computed noun-cooccurrence networks for the set of articles at each k-mean partition (Figure 6)

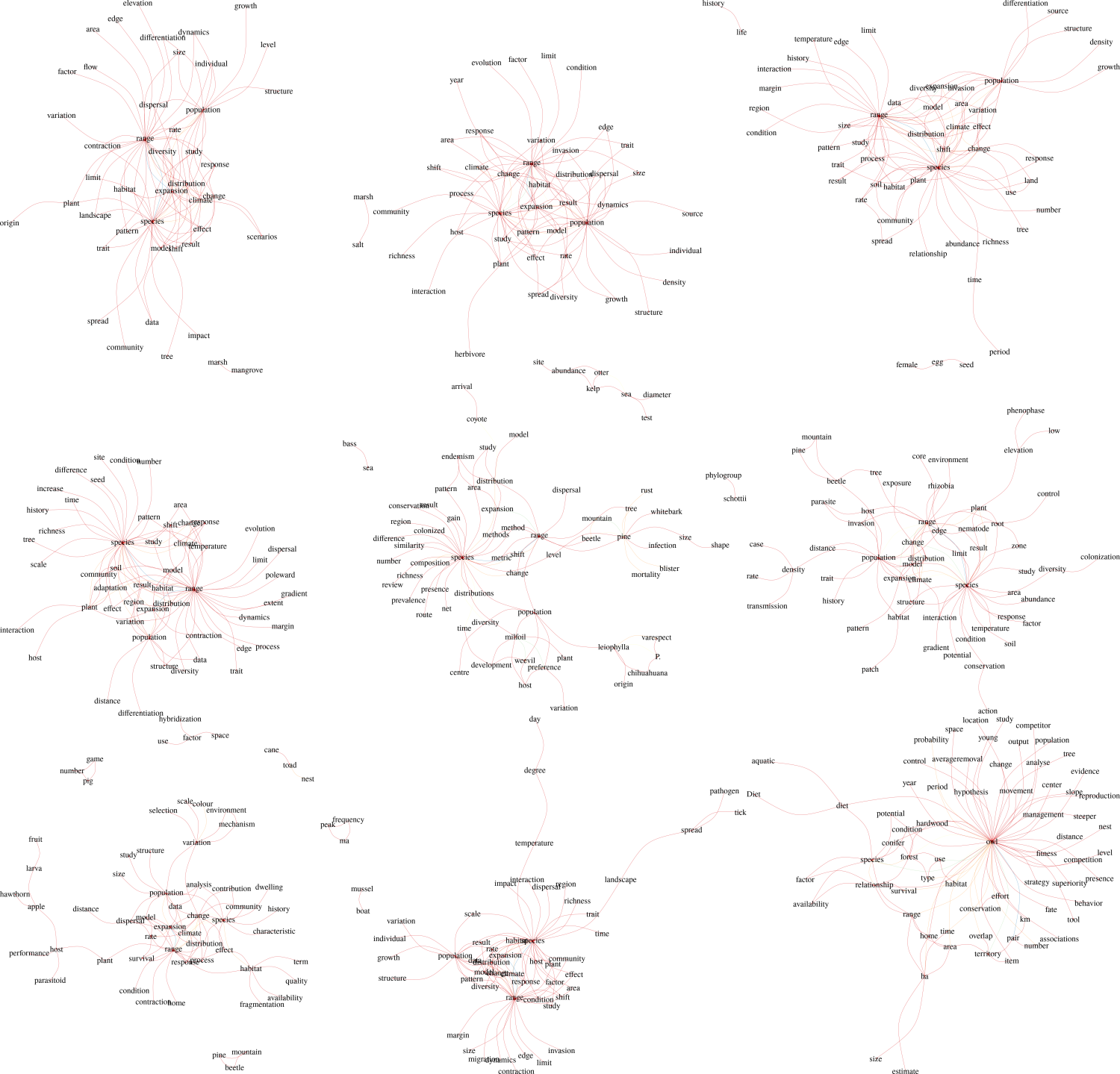


Figure 6: Network of word co-occurrences for dissimilar clusters of with similar document abstract contents between 1979-2021 that matched ‘range expansions’ and are published in general journals of Ecology

## Expanded literature search

The results of the keyword extraction showed there is a strong association between range expansion and climate change. To account for this association as well as expanding the literature search to match the objectives of the review, we used as keywords the following to make repeated searches in WoS database.

|  |  |
| --- | --- |
| Group | Term |
| 1 | rang shift |
| 1 | rang expans |
| 1 | rang edge |
| 2 | climat chang |
| 3 | land use |
| 3 | land cover |
| 4 | movement |
| 4 | dispers |
| 5 | demography |
| 5 | population structur |

We selected keywords for the renovated search effort by pairing terms from distinct group combinations but always including the group 1: In total we made 19 independent search with unique keyword pairs. We limited the search as before to include only documents in the natural sciences. Similar as before, articles were uniquely identified by its DOI (Digital Object Identifier). We combined the search results from WoS and remove duplicated DOIs. In total we obtained 1204 unique DOIs. We compared this new article set with the one from the previous broader search with only “range expansion” as keyword. This give us an idea of the added search returns added from a more complex search effort. In total the broad and the complex search efforts shared 887 DOIs, which represented the 55% of the total article sum. Thus, to maximize the breadth our literature search effort we merged both datasets into a single one containing 1589 unique DOIs.

## Topic modelling:

We modeled the topics present in the corpus of relevant literature. We did this separately for the set of titles and abstracts for the DOIs. We first pre-process the text corpus. We removed numbers, punctuation marks, and English stopwords. Finally, we stemmed the terms and tokenize terms into bigrams. This resulted 11318 terms for the document term matrix on the titles and 149494 terms for the document term matrix of the abstracts. Document term matrices are typically extremely sparce and classic clustering methods based on vector dissimilarity do not work well in sparce matrices. To classify documents into text categories, we infer topics from the corpus text with Latent Dirichlet Allocation (LDA). In an LDA model, terms are probabilistically assigned to a set number of topics. A topic thus has its own set terms, but terms can be shared among topics. An LDA model also assigns documents to topics with probabilities driven by the representation of the terms in the document in the terms in the topic words. We applied the LDA model for a series of 5 to 15 topics and selected the model with the best coherence. That is that the model where documents and terms are more evenly spread among topics. Based on this, we selected a total of 12 topics to classify our set of 1589 documents. We plotted the terms with the 30 highest posterior distributions for each topic in wordclouds (Figure 7-8)

### Topics from title terms

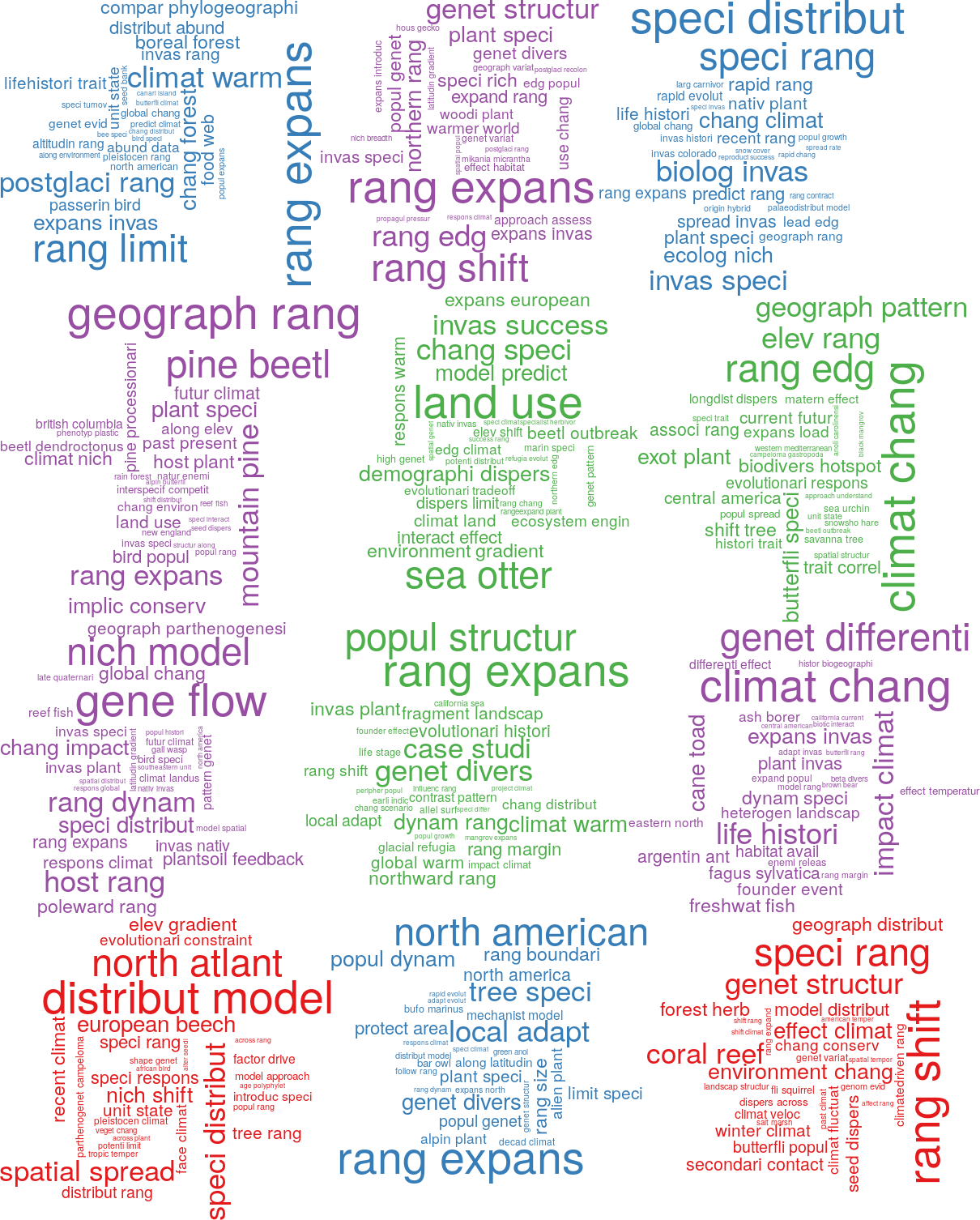


Figure 7: Wordclouds of the most relevant terms for each topic classification based on the terms from the titles of 1589 documents included in the corpus

### Topics from abstract terms

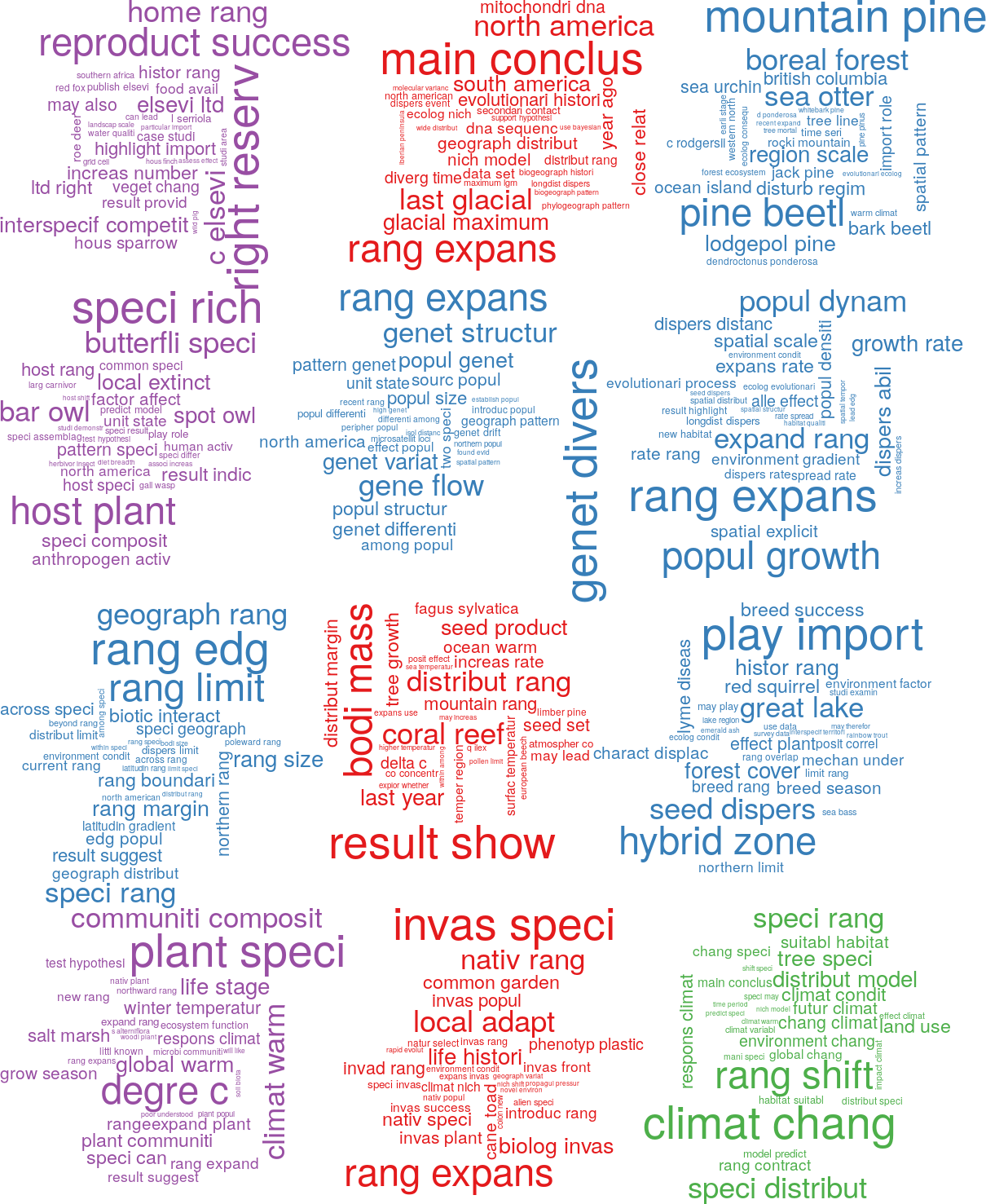


Figure 8: Wordclouds of the most relevant terms for each topic classification based on the terms from the abstracts of 1589 documents included in the corpus

## Selection of candidate articles

From the list of DOIs, we obtained the citation records of each article. We did this to categorize articles based on their overall academic influence. We obtained the citation records by matching the DOI list with the crossref database using the rcrossref package. Because number of citations accumulate logistically overtime, we model a raw measure of “article influence” as the residuals from a loess model with citation count as response variable and year of publication as the predictor. Positive residuals of this model will indicate articles that has accumulated citations at a faster rate than expected from its local averages (Figure 9).

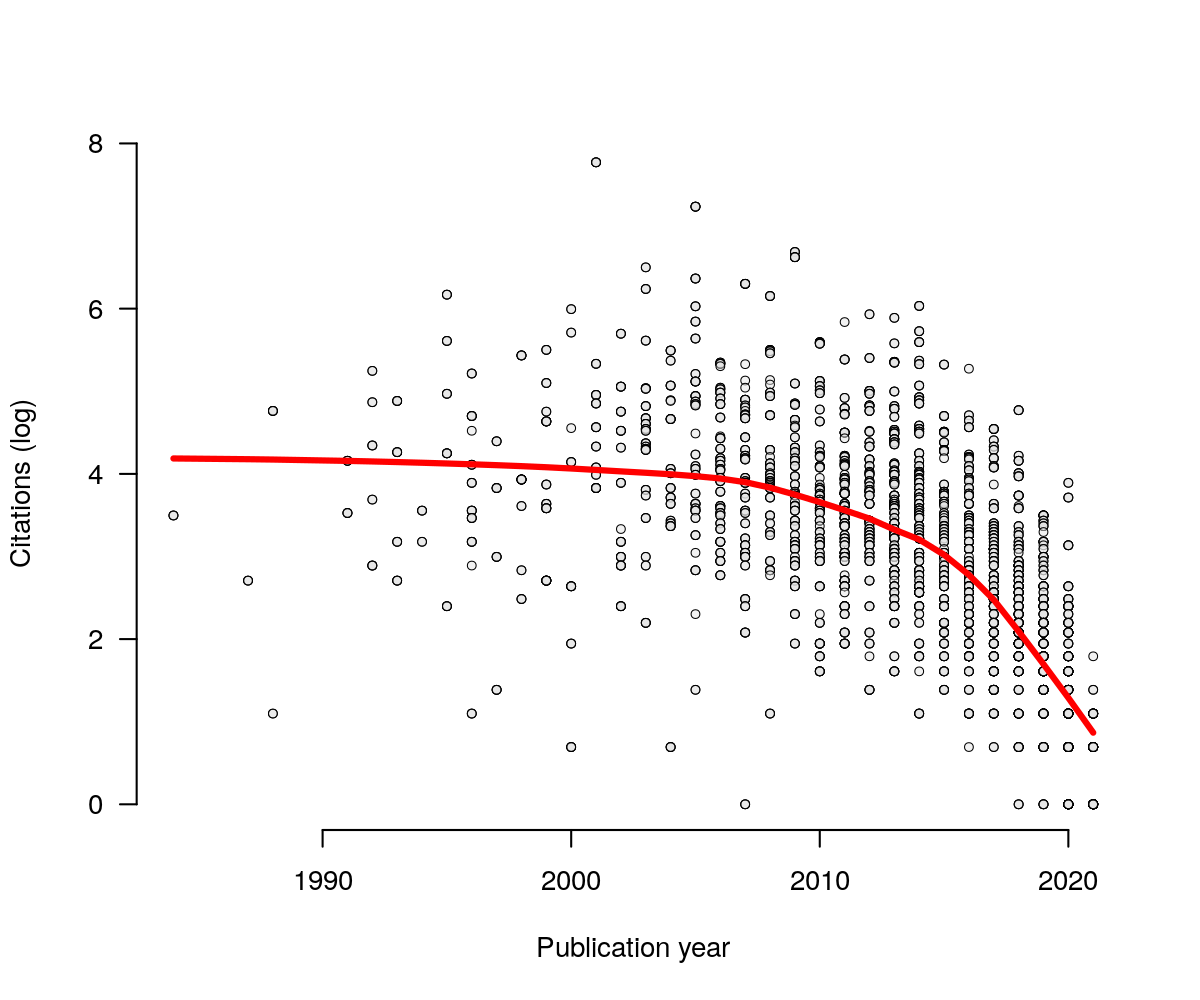


Figure 9: Relationships between citation count and year of publication of the articles extracted from WoS. Red line shows the predicted trend from a smoothed local weighted regression model LOWESS

To select relevant articles for screening we threshold the distribution of such article influence variable. We only select articles within the top 25% quantile (Figure 10). This reduced the literature corpus to 304 relevant articles ranked and classified into 12 distinct topics (Figure 10).

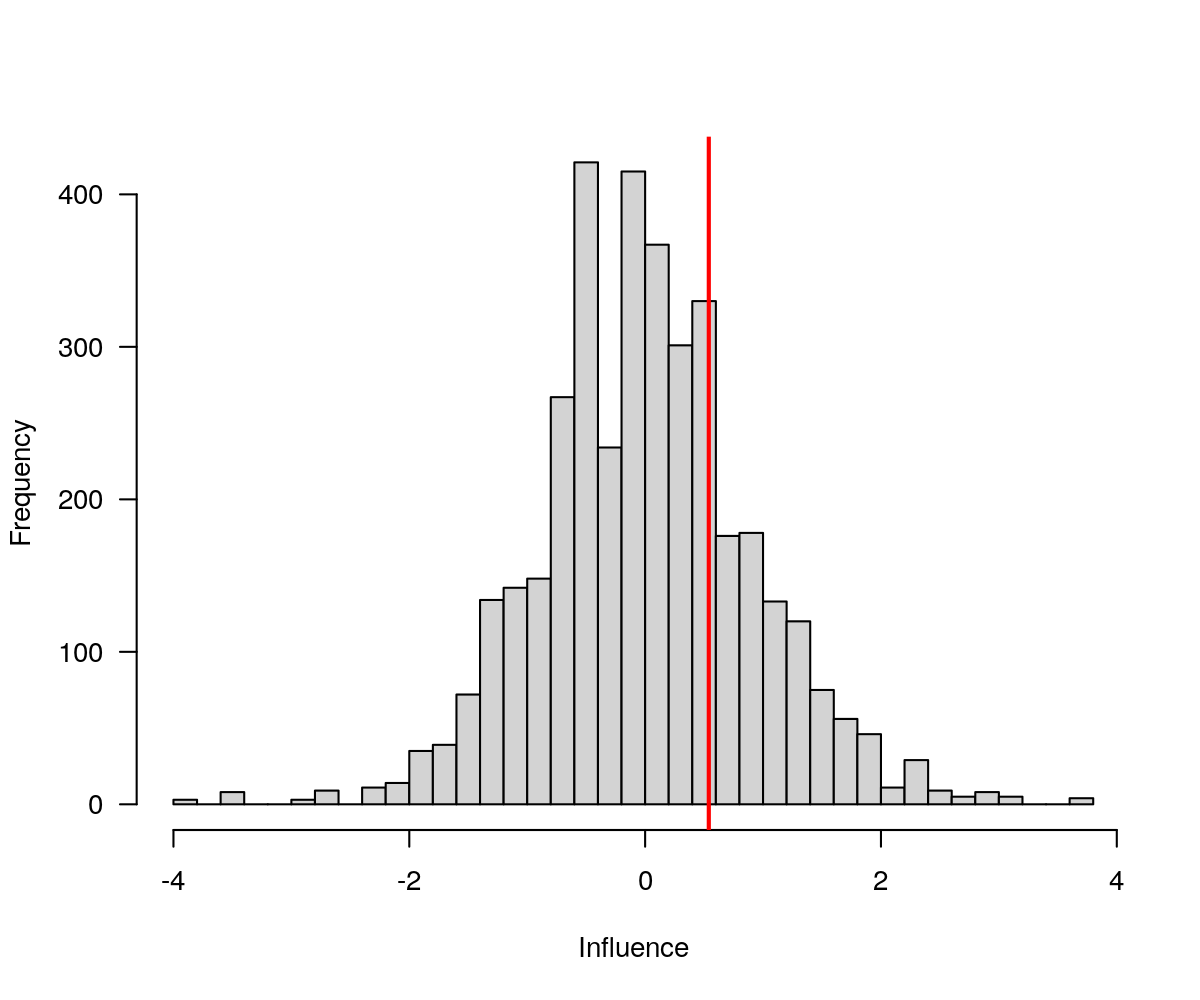


Figure 10: Distribution of the influence metric across articles in the corpus. Vertical red line shows the division of the 0.25 largest quantile

### Network of knowledge

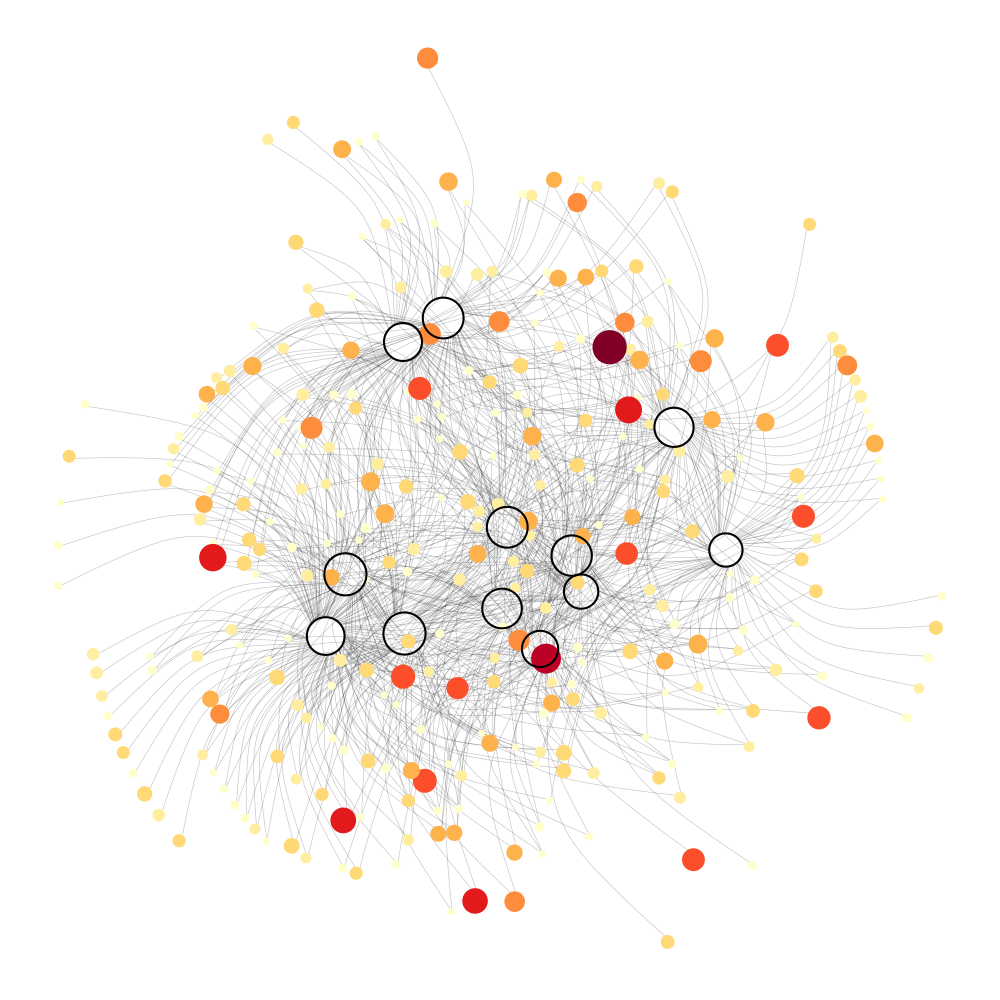


Figure 9: Network of relevant articles and their relationships with the inferred distinct topics in the corpus. Colored nodes in the networ represent a separate article, hollow nodes correspond to a topic (based on the abstract topics). The size of article nodes represent the overall influence of the article within the corpus of text. Size of topic nodes represent the number of articles included on each topic. An article can be present in more than one topic.

Additionally, we examine the effects of varying journals for the overall impact of a publication within the “range expansion” subset of ecological knowledge. Therefore, we classified the Journals in the final corpus based on the aggregated journal influence and overall journal popularity (or productivity). To create this metrics, we aggregated the article influence metric by its journal means (journal influence) and relate it the total number of articles (journal popularity) (Figure 10).

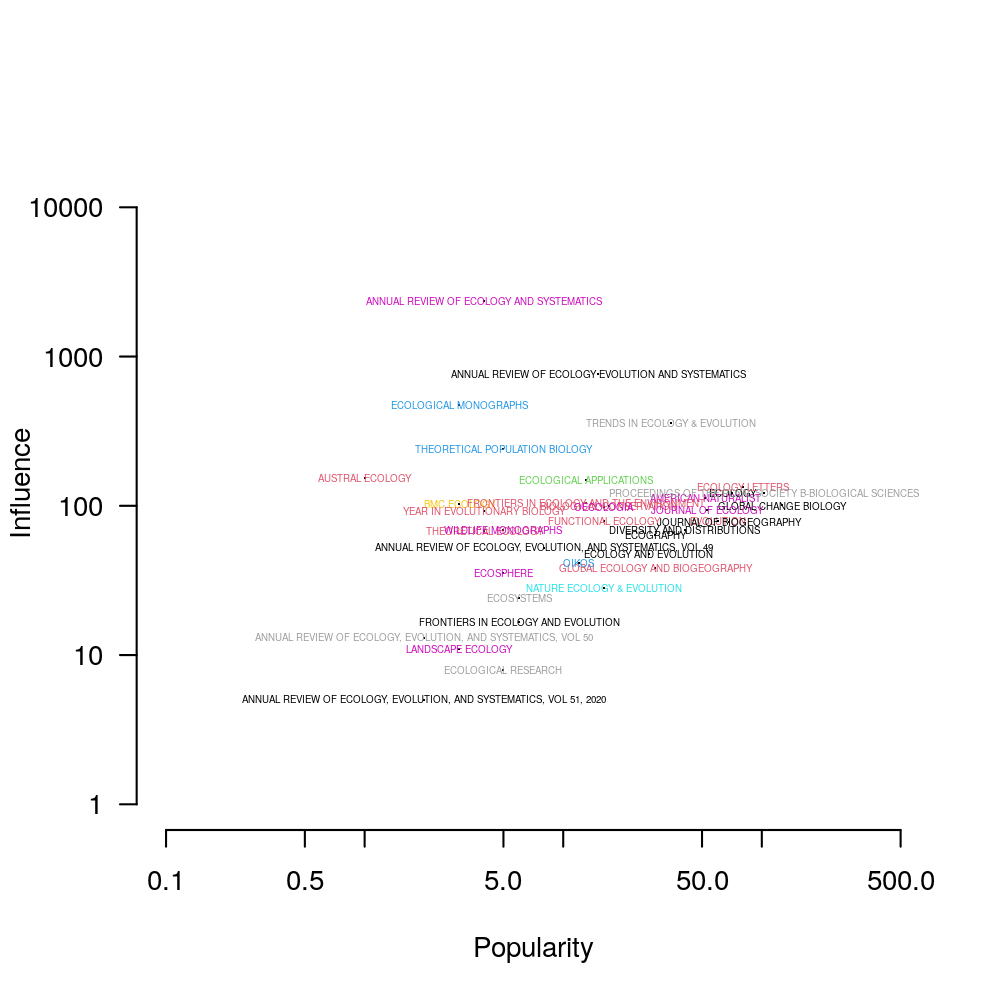


Figure 9: Journal location in a popularity and influence space