HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 320

MEASURING THE DIVERSITY OF ECOLOGICAL RESEARCH

IS LATENT SEMANTIC ANALYSIS A SUITABLE TOOL?

VON
ANNETTE SCHEINER

MEASURING THE DIVERSITY OF ECOLOGICAL RESEARCH

IS LATENT SEMANTIC ANALYSIS A SUITABLE TOOL?

VON
ANNETTE SCHEINER

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 320

Scheiner, Annette

Measuring the diversity of ecological research : is latent semantic analysis a suitable tool? / von Annette Scheiner. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2012. - 65 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 320)

ISSN 1438-7662

Abstract:

Latent Semantic Analysis (LSA) has recently been proposed as a tool for analyzing research diversity and for extracting latent themes from a set of documents. Using ecological research as an example, this study scrutinized the suitability of LSA for achieving these goals. In addition to the already tested calculation of research diversity based on co-citation of references, three other document properties – title, abstract, and keywords – were used and compared to the results of the reference-based approach. The results of the analyses suggest that LSA is – in its current state – no very suitable tool for either analyzing research diversity or extracting latent themes from bibliographies. Therefore, other options of measuring research diversity should be evaluated in the future.

Diese Veröffentlichung geht zurück auf eine Master-Arbeit im postgradualen Fernstudiengang Master of Arts (Library and Information Science) an der Humboldt-Universität zu Berlin.

Online-Version: http://edoc.hu-berlin.de/series/berliner-handreichungen/2012-320

Contents

List of Figures

List of Tables

1 Introduction

Diversity is one of the most fascinating aspects of ecology. Ever since Darwin – and probably already before – scientists have been amazed by the diversity of life and have strived to explain why some spots on earth have a more diverse flora and fauna than others. To this end, several methods of measuring and calculating diversity have been developed (e.g. Simpson's diversity index, Shannon's diversity index or the Berger-Parker-Index). These methods have meanwhile been adapted by other scientific disciplines, like demography and economics, and are also used in information science to study the diversity of research fields (e.g. Grupp, 1990; Havemann *et al.*, 2007; Stirling, 2007; Mitesser, 2008).

The research field of *ecology*, a term coined by Ernst Haeckel in 1866 (who called it "Oekologie" in German), has flourished and diversified ever since. Research topics embrace both experimental and theoretical work on biotopes around the world from the largest living plants and animals to the tiniest protozoans. Even genetic aspects of ecology as well as genetic diversity have been studied for quite some time now. So what would come more natural for an ecologist new to the field of information science than to study the diversity of ecological research?

Although examining research diversity in ecology is sufficiently fascinating by itself – and several aspects of it have already been studied e.g. by Rodriguez & Moreiro (1996), Rivera (2003), and Neff & Corley (2009) – some other interesting questions are currently being discussed in terms of research diversity, one of which is whether funding policies of research foundations or national funding agencies have an influence on research diversity. One popular hypothesis, which states that research diversity might decline if funds can be raised more easily with research topics that match the most popular topics of the current funding directive, has already been studied in different contexts (Harley & Lee, 1997; Adams & Smith, 2003; Gläser & Laudel, 2007; Whitley, 2007; Gläser *et al.*, 2008; Mitesser *et al.*, 2008).

However the question which method is most suitable for conducting such an analysis is not yet settled. Early studies (e.g. Marshakova, 1973; Small, 1978) have analyzed the *co-citation* of (frequently cited) papers to elucidate the most important concepts of research. Another option is to analyze the *bibliographic coupling* between papers and thus try to extract clusters of related papers that represent different research topics. Neff & Corley (2009) have recently published a "*bibliometric exploration of the evolution of ecology*", in which they analyzed trends in the methods and topics in ecological research papers from 1970 to 2005 by means of *co-word analysis*.

Lately, M. Heinz, F. Havemann, J. Gläser, and O. Mitesser have been discussing *Latent Semantic Analysis* (LSA) as a new and promising approach to analyze research diversity. Consequently, Mitesser (2008) scrutinized two alternatives (deterministic and probabilistic) of LSA in his master's thesis in order to analyze the diversity of two research fields: electrochemistry and scientometrics. This approach proved to be quite promising and stimulated further studies (e.g. Havemann *et al.*, 2009). LSA was developed in the late 1980ies in the context of information retrieval under the auspices of T. Landauer and was later patented (US Patent 4839853) and published by Deerwester *et al.* (1990). By means of LSA, documents were compared by analyzing which terms/words they had in common and by deriving a set of concepts related to the documents and terms in order to improve retrieval quality.

Such an approach seems also quite useful for extracting research topics from research papers by analyzing which cited references they have in common. The mathematical base of this analysis is a *Singular Value Decomposition* (SVD), in which the document-reference-matrix of a sample (which represents the occurrence of the cited references in the documents) is decomposed into three matrices, containing the left and right singular vectors, as well as the singular values of the document-reference-matrix.

While the left and right singular vectors represent the relationships between the derived topics and the references and documents, respectively, the singular values represent the size of the topics and were used by Mitesser (2008) and Mitesser *et al.* (2008) to calculate the diversity of the sample by the Shannon diversity index (Shannon, 1948). Another option would be to use the Simpson index (Simpson, 1949).

In ecology the *Shannon index* (Eq. 1.1) takes into account the number of species and their evenness. The diversity increases when more different species are present or when the numbers of individuals in each species are distributed more evenly. It is calculated from the relative abundances p_k of the K species in the habitat as:

$$H = -\sum_{k=1}^{K} p_k \log p_k. \tag{1.1}$$

The Shannon index reaches its maximum for a given set of species ($H_{max} = \log K$) when each species is present in equal numbers and gets minimal when there is only one species. In terms of research diversity, the different research topics correspond to the different species and the documents represent the individuals that belong to each species. Thus, the Shannon diversity increases when there are more research topics or when the number of documents that deal with each topic are distributed more evenly. It reaches its maximum when each document deals with a completely independent topic.

The *Simpson index* (Eq. 1.2) also takes species number and evenness into account and represents the probability that two randomly selected individuals in the habitat will belong to different species k and l:

$$S = \sum_{k,l=1}^{K} p_k(1-\delta_{kl})p_l = 1 - \sum_{k=1}^{K} p_k^2. \tag{1.2}$$

The Simpson index peaks ($S_{max} = 1 - 1/K$) when all individuals are distributed equally among the species, and it equals 0 when all individuals belong to the same species. Like with the Shannon index, the Simpson diversity increases when more similar numbers of documents deal with each topic and peaks when each document deals with a completely independent topic.

In the present study, both diversity indices will be used in order to compare whether they give substantially different results and possibly to decide which one is more suitable for calculating research diversity.

Mitesser (2008) has already analyzed some basic properties of the Shannon index in relation to the type of document-reference-matrix that is commonly found in bibliographic studies – a weakly coupled thin matrix, i.e. a matrix with many more zeros than ones, in which the majority of references is cited only by one document – and has shown that diversities close to the maximum are to be expected. Moreover, he has nicely demonstrated that the number of references per paper, which he observed to increase over time, was not responsible for the increasing diversity of research in electrochemistry and scientometrics.

Nevertheless, it seems recommendable to apply some standardization to the document-reference-matrix in order to minimize the effects of different numbers of references per document and of different citation frequencies of the references. The details of the standardization procedures used in this study are explained in section 2.4.

As described above, LSA has its origin in comparing the occurrence of words in documents. In scientometrics it was first used in relation to the occurrence of references in documents, as the standard analyses like bibliographic coupling and co-citation were based on the linkage of documents by citations, either by being cited in the same context (co-citation) or by citing the same references (bibliographic coupling). As the *Web of Science* (from which all data were downloaded, cf. section 2.1) offers much more information for each bibliographic record than the reference list, it comes natural to broaden the approach to include other term-based information in the analysis. Therefore, three term-based properties of the documents – title, abstract and keywords – will be used to analyze diversity and to compare whether they (all) yield similar degrees/patterns of diversity than initially obtained by using the references-based approach.

In these cases, natural words are the basis of the analysis, and not standardized representations of documents like in the case of references. Therefore, two additional methods of standardization are needed. On the one hand, natural language contains a lot of words that have no special meaning, so-called “stop words”, like e.g. and, or, not, be, moreover, accordingly, obviously, etc. In order to remove these words, one can choose from a variety of stop word lists from different sources, for example the one being used in the present study, which is available from: ftp://ftp.cs.cornell.edu/pub/smart/english.stop. On the other hand, natural language itself brings along a kind of intrinsic diversity, as words that belong to the same word stem can have different endings due to declension and conjugation. Several algorithms have been developed to get rid of these endings and to reduce words to their word stem, the earliest published one being designed by Lovins (1968). A later, today most commonly used stemmer developed by Porter (1980) – and usually referred to as “Porter-Stemmer” – will be used in this study. The practical application of stop word removal and word stemming is explained in section 2.4.

As mentioned above, Mitesser (2008) has already studied some fundamental properties of document-reference-matrices and their consequences on diversity. Two yet unresolved questions are, (1) whether a temporal trend in the number of *different* references in the pool of documents could lead to a corresponding trend in diversity, and (2) which variability could be generated from an existing document-reference-matrix by exchanging pairs of references between documents and thus keeping the original lengths of the individual reference lists as well as the citation frequencies of the references. Some basic attempts will be made to address both highly interesting questions here as well (details of the implementation can be found in section 2.5), however their complete scrutiny would go beyond the scope of this master's thesis.

Finally, the suitability of LSA-based analyses for identifying “real” research topics from the sampled bibliographies based on the eigenvalues resulting from the SVD will be analyzed for two examples, one based on cited references and the other based on title words. This is especially interesting from the ecologist's point of view, as it should be more easy to judge whether the derived topics are really belongig to the same kind of research area or are being falsely detected as similar, e.g. because they refer to the same statistics software or a methodological textbook, but otherwise deal with completely different topics.

2 Material and Methods

2.1 Data

All data were downloaded from the *Science Citation Index* of the *ISI Web of Science* (WoS) in 2009 and 2010 based on the journals listed in the *Journal Citation Reports* 2008 and 2009 in the subject category *ecology*. This “detour” was necessary, as the WoS does not allow to perform searches by subject category. The datasets were downloaded in packages of 500 titles (which was the maximum possible download size) and were stored as *tab-delimited (Windows)* text files. For each title, the download contained metadata in a number of fields, of which the most useful were *author*, *title*, *cited references*, *abstract*, *author keywords*, and *keywords plus*.

Data in the WoS reach as far back as 1900, so the goal was to include as long a time series as possible. The earliest entries in the subject category *ecology* dated back to 1945. However, then only eight journals were indexed and the number of journals grew dramatically over time until 2009, when 112 journals were indexed (Fig. 2.1). This is an increase of 1300% in 65 years. Due to this immense change over time, it was not possible to study all journals together in one time series. Therefore, three separate time series were analyzed as a compromise between analyzing as long a time series as possible and including as many journals as possible. The three time series spanned 40, 30, and 20 years, respectively, i.e. 1970-2009, 1980-2009, and 1990-2009. As the data analyses were intended to focus on the immediate output of research, only publications of the type “article” were included. Consequently, some journals in which mainly review articles are published could not be included in the analyses due to low article numbers.

The longest time series consisted of 14 journals (Table 2.1), which were also included in the two shorter time series. The medium time series thus comprised 27 journals, of which the 13 additional ones are listed in Table 2.2. Finally, the shortest time series contained 9 additional journals (Table 2.3), 36 journals in total.

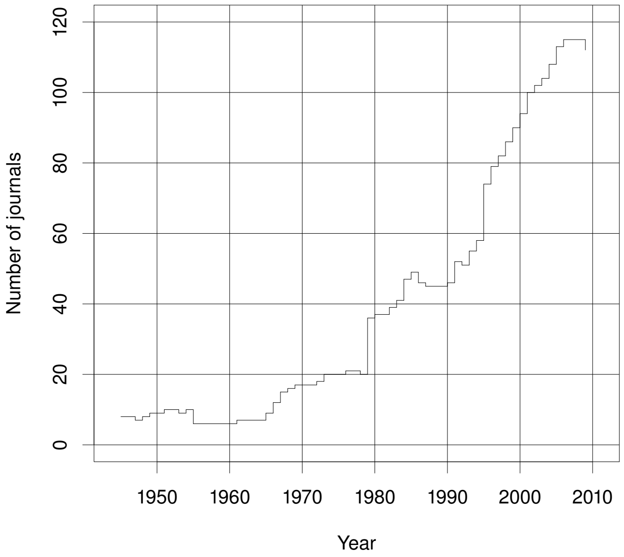

Figure 2.1: Number of journals belonging to the subject category *ecology* indexed in the *Science Citation Index* from 1945 to 2009.

In each time series there was a clear trend of an increasing number of articles over time. From 1970 to 2009 the number of articles per year in the 14 analyzed journals rose from about 770 to more than 2.300 (Fig. 2.2), which is an increase of almost 200%. Similar trends, but less steep increases were observed in the two

Table 2.1: Journals that were analyzed from 1970 to 2009. Listed are the number of issues published per year, the number of publications and the number of publications of the type "article" over the whole time period.

Journal title	Issues per year	Publications	Articles
American Midland Naturalist	2	5.109	3.846
American Naturalist	6	6.819	4.764
Ecology	6-12	12.494	10.159
Evolution	4-12	8.047	6.404
Heredity	4-12	7.291	4.582
Journal of Animal Ecology	3-6	4.344	3.675
Journal of Applied Ecology	3-6	4.216	3.587
Journal of Ecology	4-6	4.723	3.931
Journal of Natural History	4-48	3.720	3.161
Journal of Soil and Water Conservation	6	4.315	2.341
Journal of Wildlife Management	4-8	7.744	5.930
Oecologia	8-16	10.559	9.663
Oikos	6-12	7.312	5.976
Pedobiologia	6	2.378	1.908

Table 2.2: Journals that were analyzed in addition to the first set from 1980 to 2009. Listed are the number of issues published per year, the number of publications and the number of publications of the type “article” over the whole time period.

Journal title	Issues per year	Publications	Articles
African Journal of Ecology	4	1.771	1.563
Annales Zoologici Fennici	4-6	1.445	979
Behavioral Ecology and Sociobiology	4-12	4.054	3.886
Biochemical Systematics and Ecology	4-8	3.568	3.324
Biological Conservation	12	5.344	4.871
Ecological Modelling	16-24	5.913	4.662
Interciencia	6-12	2.574	1.794
Journal of Arid Environments	12	3.468	3.135
Journal of Biogeography	6-12	3.139	2.353
Journal of Chemical Ecology	6-12	5.715	5.380
J. of Exp. Marine Biology and Ecology	13-29	7.069	6.708
Microbial Ecology	4-8	2.474	2.201
Theoretical Population Biology	6-8	1.949	1.859

other time series. From 1980 to 2009 article numbers in the 27 analyzed journals increased from about 1.900 to more than 4.500 (Fig. 2.3), which is an increase of almost 140%. In the shortest time series, different numbers of articles from the 36 journals were analyzed by references, titles, abstracts and keywords, as not all documents had an abstract or keywords. But the trend was the same in all cases. To name one example, the number of articles analyzed by references rose from 3.200 in 1990 to almost 6.000 in 2009 (Fig. 2.4), which is an increase of about 90%.

The increasing article numbers are probably due to two reasons. Although this was not analyzed in the present study, it is widely known that research articles have grown shorter and shorter over time as journals have continually reduced their maximum page limit per article. Fourty years ago, research articles

Table 2.3: Journals that were analyzed in addition to the first two sets from 1990 to 2009. Listed are the number of issues published per year, the number of publications and the number of publications of the type “article” over the whole time period.

Journal title	Issues per year	Publications	Articles
Agriculture Ecosystems & Environment	12-20	4.084	3.060
Biotropica	4-6	2.527	1.809
Environmental Biology of Fishes	8-12	3.586	2.746
Journal of Freshwater Ecology	4	1.716	1.545
Landscape and Urban Planning	6-20	2.351	1.447
Marine Ecology – Progress Series	10-25	11.985	11.108
New Zealand Journal of Ecology	2-3	791	480
Paleobiology	4	1.242	959
Polar Biology	8-12	2.878	2.728

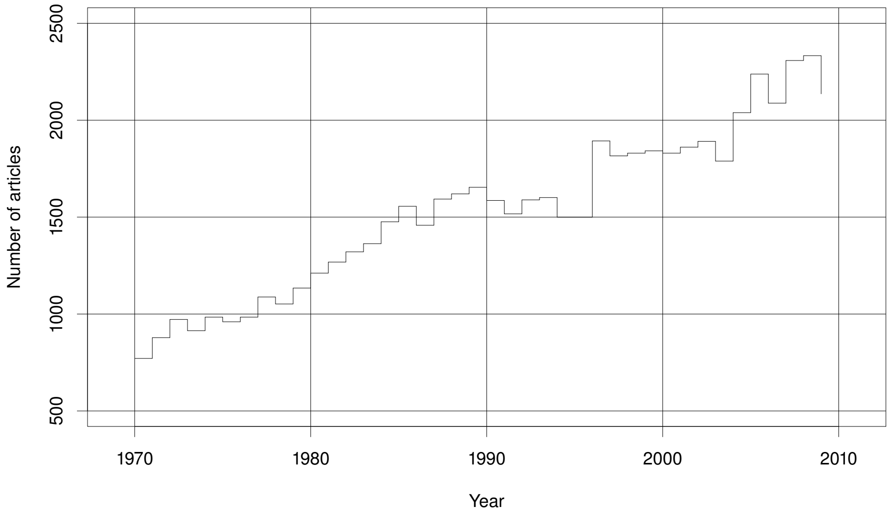

Figure 2.2: Number of articles per year in the 14 journals analyzed in the 40-year time series (1970-2009).

of more than 20-30 pages in length were quite common. Today, most journals “force” their authors to keep their papers within strict page limits (usually 10 pages at the most) and take quite high charges for any page exceeding that limit.

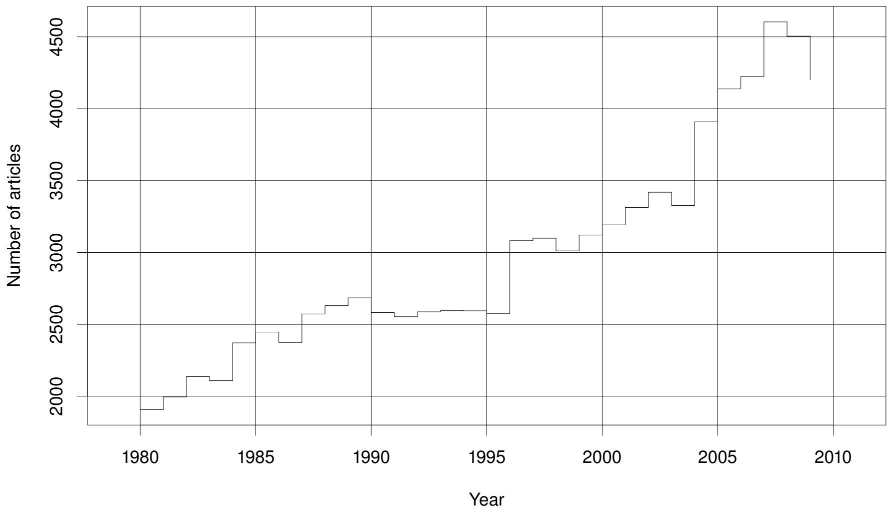

Figure 2.3: Number of articles per year in the 27 journals analyzed in the 30-year time series (1970-2009).

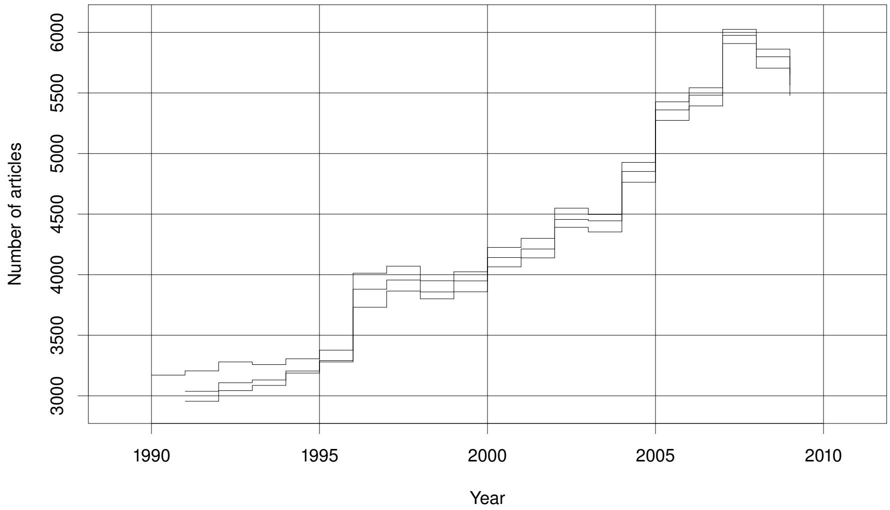

Figure 2.4: Number of articles per year in the 36 journals analyzed in the 20-year time series (1990-2009). Due to the fact that not every article had an abstract or keywords, the number of articles that were analyzed by abstract words (dashed line) and by keywords (dash-dotted line) was usually smaller than the number of articles analyzed by references and titles (solid line). Keywords and abstract were available in the *Web of Science* only from 1991 onwards.

A second reason, which can be guessed already from Tab. 2.1-2.3, are changes in the numbers of issues published per year. There is hardly any journal which has not increased its issue numbers over time, and e.g. the *Journal of Ecology* and the *Journal of Animal Ecology* have increased their issue numbers from four to six per year exactly in 1995 and 1996 respectively, where a major "jump" in article numbers per year can be observed in all time series (cf. Fig. 2.2-2.4). It would be certainly worthwhile to scrutinize the development of both article lengths and issue numbers per year in more detail and to elucidate the reasons for the observed patterns, but that is beyond the scope of this study.

Not only the number of articles published per year rose over time in all three time series, but also the number of different references that were cited (Fig. 2.5), as well as the numbers of different title words (Fig. 2.6), abstract words (Fig. 2.7), and keywords (Fig. 2.7) that were used in the articles. The steepest increase was observed in the number of different references that were cited in the time series from 1970 to 2009. Their number increased from about 15.000 to 80.000, which is an increase of approximately 430%. In the two shorter time series, the number of different references cited increased not as steeply, but still by 260% and 124%, respectively. Similar patterns were observed for the number of different title words, which increased by 80%, 70%, and 36% in the three time series, respectively. In the shortest time series (1990-2009), the number of different words used in the abstract rose by 45% from about 20.600 to around 29.800 and the number of different keywords used increased by 75% from ca. 8.200 to approximately 14.400.

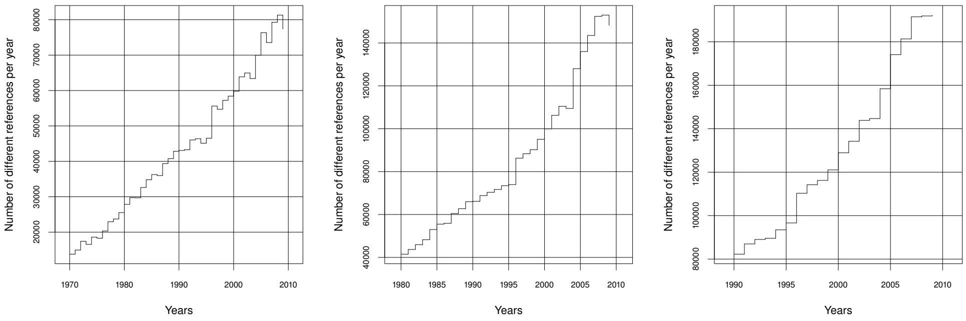

Figure 2.5: Number of different references cited in the three time series (from left to right: 1970-2009, 1980-2009 and 1990-2009).

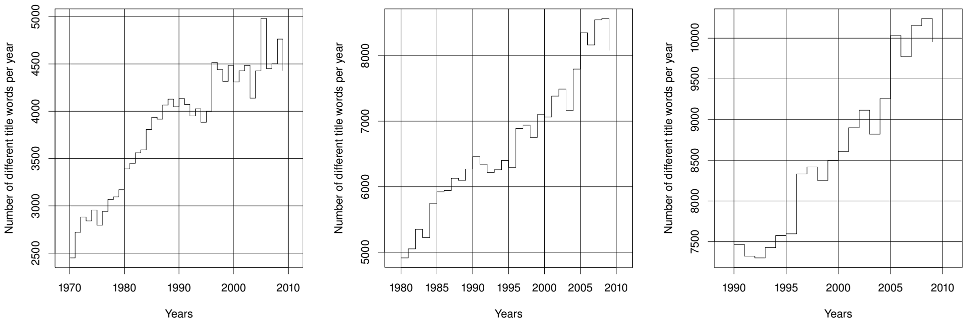

Figure 2.6: Number of different title words used in the three time series (from left to right: 1970-2009, 1980-2009 and 1990-2009).

2.2 Implementation

Both data preparations and analyses were performed with the free statistics software *R* (R Development Core Team, 2009, 2011). The additional package *corpcor* (Schaefer *et al.*, 2010) was included for calculating the SVD and the package *Snowball* (Hornik, 2009) was included for word stemming. All *R*-programs were run in batch mode on a server (running Ubuntu 9.10) at the Humboldt Univer-

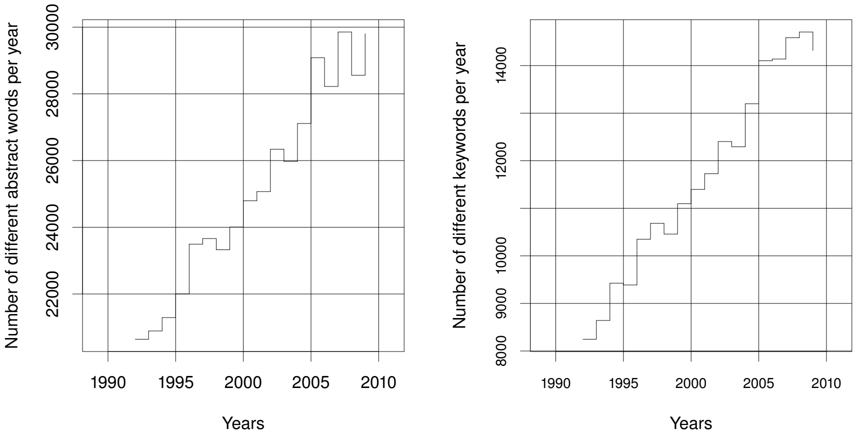

Figure 2.7: Number of different abstract words (left) and key words (right) used in the shortest time series

sity Berlin by using the standard UNIX-programs *SSH* and *NOHUP*. All graphs were also produced in *R* under Ubuntu 10.04 on a customary netbook.

After download, the text files (each containing metadata of 500 documents) were treated according to the following rules, in order to prepare the data for further analysis in *R*:

- All degree symbols (°) were substituted by the string *KRINGELCHEN* and all quotation marks were substituted by the string *ANFZO* (an abbreviation of the German “Anführungszeichen oben”).
- All tabulators, which were initially chosen as field separators in the downloaded files, were replaced by °|° in order to make | the new field separator and ° the intermediate for encoding text (characters).
- All empty fields were substituted by NA, which is the indicator for empty fields in *R*.
- Finally, all degree symbols (°) were replaced by quotation marks again, which were then the final indicator of text.

For each time series, separate scripts were programmed for analyzing references, titles, abstracts, and keywords, respectively. The basis of these scripts was developed by O. Mitesser and can be found in the appendix (“Anhang A”) of his master’s thesis (Mitesser, 2008). The majority of alterations that were made in the current scripts were standardizations and randomizations, which will be explained in section 2.4 and section 2.5. Some example scripts are included in the Appendix 5 of this study, as well. The method as such (LSA, SVD) is explained in full detail in (Mitesser, 2008), but the basic principles of the analysis will be illustrated in the following section.

2.3 Basic principles of the analysis

As there were different numbers of articles published in the different years of analysis, samples of a constant size (500 articles) were randomly chosen from the dataset in each year. For the SVD, which is the core of the LSA, a document-reference-matrix was generated from the sample data, in which each row corresponds to a document and each column to a reference. (The same procedure was applied also for the document-term-matrices in the case of title, abstract and keywords, but for the sake of brevity, the general procedure will be explained for the references only.) For this purpose, a vector of all references that were cited in the 500 documents was generated and duplicate references were removed. The dimension of the resulting matrix was then 500 rows times as many columns as there were different references in the sample.

The matrix describes, which reference is cited in which document. Thus, in each row a column contains a “1”, when the respective reference was cited in the document and a “0” when it was not cited. As already mentioned in the introduction, such a matrix is usually sparse, i.e. it contains many more zeros than ones, as the majority of references is cited only in one document.

For the SVD, the matrix has to be transposed, so that the documents correspond to the m columns and the references to the n rows. The SVD (by means of the R-function *fast.svd*) decomposes the reference-document-matrix X $(n \times m)$ into three separate matrices which can be described like this:

$$X = U\Lambda^{1/2}V^{T}. \tag{2.1}$$

The r columns of the matrix U $(n \times r)$ contain the standardized eigenvectors of the matrix XX^T, which describes the co-citation relationships between the references. U thus represents the relationship between the r derived topics and the n cited references. The r columns of the matrix V $(m \times r)$ contain the standardized eigenvectors of the matrix X^TX, which describes the bibliographic coupling of the documents. V thus represents the relationship between the r derived topics and the m documents. Usually, the SVD yields as many topics as there were documents in the matrix, so V is a square matrix. The third matrix, $\Lambda^{1/2}$ $(r \times r)$, is a diagonal matrix and contains the square roots of the r eigenvalues $\lambda_k > 0$, which are common to both matrices XX^T and X^TX.

These eigenvalues are in the following used to calculate the diversity of each sample with Eq. 1.1 and Eq. 1.2, because the relative “abundances” p_k of each of the r topics in the sample can be obtained by dividing the eigenvalue λ_k of each topic by the sum of all eigenvalues in the sample:

$$p_k = \frac{\lambda_k}{\sum\limits_{k=1}^{r} \lambda_k} \tag{2.2}$$

The whole procedure was repeated 50 times for each year to obtain means and standard deviations for the measured diversities.

2.4 Standardization

The basic principles were so far the same as developed by Mitesser (2008) in his master's thesis. In the present study, several standardization procedures were included for different reasons.

First of all, both the columns and the rows of the document-reference-matrix (likewise also the document-term-matrices in the case of title, abstract and keywords) were standardized, in order to (1) reduce the influence of highly cited references and (2) to alleviate the effects of differences in the length of the reference list.

In a first step, the n columns of the matrix X were standardized by replacing the ones in the m rows by the *inverse document frequency* (idf) of the references, i.e. the references were weighted by their general importance in the set of documents. This was accomplished by dividing the total number of documents m by the number of documents citing the reference and then taking the logarithm of this fraction:

$$idf_j = \log_{10} \frac{m}{\sum\limits_{i=1}^{m} X_{ij}} \tag{2.3}$$

In a second step, the m rows of the matrix X were standardized by dividing each cell in a row by the square root of the sum over the squared idf values from the whole row:

$$\tilde{X}_{ij} = \frac{X_{ij}}{\sqrt{\sum\limits_{j=1}^{n} X_{ij}^2}} \tag{2.4}$$

The second kind of standardization was necessary only when natural words were analyzed, i.e. in the case of title, abstract, and keywords. Each of the three was read from the text files into *R* as one single string per article, in which the different words were separated by spaces and punctuation marks. Using regular expressions the strings were split into single words by means of the *R*-function *strsplit*. A similar string splitting procedure was already used in the references-based analysis, but there it was clearly defined that the different references were always separated by a semicolon.

After splitting the strings, all terms had to be changed to lowercase letters, as *R* is a case-sensitive language and also because the word stemming procedure (see below) could deal only with lowercase words. The fact that *R* is case-sensitive mattered especially for word comparison. On the one hand, the function used for removing stop words had to compare the words that resulted from string splitting with the words in the stop word list, in order to remove those from the term list, that were contained in the stop word list. As especially the older entries in the WoS, but also the entries of some journals, were all in uppercase letters or sometimes all title words began with uppercase letters, it was the simplest solution to change all words to lowercase in order to be able to reliably remove stop words. On the other hand, the document-term-matrix was generated by comparing which term from the list of different words was contained in the list of words derived from each document. Also in this case it was important to be able to execute this comparison reliably.

Word stemming was performed with the *R*-function *SnowballStemmer* after splitting strings and changing them to lowercase letters. Some random samples were checked in order to test the "success" of the stemming procedure but no systematic checks were performed. Details of the stemming algorithm and its development can be found on the website of its developer, Martin Porter: http://snowball.tartarus.org/.

2.5 Randomization

In order to scrutinize whether the eigenvalues resulting from the SVD can reliably be used to measure the diversity of a research field, three alternatives of randomizations were tested. All tests were performed using the data set of the longest time series (1970-2009) based on references.

In the first alternative, the list of unique references was derived as usual and the length of the reference list of each document was determined. Then, the document-reference-matrix was established by randomly assigning as many references to each document as its reference list had originally contained. Finally, the SVD was performed and diversities were calculated as described above. This randomization procedure kept the distribution of the lengths of the reference lists constant, but changed the overall citation frequencies of the references. The latter was on the one hand due to the fact that by way of sampling from the list of unique references, each of them had the same chance of being chosen, and on the other hand, not every reference from the list got sampled at all, as there were commonly between 10.000 and 20.000 different references per sample and each individual reference list was typically only between 20 and 50 references long.

Therefore, the second randomization scenario was changed compared to the first in terms of the sampling method. Again, the references were randomly assigned to each document based on the length of its reference list, but the sampling probabilities of the list of unique references were chosen based on their citation frequencies in the sample. Thus, references that were cited by more than one document in the original sample had a higher probability of being sampled in the randomization process. This approximated the citation frequencies of the references in the original data set, but now the very rare references had an even lower chance of being sampled at all.

Finally, in the third alternative, both the lengths of the reference lists and the citation frequencies of the references were kept the same as in the original sample. This was accomplished by first deriving the document-reference-matrix from the original data and then swapping references between pairs of documents. In 100 cycles one pair of references was swapped between each of the 250 pairs of documents. Both the document pairs and the pairs of references that were to be swapped were randomly assigned in each cycle of the randomization.

3 Results

In the first four sections of this chapter, the results from the diversity calculations in the three time series are presented separately by cited references (section 3.1), title words (section 3.2), abstract words (section 3.3), and keywords (section 3.4). The fifth section (section 3.5) deals with the results of the randomization experiments, that were exemplary performed based on cited references in the longest time series (1970-2009). The last section (section 3.6), explores the suitability of LSA for deriving research topics based on cited references and title words using two examples from the longest time series.

3.1 Diversity based on cited references

The observed trends in diversity based on cited references were similar in the three studied time series. No major differences could be observed between the two diversity indices *Shannon* and *Simpson*. Interestingly, in the longest time series (Fig. 3.1), diversity seemed first to decline from 1970 until 1984 before it began to increase continually until the end of the time series. The effect size, however, was very small as the observed diversities ranged only between 99.78% and 99.87% of the maximum possible diversity in case of the Shannon index and between 99.7955% and 99.7980% for the Simpson index.

From the right panel of Fig. 3.1 it can be seen that both the mean number of references per article and the mean number of different references per article increased over time from about 20 to 50 (i.e. by 150%) in the former and from about 20 to 45 (i.e. by 125%) in the latter case. The difference in these trends already gives a hint that there has been a change in citation frequencies such that in earlier years almost every reference was cited only by one document and later on the number of references that were cited more than once increased.

In the other two time series, the initial decline in diversity was not observed, but both from 1980 to 2009 (Fig. 3.2) and from 1990 to 2009 (Fig. 3.3) diversity values increased continuously over time. Here as well, the effect sizes were very small. Between 1980 and 2009, the diversity values ranged from 99.84% to

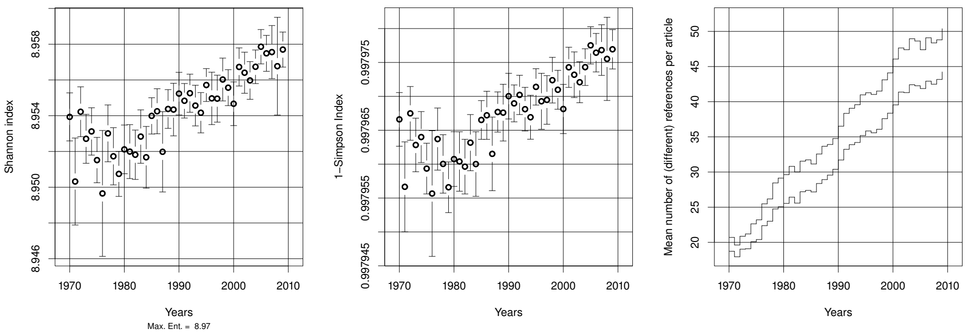

Figure 3.1: Shannon and Simpson diversity indices (left, center) based on cited references in the longest time series (1970-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of references per article (thick line) as well as of *different* references per article (thin line) are shown.

99.91% of the maximum Shannon diversity and from 99.7970% to 99.7990% of the maximum Simpson diversity. Between 1990 and 2009 the diversity values ranged from 99.88% to 99.92% of the maximum Shannon diversity and from 99.7982% to 99.7992% of the maximum Simpson diversity.

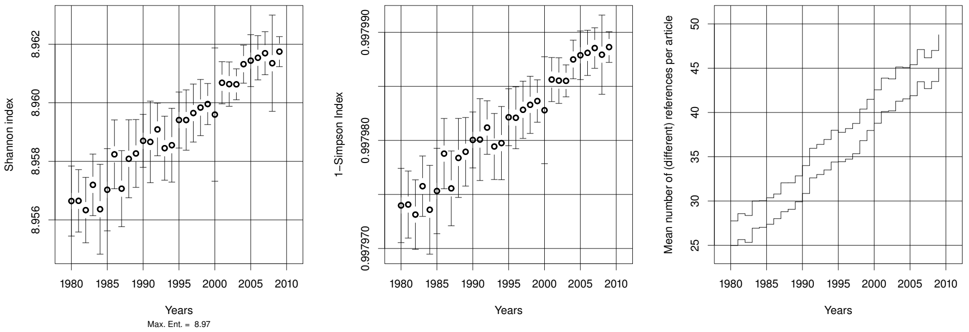

Figure 3.2: Shannon and Simpson diversity indices (left, center) based on cited references in the medium length time series (1980-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of references per article (thick line) as well as of *different* references per article (thin line) are shown.

In the medium and short time series, the increases in the number of references and different references per article were not as different as in the longest time series (right panels of Fig. 3.2 and Fig. 3.3). From 1980 to 2009, the number of references per article increased from about 28 to 48 (i.e. by 71%) while the number of different references per article increased from about 25 to 45 (i.e. by 80%). From 1990 to 2009, the number of references per article increased from about 33 to 47 (i.e. by 42%) while the number of different references per article

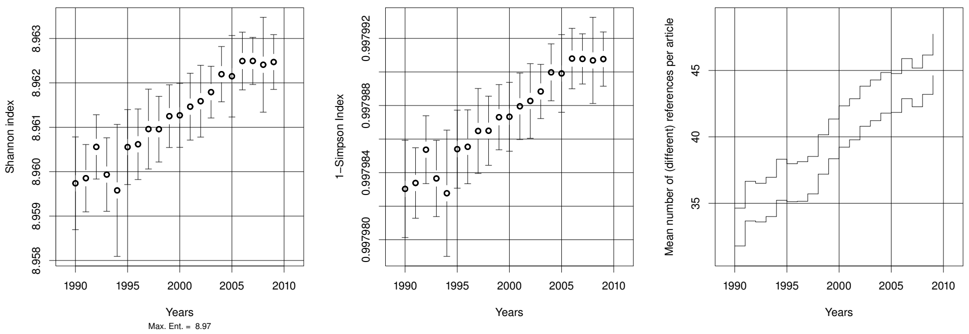

Figure 3.3: Shannon and Simpson diversity indices (left, center) based on cited references in the shortest time series (1990-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of references per article (thick line) as well as of *different* references per article (thin line) are shown.

increased from about 31 to 45 (i.e. by 45%). Thus, there were probably no major changes in the number of multiply-cited references over time in these two time series.

3.2 Diversity based on title words

The observed trends in diversity based on title words were somehow similar in the three studied time series but very different from the trends based on cited references. In the case of title words, too, no major differences could be observed between the two types of diversity indices. In all three time series a clear decline in diversity could be observed towards the end of each time series, while the earlier years showed somewhat mixed trends.

In the longest time series, diversity first seemed to increase for some years and then it oscillated around a certain value for several years before it finally declined from about year 2000 onwards (Fig. 3.4). Again, the effect size was very small as the observed diversities ranged only between 97.32% and 97.88% of the maximum possible Shannon diversity and between 99.725% and 99.740% of the Simpson diversity. The diversities based on title words were persistently lower than the ones based on cited references.

The mean number of title words per article increased over time from about 7.4 to 9.4 (i.e. by 27%) while the number of different title words per article stayed rather constant at around 3.7-4.0 (right panel of Fig. 3.4). This also suggests a slight increase in the number of multiply used title words over time.

In the medium time series, diversity first increased from 1980-1990 and from then on declined until the end of the time series (Fig. 3.5). Diversities ranged

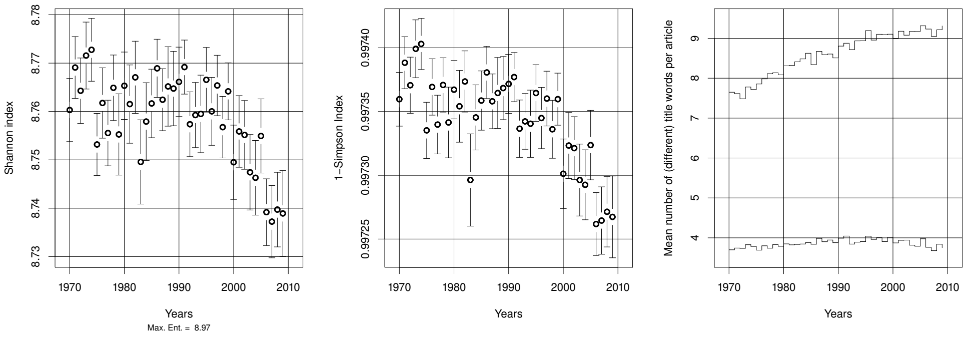

Figure 3.4: Shannon and Simpson diversity indices (left, center) based on title words in the longest time series (1970-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of title words per article (thick line) as well as of *different* title words per article (thin line) are shown.

between 97.94% and 98.16% of the maximum possible Shannon diversity and between 99.742% and 99.750% of the Simpson diversity. The mean number of title words per article increased over time from about 8.3 to 9.8 (i.e. by 18%) while the number of different title words per article increased only from about 4.1-4.4 (i.e. by 7.3%) (right panel of Fig. 3.5). This again suggests a slight increase in the number of multiply used title words over time.

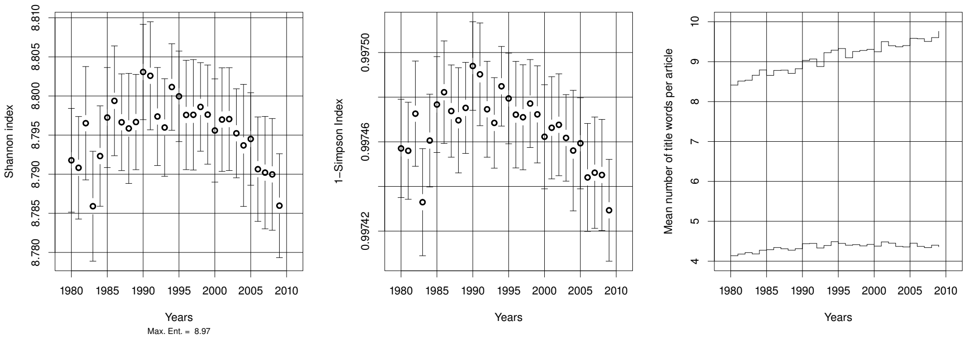

Figure 3.5: Shannon and Simpson diversity indices (left, center) based on title words in the medium length time series (1980-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of title words per article (thick line) as well as of *different* title words per article (thin line) are shown.

The shortest time series (Fig. 3.6) rather resembles the longest again, with a short decrease of diversity in the beginning, then a rather long plateau until 2002 after which diversity decreased again. Diversities ranged between 98.04% and 98.22% of the maximum possible Shannon diversity and between 99.746% and 99.752% of the Simpson diversity. The mean number of title words per article increased over time from about 9.0 to 9.8 (i.e. by 8.9%) while the number of different title words per article stayed rather constant around 4.4-4.6 (right panel of Fig. 3.5), which might correspond to a slight increase in the number of multiply used title words over time.

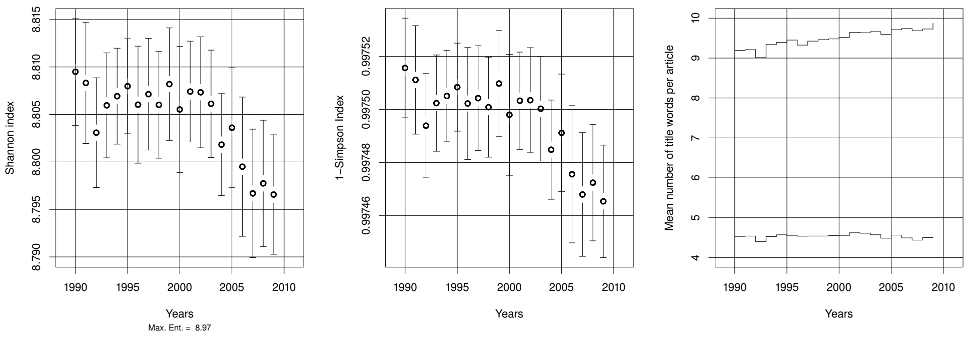

Figure 3.6: Shannon and Simpson diversity indices (left, center) based on title words in the shortest time series (1990-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of title words per article (thick line) as well as of *different* title words per article (thin line) are shown.

3.3 Diversity based on abstract words

Abstract-based diversity could only be studied in the shortest time series, as abstracts were only available in sufficient numbers from 1991 onwards. In this case, diversity clearly and continuously decreased over time (Fig. 3.7). As expected, the effect size was very small with Shannon diversities ranging between 98.05% and 98.33% and Simpson diversities between 99.685% and 99.715% of the maximum. The mean number of abstract words per article increased over time from about 76 to 92 (i.e. by 21%) while the number of different abstract words per article stayed rather constant around 14.9-15.6 (right panel of Fig. 3.7), which suggests a slight increase in the number of multiply used abstract words over time.

3.4 Diversity based on keywords

Finally, diversity based on keywords was also only studied in the shortest time series, as both author keywords and *Keywords Plus*® (i.e. keywords that were assigned to the article metadata by Thomson Reuters) were available in sufficiently high numbers only from 1991 onwards. Keyword-based diversity increased over time but seemed to reach some kind of saturation from 1995/1996 onwards,

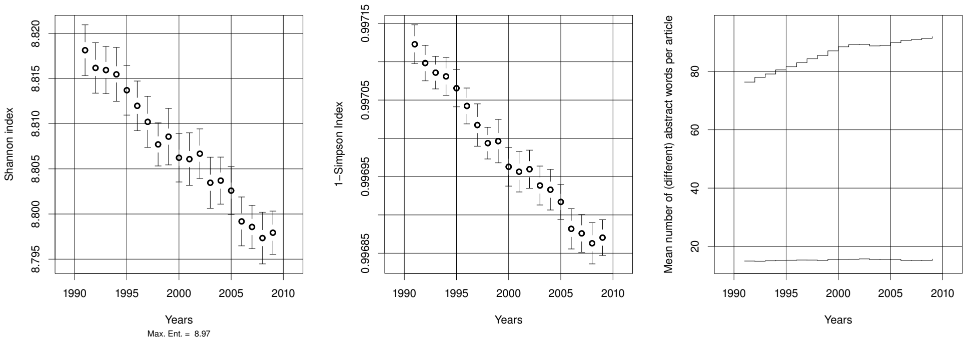

Figure 3.7: Shannon and Simpson diversity indices (left, center) based on abstract words in the shortest time series (1991-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of abstract words per article (thick line) as well as of *different* abstract words per article (thin line) are shown.

depending on which diversity index was considered. Shannon diversities ranged from 97.78% to 98.10% and Simpson diversities from 99.730% to 99.740% of the maximum (Fig. 3.8). Looking at the right panel of Fig. 3.8, the number of keywords per article rose from about 14 to 22 (i.e. by 57%), whereas the number of different keywords per article increased from about 5.6 to 7.4 (i.e. by 32%). The steeper increase in the former parameter suggests an increase in the number of keywords that were used by more than one article in the sample.

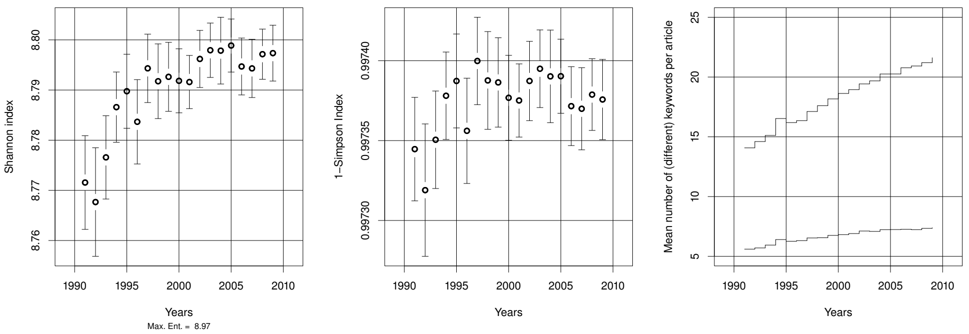

Figure 3.8: Shannon and Simpson diversity indices (left, center) based on keywords in the shortest time series (1991-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. On the right, mean numbers of keywords per article (thick line) as well as of *different* keywords per article (thin line) are shown.

3.5 Randomization experiments

In the first two randomization scenarios, a continuous increase of both Shannon and Simpson diversity over time was observed (Fig. 3.9, rows 1-2). Compared with the original data set, the initial decrease of diversity until 1984 had disappeared after randomization. In the first scenario, Shannon diversities ranged from 99.61% to 99.83% and Simpson diversities from 99.7910% to 99.7965% of the maximum values. In the second scenario Shannon and Simpson diversities ranged from 99.59% to 99.80% and from 99.7908% to 99.7960%, respectively. In both scenarios, the observed diversities were consistently lower than in the original data set, with the values in the second scenario being always lower than in the first scenario (Fig. 3.10).

As can be seen from the right panels of Fig. 3.9, the number of references per article stayed the same as in the original data set (minor differences are due to variation in the samples) and likewise increased from about 20 to 50 (i.e. by 150%) over time. However, as already suggested in section 2.5, the number of different references per article was consistently lower than in the original data set (cf. Fig. 3.1 or the right panel of the third row in Fig. 3.9, which is basically the same). In the first scenario it increased from about 12.5 to 30 (i.e. by 140%), and in the second it increased from about 12 to 29 (i.e. by 142%).

In the third randomization scenario, the trend in diversity over time (Fig. 3.9, third row) resembled the trend in the original data set, however, diversity values were consistantly higher after randomization. Shannon diversities ranged from 99.87% to 99.92% and Simpson diversities from 99.7983% to 99.7992% of the maximum values. As in this scenario both the lengths of the reference lists and the citation frequencies of the references were kept the same than in the original data, both the number of references per article and the number of different references per article showed the same trend than in the original time series (cf. Fig. 3.1, again minor differences are due to sample variation).

3.6 Extracting research topics by means of LSA

Initially, LSA was developed to extract topics or themes from documents based on the terms they contain. Thus, after looking at all the LSA-based diversity calculations in the previous sections, this last section shows the results of two tests of extracting topics from research papers using cited references or title words.

Each of the tests was based on a sample of 500 articles from year 2009, which was in both cases taken from the longest time series, i.e. documents were taken

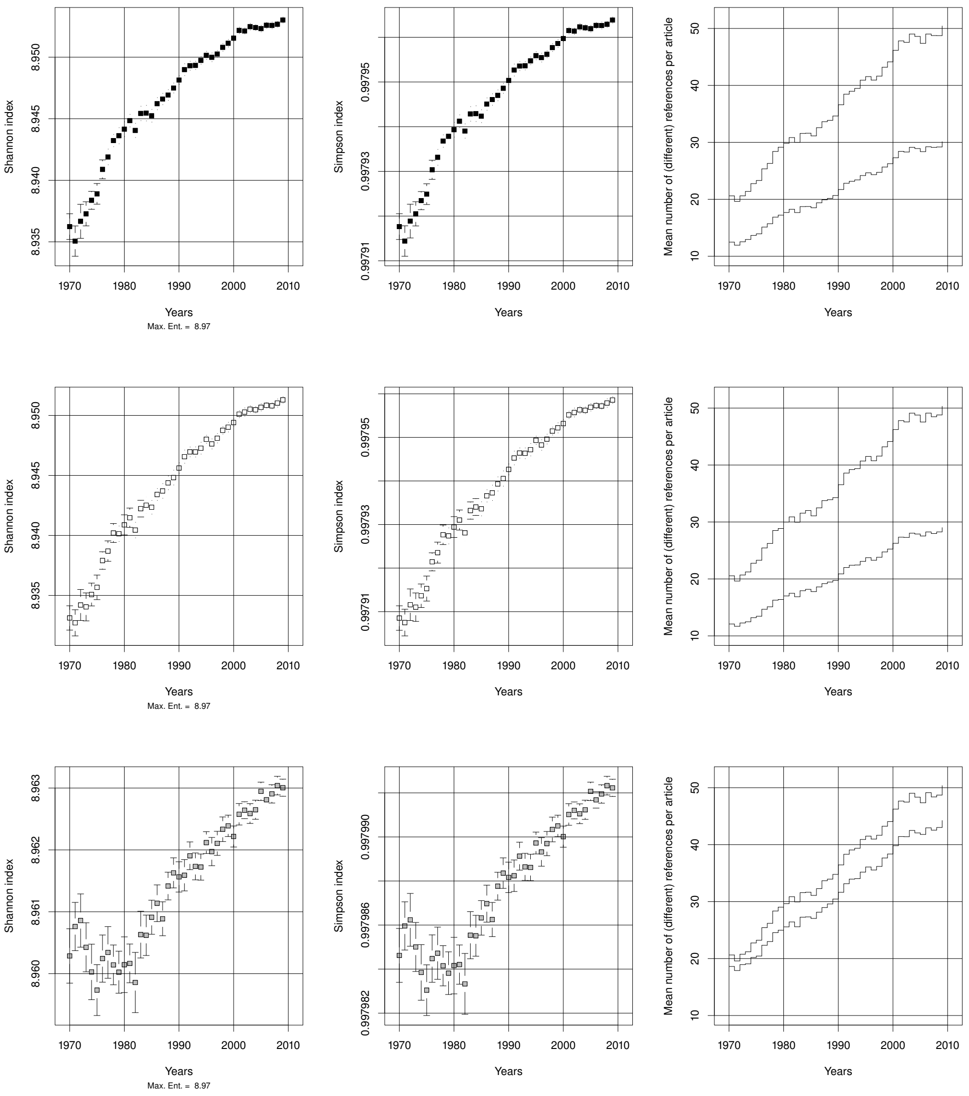

Figure 3.9: Shannon and Simpson indices (left, center) of the three randomizations scenarios based on references (1970-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. The three rows show the three scenarios (1 = first row, black squares; 2 = second row, open squares; 3 = third row, grey squares). In each row, the third panel shows the number of references per article (thick line) and the number of different references per article (thin line).

from a set of 14 journals (Table 2.1). The first test used the cited references and the second used the title words for the LSA.

The largest 40 (out of 500) eigenvalues of each SVD are shown, the ones based on references in Table 3.1 and the ones based on title words in Table 3.2. As can be seen from the tables, the eigenvalues are quite low in both examples and in case of the references-based approach, there is no topic that has a larger

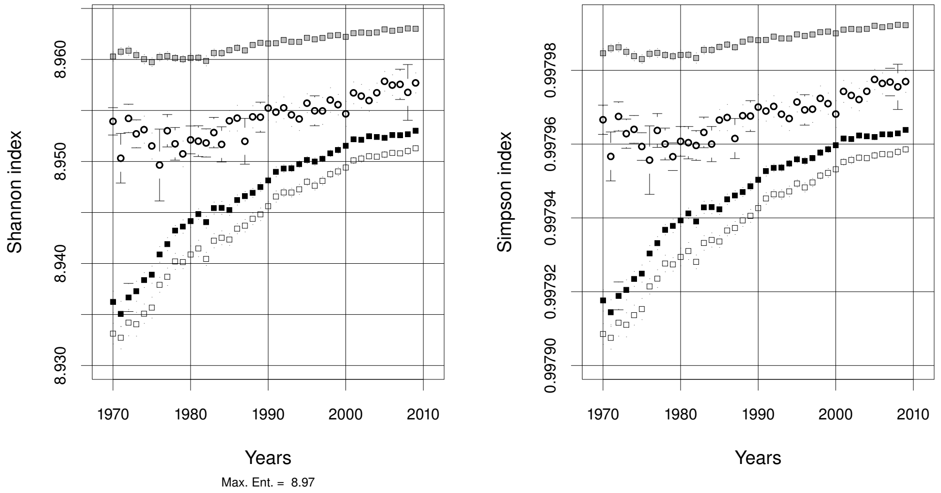

Figure 3.10: Shannon and Simpson indices (left, center) based on references (1970-2009). Plotted are the means and standard deviations from 50 samples (sample size = 500 articles) per year. Original data (open circles) are compared with the three randomization scenarios (1 = black squares, 2 = open squares, 3 = grey squares).

share than 5% in more than 10 articles. In the example based on title words the four largest topics have a share larger than 5% in 10-13 articles.

A look at the maximum share of topics in those articles in which they have the largest share, shows rarely values larger than 20%, and only in the references-based example three topics have a share larger than 50%. Moreover, all values in the second-last column, which shows the share of those articles, in which a topic was maximum, among all papers that had a share in that topic at all, are way below 5%. Thus, based on this method there are commonly about 100-200 documents that have a share in the same topic.

Finally, the last column shows the topics that were “manually” extracted from the titles of those articles in which each topic had a maximum share. In those cases, in which the topic was only maximum in one article, or in which no proper topic could be figured out, the respective rows are left empty. For example, in the largest topic of the references-based example, the titles of the ten articles, in which the topic was maximum, were the following:

1. “Sapling herbivory, invertebrate herbivores and predators across a natural tree diversity gradient in Germany’s largest connected deciduous forest”

2. “Species interaction mechanisms maintain grassland plant species diversity”

3. “Disperser limitation and recruitment of an endemic African tree in a fragmented landscape”
4. “Local neighborhood and species’ shade tolerance influence survival in a diverse seedling bank”
5. “Interspecific variation in seedling responses to seed limitation and habitat conditions for 14 Neotropical woody species”
6. “Spruce colonization at treeline: where do those seeds come from?”
7. “Abiotic and biotic drivers of seedling survival in a hurricane-impacted tropical forest”
8. “Beyond description: the active and effective way to infer processes from spatial patterns”
9. “On the emergent spatial structure of size-structured populations: when does self-thinning lead to a reduction in clustering?”
10. “Recruitment in tropical tree species: revealing complex spatial patterns”

Although most of the titles cleary deal with closely related things, it is not as straightforward to write down “the” common topic. Thus, “plant species diversity” and “spatial patterns” are somehow umbrella terms, that embrace all the topics but are not that very specific themselves. This problem gets worse with decreasing eigenvalues of the topics.

Table 3.1: List of the fourty largest topics extracted by LSA based on references from a sample of 500 documents in year 2009. Given are the rank (Rk) and the eigenvalue (Ev) of each topic, the number of documents in which the share of the topic is larger than 5% (Nr), and the number of documents in which the topic has the largest share (Nmax). The following columns all refer to the documents, in which the topic has the largest share. For those, the maximum, median and minimum share of the topic (MaxS, MedS, and MinS, respectively) are given. Furthermore, the share of those papers (Nmax) in all papers that have any share in the topic is presented in column 8 (Share). The last column shows the (possible) topic that was manually derived from the titles of the papers in which the topic had the largest share.

Rk	Ev	Nr	Nmax	MaxS	MedS	MinS	Share	Topic
1	1.249	4	10	14.3	3.9	1.0	2.2	plant species diversity, spatial patterns
2	1.208	8	11	17.4	4.9	0.5	2.4	diversity
3	1.188	2	9	4.2	3.1	0.3	1.9	habitat selection, survival
4	1.177	2	3	64.7	64.1	0.4	0.6	reproduction in oaks
5	1.176	2	3	68.2	67.7	0.7	0.6	parasitic worms
6	1.173	9	8	20.9	9.8	5.3	1.7	plant-pollinator/seed disperser-interactions, mutualism
7	1.167	7	1	1.3	1.3	1.3	0.2	
8	1.158	7	10	22.2	6.6	1.4	2.1	competition and facilitation in plant communities
9	1.156	4	7	10.4	4.1	0.1	1.5	mixed-effects models (method)
10	1.148	2	5	22.1	4.2	0.5	1.1	habitat use of birds
11	1.144	7	9	11.1	3.7	0.1	1.9	predator-induced effects on physiology
12	1.139	7	1	6.0	6.0	6.0	0.2	
13	1.136	5	6	16.4	2.1	0.0	1.3	reproductive strategies in insects, Wolbachia infection
14	1.128	3	12	10.2	2.4	0.3	2.5	phenotypic evolution
15	1.125	6	4	10.4	8.5	5.7	0.8	biogeography of plants
16	1.124	5	7	27.4	4.4	1.5	1.5	estimating population size in natural populations
17	1.123	5	5	16.2	8.2	1.2	1.1	self-incompatibility in plants
18	1.121	4	4	6.1	3.5	0.2	0.8	spatial heterogeneity
19	1.121	6	7	17.9	7.6	0.2	1.5	mycorrhiza
20	1.114	9	9	9.8	5.1	0.1	1.9	coevolution
21	1.112	6	0					
22	1.107	0	10	4.9	1.6	0.2	2.1	genetic population structure
23	1.103	4	4	5.7	4.6	2.8	0.8	habitat selection, home range size, foraging niche in herbivores
24	1.103	3	1	5.7	5.7	5.7	0.2	
25	1.100	1	7	5.0	3.2	0.1	1.5	competition
26	1.099	3	6	5.4	2.7	0.0	1.2	
27	1.096	5	7	8.5	6.3	1.2	1.5	functional community ecology
28	1.091	5	5	15.1	1.6	0.0	1.0	evolution
29	1.090	2	5	10.3	3.7	0.1	1.0	natural enemies
30	1.087	5	7	13.6	2.3	0.2	1.5	inbreeding, outbreeding, fertility
31	1.086	4	4	18.0	14.8	0.3	0.8	habitat heterogeneity
32	1.085	4	8	10.1	3.5	0.0	1.7	seed dispersal
33	1.083	3	4	14.5	10.4	2.7	0.8	competition in plants
34	1.082	0	14	3.7	1.9	0.0	2.9	reproductive strategies
35	1.082	2	2	1.4	1.3	1.1	0.4	
36	1.079	1	4	1.6	0.8	0.3	0.8	invasive species
37	1.076	2	7	12.4	2.3	0.6	1.5	water availability
38	1.076	4	4	8.5	3.6	0.0	0.8	sexual selection
39	1.075	2	2	57.8	57.8	57.8	0.4	Platygastridae of the British Isles
40	1.075	0	13	4.6	0.8	0.0	2.7	population ecology, density-dependence

Table 3.2: List of the fourty largest topics extracted by LSA based on title words from a sample of 500 documents in year 2009. Given are the rank (Rk) and the eigenvalue (Ev) of each topic, the number of documents in which the share of the topic is larger than 5% (Nr), and the number of documents in which the topic has the largest share (Nmax). The following columns all refer to the documents, in which the topic has the largest share. For those, the maximum, median and minimum share of the topic (MaxS, MedS, and MinS, respectively) are given. Furthermore, the share of those papers (Nmax) in all papers that have any share in the topic is presented in column 8 (Share). The last column shows the (possible) topic that was manually derived from the titles of the papers in which the topic had the largest share.

Rk	Ev	Nr	Nmax	MaxS	MedS	MinS	Share	Topic
1	2.461	13	16	6.8	3.1	1.0	3.2	poulation dynamics (in unstable environments)
2	1.874	10	19	21.4	5.8	1.5	3.8	diversity of communities
3	1.779	11	17	26.0	5.6	0.4	3.4	biology/ecologgy of trees/forests
4	1.744	12	14	9.0	3.7	0.5	2.8	multitrophic interactions
5	1.722	9	19	8.7	4.1	0.5	3.8	predation
6	1.688	6	13	10.7	3.5	1.7	2.6	reproductive success
7	1.668	7	14	9.8	4.3	1.3	2.8	genetic variation/diversity
8	1.655	8	16	7.8	5.1	1.3	3.2	sexual selection
9	1.635	9	17	10.7	4.9	0.9	3.4	dispersal in hetereogeneous landscapes
10	1.622	3	10	6.7	3.3	0.6	2.0	density dependence
11	1.593	3	8	7.0	2.4	0.8	1.6	(a)biotic interactions
12	1.585	2	11	8.0	1.8	0.7	2.2	
13	1.575	4	11	8.6	3.6	0.9	2.2	
14	1.563	2	5	9.3	2.5	1.3	1.0	
15	1.556	4	12	11.5	3.8	0.6	2.4	population genetics
16	1.540	3	12	5.6	3.0	0.8	2.4	distribution patterns
17	1.530	1	4	5.3	3.1	1.5	0.8	
18	1.521	1	8	5.8	2.2	0.3	1.6	
19	1.508	0	5	4.4	3.4	1.9	1.0	reproduction
20	1.506	2	6	6.7	3.1	1.1	1.2	genetics
21	1.495	1	4	4.3	3.7	3.1	0.8	
22	1.490	0	4	3.7	3.0	2.3	0.8	
23	1.484	0	6	3.2	2.4	1.6	1.2	
24	1.470	1	8	6.6	3.0	0.4	1.6	community ecology
25	1.469	2	9	7.2	2.9	1.1	1.8	biodiversity, interactions in communities
26	1.466	1	0					
27	1.456	2	6	10.8	3.3	1.1	1.2	(home) range
28	1.447	3	7	6.8	2.4	1.2	1.4	forests
29	1.443	0	3	4.4	1.5	1.2	0.6	
30	1.439	1	5	5.3	3.0	1.4	1.0	adaptation
31	1.431	1	6	5.2	2.2	1.5	1.2	competitive effects
32	1.428	0	4	3.1	2.8	1.6	0.8	
33	1.423	0	11	3.9	2.5	1.2	2.2	sex
34	1.417	2	8	5.1	3.0	1.1	1.6	size
35	1.408	0	4	4.2	3.2	0.3	0.8	
36	1.402	0	8	4.8	3.0	2.0	1.6	
37	1.398	2	9	5.5	2.9	1.3	1.8	
38	1.390	2	3	5.7	5.6	2.5	0.6	
39	1.389	0	8	3.8	2.0	0.8	1.6	
40	1.381	1	5	7.4	3.1	1.0	1.0	

4 Discussion

This thesis set out to scrutinize the suitability of *Latent Semantic Analysis* for measuring the diversity of research, using the field of ecology as an example. To this end, three time series, beginning in 1970, 1980, and 1990, respectively, and all ending in 2009 were studied and diversities based on cited references, title, abstract, and keywords were calculated by means of LSA and the Shannon and Simpson indices.

The results of the "pilot study" by Mitesser (2008) suggested that LSA may well be suitable for measuring diversity, but he already recommended some further scrutiny in order to make sure that the observed diversity patterns were not due to merely statistical properties of the document-reference-matrix (or document-term-matrix) only. Therefore, in addition to comparing the results of the diversity calculations based on different document properties, three randomization scenarios were analyzed in order to establish, which proportion of the measured diversity is due to statistical properties of the matrix. Some properties in question were the dimension of the matrix, i.e. the number of different references in the sample, the citation frequencies of the different references, or the length of the reference lists of the documents, although the influence of the latter has already been analyzed by Mitesser (2008) and the impact of the last two properties has been attenuated by the standardization procedures described in section 2.4.

4.1 Trends in diversities

Supposed, the eigenvalues derived by LSA were a sound basis for measuring research diversity, it obviously makes a difference, which document property is used for the analysis, as the diversity calculations based on cited references, title, abstract, or keywords yielded overall quite different results, whereas the trends for references and title words were pretty much the same in the three different time series. Based on references and keywords, diversity seemed to increase over time, whereas it decreased when title or abstract words were examined. Possible reasons for these differences are not obvious and it is difficult to decide which of the four parameters is the most useful to estimate research diversity.

The trends observed in the references- and keywords-based approaches support the observation made also by Neff & Corley (2009) that ecology is still a flourishing and expanding field. Looking at the trend in keyword-based diversity, it might however have reached a plateau in the latest years. Diversities estimated by words from the title and abstract, on the contrary, favor the hypothesis that research diversity is declining – possibly due to a concentration on topics that are favored by research funding agencies (Harley & Lee, 1997; Adams & Smith, 2003). Consequently, from the current state of the analysis, it cannot be decided which parameter is the "right" one to estimate diversity and further studies will be needed. Maybe some well-considered combination of all four (or even some other) document properties will give the best results.

Regarding potential reasons for the observed trends in diversity, it might in the case of references be concluded from Fig. 2.5 that the increasing number of different references cited over time have led to an increase in diversity. However, the same increasing trend was observed for the other three document properties, as well (cf. Fig. 2.5 and Fig. 2.6), so it is unlikely that an increase in the number of different references or terms alone can be responsible for an increase in diversity. Likewise, more different references do not necessarily have to correspond to more different topics, as certain references might always be cited together and thus may always count as parts of the same topic.

It is very interesting that the increases in the number of different references, title words, abstract words and keywords were very much alike, as it cannot be expected that their numbers could go on growing in the same fashion forever. In the case of references, every published article automatically adds to the pool of potential references that can be cited by articles that are written and published later on. For example, the number of articles published each year in the 14 journals that were studied in the longest time series rose by almost 200% in 40 years (the increase being much steeper if all ecology journals were taken into account). Alone in these 14 journals there were 61.443 articles published from 1970 to 2009, which could be cited by articles in 2009. Supposed an article published in one of these journals in 2009 had cited only articles from within this 40-year data set, the number of articles which it could cite had increased from 1970 to 2009 by 7870%. Usually, older articles are less and less cited over time, but still, they at least could be cited. In this context it would be very fascinating to study the "age distribution" of the cited references as well as if and how it possibly changes over time.

Considering title words, the observed (more or less linear, but certainly not exponential) increases in the three time series did not correspond to the exponential increase described by Neff & Corley (2009). However in their study, articles from all ecology journals in the timespan 1970 to 2005 were included, so maybe the observed exponential increase was biased by the ever growing number of new journals that were released over time. Independent of the shape of the increase, it seems unlikely that the number of different words from title and abstract as well as keywords could keep increasing all the time like the number of different references. Thus, independently of the suitability of LSA for measuring diversity, it will be interesting to keep an eye on this development also in the future, in order to see, at what point in time the number of different words used in the documents will reach a plateau. According to the *Oxford Dictionaries*[1], there are at least 250.000 words in the English language, 20% of which are no longer in current use, which leaves us with 200.000 words. As the largest number of different words was close to 30.000 in the abstracts, there is still much scope for increase in all three studied natural-word-based variables.

Concerning words from title and abstract, the observed decrease of diversity over time might be due to a potential trend of using certain "popular" words independent of the real topic of the paper. Neff & Corley (2009) likewise suggested in their study on the evolution of ecology that authors might use (title) words to "*tie the article to a hot topic*" in order to boost its chances of being published. Another hypothesis could be that titles and articles are more and more phrased in the same fashion and extravagant phrasing, which might have been fashionable in the past, disappears. This is, however, merely a vague hypothesis which requires more detailed information to be tested.

In addition, some methodological issues may also play a role in estimating diversity based on words from title and abstract, that need to be refined in the future. On the one hand, in the present study, a "standard" stop word list was used, which might not have been appropriate enough for the analysis, as several words that have a meaning in natural language, but no special meaning in terms of research diversity were not considered as stop words. Such words could be e.g. *method*, *analysis*, *experiment*, *effect*, or *statistic*, though the question of relevance might be hard to decide in some cases. On the other hand, the procedure of splitting title and abstract into single words also separates groups of words that might be coined terms, which e.g. might describe a special kind of

[1] http://www.oxforddictionaries.com/page/93 (last visited on 25.04.2011)

method or experiment. This is most likely also relevant in the case of keywords. Therefore, it will be necessary to carefully design a better approach of string splitting (and possibly also of word stemming, from which these coined terms might be excluded), in order to increase the chance of ending up with (only) truly relevant terms. A third option of fine-tuning could be the fact that in contrast to title and keywords certain (most likely very relevant) terms might be used more than once in the abstract of a document. Thus, term frequency could be considered as an additional factor in the process of generating the document-term-matrix of a sample based on abstract words.

From a researcher's point of view, the educated guess could be that keywords are the most suitable parameter for estimating research diversity, as they are intended to describe the topic of the article in a most concise way – and are sometimes even chosen based on a controlled vocabulary. In the case of the *Web of Science*, the "*Keywords Plus® are index terms created by Thomson Reuters from significant, frequently occurring words in the titles of an article's cited references*"[2]. So, interestingly, the experts at Thomson Reuters obviously also think that the references of an article are suitable for describing the content of the article itself. The observed trend in diversity based on keywords (which were a combination of author keywords and the aforementioned *Keywords Plus®*) does actually look quite reasonable, as it shows an increase in diversity in the ten years from 1991 to 2000 (at least in the case of Shannon diversity), which then levels off to a plateau. This could be a realistic development of a once quickly emerging field that has come to a point where the possibilities of the research community to deal with ever more different topics is saturated.

4.2 Randomization – or: was it really diversity that was measured?

After all these educated guesses and possible explanations and interpretations of the observed trends in diversity comes the even more exciting question, whether it really is diversity that was measured in all these experiments. Although some more detailed analyses will be needed, the three randomization scenarios that were performed concerning the reference-based diversity in the longest time series (1970-2009) already shed some light on that important question.

[2] http://images.isiknowledge.com/WOK46/help/WOS/h_fullrec.html#keywords_plus_fr (last visited on 25.04.2011)

As can be clearly seen from Fig. 3.10, the first and the second scenario were no adequate null models for the question how much variation in the measured diversity is due to statistical properties of the matrix only, because their diversity values were always lower than the ones derived from the original data. Diversity increases by definition when all individuals are randomly assigned to all species, or likewise when the references are randomly assigned to all documents, thus some things had to be going wrong that led to this unexpected result. As already mentioned in section 2.5, the first two randomization scenarios changed the citation frequencies of the different references. In the first scenario each reference had the same probability of being sampled in the randomization process, wherefore the sampling probabilities were adjusted to the observed citation frequencies in the second scenario. But nevertheless, the numbers of different references per article were consistently lower in both scenarios than in the original data. This was due to the fact that in both cases, it never happened that each of the different references from the list got sampled at least once in the random generation of the document-reference matrix. This, naturally, had to have a decreasing effect on diversity.

Thus, all hopes were on the third randomization alternative, in which both the length of the reference lists and the citation frequencies of the references were kept constant. In this case it was hypothesized that (1) the diversity values would increase after randomization and (2) the observed trend in diversity over the years would disappear. The first hypothesis was fully supported, as can be seen from Fig. 3.10. As the only difference between the original data and the randomization was that any non-random co-citation patterns were gone, anything else but an increase in diversity would have been impossible. The second hypothesis, however, proved to be wrong. The trend stayed the same and was even a bit more pronounced than in the original data and, moreover, the sample-based variation was much lower after randomization. This was indeed unexpected, but corroborated the apprehension already expressed by Mitesser (2008) that the observed trends in diversity were merely due to statistical properties of the document-reference matrix. So, obviously, it was not (only) diversity that was measured by this LSA-based method, but simply changes in the statistical properties of the data (length of the reference list and citation frequencies of the references) were responsible for changes in the measured “diversity”.

The next question to answer is, how much co-citation is already determined by the shape of the distribution of the number of references per article and of the citation frequencies of those references in the articles and how much scope for variation is left due to “intended”, thematic co-citation. In order to thoroughly answer this question, a quite laborious set of analyses needs to be carried out, which would have led too far within the scope of this thesis, but which is certainly worth wile to be performed in the future. A rudimentary glimpse on the matter was caught by taking 50 samples from year 2009 in the longest time series (i.e. based on a set of 14 journals) and calculating diversities of each sample both before and after randomization. The mean (±*sd*) Shannon diversity was 8.9580 (±0.0008) before and 8.9630 (±0.0001) after randomization, the mean of the differences between the data pairs (original/random) being 0.0050 (±0.0008). Although this is only a one-year sample, the values suggest that the difference between observed and randomized values is large enough and the sample-based variation in both cases small enough that variation due to thematic co-citation is detectable. So, maybe it would be possible to measure changes in the "real" diversity by always calculating diversity both from the original and the randomized data in a sample and then studying the changes in the difference between the two. As this is a very time-consuming task, the question remains, whether there is no better way of measuring research diversity than by LSA.

4.3 Extracting research topics by means of LSA

Irrespective of the presumable inadequateness of abstracting the eigenvalues of the topics derived by LSA to calculate the diversity of a sample of documents, LSA might do well for extracting "real" research topics from those documents. This has already been tested by Havemann *et al.* (2009) for articles from the field of scientometrics. A general problem with a LSA-based derivation of research topics might, however, be the fact that in the vast majority of cases, the LSA produces as many topics as there are documents in the set and thus the discriminatory power of the analysis is not very high. This can easily be seen from looking at Table 3.1 and Table 3.2. Even when only the 40 "largest" topics are considered, there are quite a few that sound very similar.

Keeping in mind that commonly about 100-200 papers have at least some share in a topic and there usually are less than 10-15 documents in which a topic has a larger than 5% share, it is easily comprehensible that there have to be topics that are quite similar to each other. Thus, although it is advantageous that not every document has to be assigned to one single topic only (in contrast to nature, where each individual can by definition only belong to one species), it is as well disadvantageous when the consequence is that there are as many topics as documents.

Moreover, research topics have – at the current state of the art – to be "manually" extracted from the titles of the articles in which the topic had the maximum share. This is a very time-consuming task and has to be accomplished by an expert in the studied research field, as it might often be difficult for an "outsider" to discover the common topic. Even for an expert this can be quite challenging at times, as can be seen from the rather many empty lines in Table 3.1 and Table 3.2. It is not hard to imagine that the number of empty lines would keep increasing when all of the 500 topics would be considered.

Based on the references, it is a common problem that articles are classified as similar because they co-cite one or more methods papers or (statistics) textbooks, but otherwise deal with completely different topics. In the term-based approaches, some previously discussed methodological imperfection in terms of data standardization (string splitting, stop word removal, word stemming) might lead to a similar problem when topics are derived based on title, abstract, or keywords.

4.4 Conclusion

In conclusion, the results of this study suggest that LSA in its current way of application is not a very suitable tool neither for measuring research diversity nor for deriving research topics from a sample of documents. Calculating diversity based on LSA might be possible in the context of a null model obtained by adequate randomization, which merits further scrutiny in the future. Still, it seems recommendable to focus on other options of measuring research diversity.

5 Summary

The aim of this master's thesis was to scrutinize the suitability of *Latent Semantic Analysis* – a rather new method in the field of scientometrics – for analyzing research diversity and for extracting latent themes from a set of documents. To this end, a large dataset was downloaded from the *ISI Web of Science*, all data belonging to the subject category *ecology*, which was chosen as an exemplary research field. As a compromise between analyzing as long a time series as possible and including documents from as many journals as possible, three different time series of 40, 30, and 20 years in length were used, which included articles from 14, 27, and 36 journals, respectively. All data were prepared and analyzed with the free statistics software *R*.

The core of the *Latent Semantic Analysis* is a *Singular Value Decomposition*, which decomposes the document-reference- or document-term-matrix into a set of eigenvectors and eigenvalues, of which the eigenvalues were used to calculate diversity and a combination of eigenvalues and eigenvectors to extract latent themes from two example bibliographies. Two diversity measures were used and compared: the *Shannon* and the *Simpson* diversity index, which are commonly used for calculating biodiversities of habitats in ecology.

In addition to the already tested calculation of research diversity based on co-citation of references, three other document properties – title, abstract, and keywords – were used and compared to the results of the reference-based approach. The results based on each of the document properties were consistent across the different time series, however, the comparison between the different parameters yielded quite different results, so that at this point it cannot be decided, which document property is the best one to use in order to reliably calculate research diversity.

Even more challenging was the question, whether the supposedly measured research diversity did really reflect diversity and not merely statistical properties of the document-reference- or document-term-matrix. The results of three different randomization scenarios (based on references) suggest that the major part of the observed trends in diversity were due to the shape of the distributions of the number of references per article and of the citation frequencies of the references in the bibliography. However, it might be possible to isolate the additional effect of thematic co-citation, if the diversity calculated from the original data is analyzed in the context of a suitable null model derived by adequate randomization.

Likewise, the suitability of *Latent Semantic Analysis* for extracting latent themes from a set of document seems rather limited, as the number of topics is commonly the same as the number of analyzed documents, which leads to a rather low discriminatory power and a small number of papers in which even the largest topics have a larger than 5% share. Moreover, the fact that currently the research topics have to be extracted manually from the titles of the papers in which a topic has the maximum share, makes this procedure very time-consuming and tedious.

In conclusion, the results of the analyses suggest that *Latent Semantic Analysis* is – in its current state – no very suitable tool for neither analyzing research diversity nor extracting latent themes from bibliographies. Therefore, other options of measuring research diversity should be evaluated in the future.

Bibliography

Adams, J. & Smith, D. (2003). Funding research diversity. *A report from Evidence Ltd to Universities UK* **1**(84036): 102.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**(6): 391–407, DOI: `http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9`.

Gläser, J., Lange, S., Laudel, G. & Schimank, U. (2008). Evaluationsbasierte Forschungsfinanzierung und ihre Folgen. In: Neidhardt, F., Mayntz, R., Weingart, P. & Wengenroth, U. (Ed.) *Wissen für Entscheidungsprozesse*, transcript, Bielefeld, pp. 145–170.

Gläser, J. & Laudel, G. (2007). Evaluation without evaluators: the impact of funding formulae on Australian university research. In: Whitley, R. & Gläser, J. (Ed.) *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, Springer, Dordrecht, pp. 127–151.

Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. *Scientometrics* **18**(3-4): 219–239.

Harley, S. & Lee, F. S. (1997). Research selectivity, managerialism, and the academic labor process: The future of nonmainstream economics in UK universities. *Human Relations* **50**(11): 1427–1460.

Havemann, F., Heinz, M., Mitesser, O. & Gläser, J. (2009). Calculating diversity of latent themes of research fields from their bibliographic coupling matrices, (unpublished).

Havemann, F., Heinz, M., Schmidt, M. & Gläser, J. (2007). Measuring diversity of research in bibliographic-coupling networks. In: *Proceedings of ISSI*, vol. 2, Madrid, vol. 2, pp. 860–861.

Hornik, K. (2009). *Snowball: Snowball Stemmers.* URL: `http://CRAN.R-project.org/package=Snowball`, R package version 0.0-7.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* **11**(1-2): 22–31.

Marshakova, I. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2– Informatsionnye Protsessy I Sistemy* **2**(6): 3–8.

Mitesser, O. (2008). *Latente Semantische Analyse zur Messung der Diversität von Forschungsgebieten.* Master's thesis, Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.

Mitesser, O., Heinz, M., Havemann, F. & Gläser, J. (2008). Measuring diversity of research by extracting latent themes from bipartite networks of papers and references. In: Kretschmer, H. & Havemann, F. (Ed.) *Proceedings of the Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*, Gesellschaft für Wissenschaftsforschung Berlin, Berlin, URL: `http://www.scientificcommons.org/36249138`.

Neff, M. W. & Corley, E. A. (2009). 35 years and 160,000 articles: a bibliometric exploration of the evolution of ecology. *Scientometrics* **80**(3): 657–682.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program* **14**(3): 130–137.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, URL: `http://www.R-project.org`.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, URL: `http://www.R-project.org/`.

Rivera, A. C. (2003). Trends in the evolution of ecology: “Spain is different”. *Web Ecology* **4**: 14–21.

Rodriguez, K. & Moreiro, J. A. (1996). The growth and development of research in the field of ecology. *Scientometrics* **35**(1): 59–70.

Schaefer, J., Opgen-Rhein, R., & Strimmer., K. (2010). *corpcor: Efficient Estimation of Covariance and (Partial) Correlation.* URL: `http://CRAN.R-project.org/package=corpcor`, R package version 1.5.7.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**: 379–423.

Simpson, E. (1949). Measurement of diversity. *Nature* **163**(4148): 688.

Small, A. (1978). Cited documents as concept symbols. *Social Studies of Science* **8**(3): 327–340.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface* **4**(15): 707–719.

Whitley, R. (2007). Evaluation without evaluators: the consequences of establishing research evaluation systems for knowledge production in different countries and scientific fields. In: Whitley, R. & Gläser, J. (Ed.) *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, Springer, Dordrecht, pp. 3–27.

Appendix

Listing 1: Example *R*-code for the LSA based on keywords in the shortest time series (1991-2009)

```
# Data sources from ISI Web of Science (column numbers are shown in parentheses):
# SO Publication Name (9)
# CR Cited References (26)
# AU Authors (2)
# PY Year Published (38)
# DT Document Type (12)
# NR Cited Reference Count (27)
# TI Document Title (8)
# JI ISO Source Abbreviation [Journal Identifier] (36)
# DE Author Keywords [Descriptors] (18)
# ID Keywords Plus (19)
# AB Abstract (20)
# UT Unique Article Identifier (51)

# load package for LSA
library(corpcor)
# load package for word stemming
library(Snowball)

# define local paths
st="~/Masterarbeit/"
source(paste(st,"files.R",sep=""))
# load variables containing journal sets
source(paste(st,"journals.R",sep=""))

# set number of files to be processed
SetNumber=length(datapaths)

# clear variables for data input (if not empty)
z=c()
zhlp=c()
stp=c()

# read stop word list
stp=read.table(paste(st,"Stopword-List.txt",sep=""),header=F,as.is=T,quote="\"")

# choose relevant columns from the text files
# other columns need not be retained in the working memory
```

```
fff=c(51,9,26,2,38,12,27,8,36,18,19,20)

# repeat for every file to be read:
for (k in  1:SetNumber){
    # return the file names
    print(datapaths[k])
    # read a single text file; special data format needed: with header, fields separated by "|"
    zhlp=read.table(paste(st,"DatenR/",datapaths[k],sep=""),sep="|",header=F,as.is=T,quote="\"")
    # bind data sets together, but only columns fff
    z=rbind(z,zhlp[,fff])
}
# add names to the columns of z
names(z)=colnames[fff]

# reduce the number of datasets to "articles", i.e. remove all documents,
# which are no articles and which have no keywords (DE) or keywords plus (ID)
z=z[z$DT=="Article" & (!is.na(z$DE) | !is.na(z$ID)),]
# print the dimensions of z
dim(z)
# remove possible duplicates (resulting from two separate, overlapping downloads)
z=z[which(!duplicated(z[,"UT"])),]
dim(z)
# choose the set of selected journals that were indexed over the whole timespan (1991-2009)
z=z[which(z$"SO" %in% jour89),]
# print the (new) dimensions of z
dim(z)

# set the sample size to 500 and the number of repetitions per year to 50
SampleN=500
SampleS=50

# initialize vectors for entropy and year data
# Evec: vector for Shannon diversity values from individual samples (caution: multiple samples per year)
# Svec: vector for Simpson diversity values from individual samples (caution: multiple samples per year)
# Yvec: vector for the years
# Rvec: vector for the mean number of keywords per article in a sample (multiple words count multiply)
# Uvec: vector for the mean number of different keywords per article in a sample (multiple words count only once)
# yVvec: vector for the extent of the vocabulary in each year (all articles)
# Vvec: vector for the extent of the vocabulary in each year (samples)
Svec=c(); Evec=c(); Yvec=c(); Rvec=c(); Uvec=c(); yVvec=c(); Vvec=c()

# analysis of vocabulary size in all articles per year
# initialize vector for the number of articles per year
ny=c()
# initialize vector for the number of unique title words per year
```

```r
un=c()

# set the time span to be analyzed
timespan=1991:2009

# analysis of the years
for (k in (timespan)){

    # choose all articles from year k and print their number
    z0=z[z$PY==k,]
    print(paste("nr. of art. in ",k,": ",length(z0$AU)))
    # write number of articles per year to a vector
    ny=c(ny,length(z0$AU))
    # initialize key word list for year k
    yKeyList=c()

    # repeat for each article in year k
    for (i in 1:length(z0$AU)){
        # keywords fields ($DE and $ID) are split at punctuation marks, spaces
        # and quotation marks (coded as "ANFZO")
        y0=strsplit(c(z0[i,]$DE,z0[i,]$ID),split="[[:punct:]]|[[:space:]]|ANFZO")
        # convert y0 to a vector
        y0=unlist(y0)
        # remove empty items from the vector
        y0=setdiff(y0,"")
        # convert all items to lower case letters
        y0=tolower(y0)
        # keywords are stemmed via the 'Porter-stemmer'
        y0=list(SnowballStemmer(y0))
        yKeyList=c(yKeyList,y0)
    }

    # convert dataframe to a vector with all keywords occurring in the dataset
    yTList=c(); for (i in 1:length(z0$AU)) yTList=c(yTList,yKeyList[[i]])

    # remove stop words
    yDiffList=setdiff(yTList,stp)

    # generate vector that contains all occurring keywords exactly once
    # DList: vector with duplicate-free keywords
    yDList=sort(unique(yDiffList))

    # determine the number of individual keywords
    yN0=length(yDList)
    # append the number of individual keywords to the vocabulary vector
    yVvec=c(yVvec,yN0)

    # repeat samples
    for (huz in (1:SampleS)){

        # take a sample of articles from the whole year
        z1=z0[sample(1:length(z0$AU),SampleN),]

        # determine the number of articles in the current year
        # this should yield SampleN
        n=length(z1$AU)

        # create dataframe with separate keyword lists
        # KeyList: dataframe for the different keyword strings from all articles
        # initialize vector for the number of keywords per article
        KeyList=c(); nkey=c()
        # repeat for all articles:
        for (i in 1:n) {
            # keywords fields ($DE and $ID) are split at punctuation marks, spaces
            # and quotation marks (coded as "ANFZO")
            y=strsplit(c(z1[i,]$DE,z1[i,]$ID),split="[[:punct:]]|[[:space:]]|ANFZO")
            # convert y0 to a vector
            y=unlist(y)
            # remove empty items from the vector
            y=setdiff(y,"")
            # count the key words
            nkey=c(nkey,length(y))
            # convert all items to lower case letters
            y=tolower(y)
            # keywords are stemmed via the ’Porter-stemmer’
            y=list(SnowballStemmer(y))
            KeyList=c(KeyList,y)
        }

        # convert dataframe to a vector with all keywords occurring in the dataset
        TList=c(); for (i in 1:n) TList=c(TList,KeyList[[i]])

        # remove stop words
        DiffList=setdiff(TList,stp)

        # generate keyword vector that contains all occurring keywords exactly once
        # DList: vector with duplicate-free keywords
        DList=sort(unique(DiffList))

        # determine the number of individual keywords
        N0=length(DList)
        # append the number of individual title words to the vocabulary vector
        Vvec=c(Vvec,N0)

        # create reference-document matrix; documents are still rows here!
        # m: matrix with article data
        m=matrix(rep(0,n*N0),nrow=n,ncol=N0)
        # write a "1" in a column when the respective keyword was contained in the respective document
        for (i in 1:n) m[i,which(DList %in% KeyList[[i]])]=1

        # remove 0 columns from the matrix (should actually not exist, but just in case)
        m=m[,apply(m,MARGIN=2,sum)!=0]
        # adjust the number of columns in the matrix
        N0=length(m[1,])

        # remove 0 rows from the matrix
        # in case the keywords of a document consisted only of stop words (though quite unlikely)
        m=m[apply(m,MARGIN=1,sum)!=0,]

        # adjust the number of rows in the matrix
        n=nrow(m)

        # normalizing the columns of the matrix to reduce the impact of highly cited papers
        for (j in 1:N0) m[ ,j] = m[ ,j]*log10(n/(sum(m[ ,j])))

        # normalizing the rows of the matrix to 1 to avoid that different numbers
        # of keywords per article influence the analysis
        for (i in 1:n) m[i, ] = m[i, ]/sqrt(sum((m[i, ])^2))

        # transponse the matrix, so that a document corresponds to a column
        m=t(m)

        # remove 0 rows from the matrix (just in case)
        # should actually not exist anymore
        m=m[apply(m,MARGIN=1,sum)!=0,]

        # core of the calculation: carry out the singular value decomposition
        g=fast.svd(m)

        # determine the sum of the singular values
        # SVSum: sum of singular values
        SVSum=sum(g$d^2)

        # determine the Shannon and Simpson diversity from the singular values
        # ent: Shannon diversity, smp: Simpson diversity
        ent= -sum((g$d^2/SVSum)*log2(g$d^2/SVSum))
        smp=1-sum((g$d^2/SVSum)^2)

        # maximum possible Shannon diversity (entmax)
        entmax=log2(length(g$d))

        # append current diversity values to the diversity vectors and the current year to the year vector
        Svec=c(Svec,smp); Evec=c(Evec,ent); Yvec=c(Yvec,k)
        # append numbers of keywords per article (Rvec) and numbers of different keywords per article
        # to the respective vectors
        Rvec=c(Rvec,length(which(m>0))/SampleN); Uvec=c(Uvec,length(m[,1])/SampleN)
    }
}

# initialize vectors for mean diversities, keyword numbers, and vocabulary sizes (sample mean!)
# and associated standard deviations
mSvec=c(); sSvec=c(); mEvec=c(); sEvec=c(); mRvec=c(); mUvec=c(); mVvec=c()

# determine mean and standard deviation for each year
for (j in 1:(length(timespan))) {
    mSvec=c(mSvec,mean(Svec[(((j-1)*SampleS)+1):(j*SampleS)]))
    sSvec=c(sSvec,sd(Svec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mEvec=c(mEvec,mean(Evec[(((j-1)*SampleS)+1):(j*SampleS)]))
    sEvec=c(sEvec,sd(Evec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mRvec=c(mRvec,mean(Rvec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mUvec=c(mUvec,mean(Uvec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mVvec=c(mVvec,mean(Vvec[(((j-1)*SampleS)+1):(j*SampleS)]))
}

# write resulting data to a text file for later use:
erg=data.frame(mEvec=mEvec,sEvec=sEvec,mRvec=mRvec,mUvec=mUvec,mSvec=mSvec,sSvec=sSvec,mVvec=mVvec,ny=ny,yVvec=yVvec)
write.table(erg, paste(st,"2011-03-23-SelectedJournals-1989-2009/",Sys.Date(),"-timeseriesKey.txt",sep=""))
write.table(data.frame(SampleN=SampleN,SampleS=SampleS,entmax=entmax,tmin=timespan[1],tmax=timespan[length(timespan)]), paste(st,"2011-03-23-SelectedJournals-1989-2009/",Sys.Date(),"-timeseriesKey.para.txt",sep=""))

# save workspace for later use:
save.image(paste(st,"2011-03-23-SelectedJournals-1989-2009/",Sys.Date(),"-LSAnormKey.RData",sep=""))
```

Listing 2: Example *R*-code for the third randomization scenario based on references in the longest time series (1970-2009)

```
# Data sources from ISI Web of Science (column numbers are shown in parentheses):
# SO Publication Name (9)
# CR Cited References (26)
# AU Authors (2)
# PY Year Published (38)
# DT Document Type (12)
# NR Cited Reference Count (27)
# TI Document Title (8)
# JI ISO Source Abbreviation [Journal Identifier] (36)
# DE Author Keywords [Descriptors] (18)
# AB Abstract (20)
# UT Unique Article Identifier (51)

# load package for LSA
library(corpcor)

# define local paths
st="~/Masterarbeit/"
source(paste(st,"files.R",sep=""))
# load variables containing journal sets
source(paste(st,"journals.R",sep=""))

# set number of files to be processed
SetNumber=length(datapaths)

# clear variables for data input (if not empty)
z=c()
zhlp=c()

# choose relevant columns from the text files
# other columns need not be retained in the working memory
fff=c(51,9,26,2,38,12,27,8,36,18,20)

# repeat for every file to be read:
for (k in  1:SetNumber){
    # return the file names
    print(datapaths[k])
    # read a single text file; special data format needed: with header, fields separated by "|"
    zhlp=read.table(paste(st,"DatenR/",datapaths[k],sep=""),sep="|",header=F,as.is=T,quote="\"")
    # bind data sets together, but only columns fff
    z=rbind(z,zhlp[,fff])
}
# add names to the columns of z
names(z)=colnames[fff]

# reduce the number of datasets to "articles", i.e. remove all documents,
# which are no articles and which have no references (NR)
z=z[z$DT=="Article" & z$NR>0,]
# print the dimensions of z
```

```
dim(z)
# remove possible duplicates (resulting from two separate, overlapping downloads)
z=z[which(!duplicated(z[,"UT"])),]
dim(z)
# choose the set of selected journals that were indexed over the whole timespan (1969-2009)
z=z[which(z$"SO" %in% jour69),]
# print the (new) dimensions of z
dim(z)

# set the sample size to 500 and the number of repetitions per year to 50
SampleN=500
SampleS=50

# initialize vectors for entropy and year data
# Evec: vector for Shannon diversity values from individual samples (caution: multiple samples per year)
# Svec: vector for Simpson diversity values from individual samples (caution: multiple samples per year)
# Yvec: vector for the years
# Rvec: vector for the mean number of references per article in a sample (multiple references count multiply)
# Uvec: vector for the mean number of different references per article in a sample (multiple references count only once)
# yVvec: vector for the number of different references in each year (all articles)
# Vvec: vector for the number of different references in each year (samples)
Svec=c(); Evec=c(); Yvec=c(); Rvec=c(); Uvec=c(); yVvec=c(); Vvec=c()

# analysis of reference size in all articles per year
# initialize vector for the number of articles per year
ny=c()
# initialize vector for the number of unique references per year
un=c()

# set the time span to be analyzed
timespan=1970:2009

# analysis of the years
for (k in (timespan)){

    # choose all articles from year k
    z0=z[z$PY==k,]
    print(paste("nr. of art. in ",k,": ",length(z0$AU)))

    # write number of articles per year to a vector
    ny=c(ny,length(z0$AU))
    # initialize references list for year k
    yRefList=c()

    # repeat for each article in year k
    for (i in 1:length(z0$AU)){
        # reference field ($CR) is split at semicola
        y0=strsplit(z0[i,]$CR,split="; ")
        # bind references together
        yRefList=c(yRefList,y0)
    }

    # convert dataframe to a vector with all references occurring in the dataset
    yTList=c(); for (i in 1:length(z0$AU)) yTList=c(yTList,yRefList[[i]])

    # generate vector that contains all occurring references exactly once
    # DList: vector with duplicate-free references
    yDList=sort(unique(yTList))

    # determine the number of individual references
    yN0=length(yDList)
    # append the number of individual references to the vocabulary vector
    yVvec=c(yVvec,yN0)

    # repeat samples
    for (huz in (1:SampleS)){

        # take a sample of articles from the whole year
        z1=z0[sample(1:length(z0$AU),SampleN),]

        # determine the number of articles in the current year
        # this should yield SampleN
        n=length(z1$AU)

        # create dataframe with separate reference lists
        # RefList: dataframe for the different reference strings from all articles
        # (duplicate references possible!)
        RefList=c(); nref=c()
        # repeat for all articles:
        for (i in 1:n) {
            # reference field ($CR) is split at semicola
            y=strsplit(z1[i,]$CR,split="; ")
            # bind references together
            RefList=c(RefList,y)
            # count the references
            nref=c(nref,length(y[[1]]))
        }

        # convert dataframe to a vector with all references occurring in the dataset
        TList=c(); for (i in 1:n) TList=c(TList,RefList[[i]])

        # generate reference vector that contains all occurring references exactly once
        # DList: vector with duplicate-free references
        DList=sort(unique(TList))

        # determine the number of individual references
        N0=length(DList)
        # append the number of individual title words to the vocabulary vector
        Vvec=c(Vvec,N0)

        # create reference-document matrix; documents are still rows here!
        # m: matrix with article data
        m=matrix(rep(0,n*N0),nrow=n,ncol=N0)
        # write a "1" in a column when the respective reference was cited in the respective document
        for (i in 1:n) m[i,which(DList %in% RefList[[i]] )]=1

        ## RANDOMIZATION ##

        # number of repetitions
        nx=100
        # repeat nx times
        for (k in 1:nx){
            # save old matrix
            m0=m

            # generate a vector with random row indices
            # successive row indices are interpreted as exchange pair
            rs=sample(n)
            # repeat for all row index pairs
            for (i in (2*(1:(n/2)))){

                # where are the "1"s in the first row of the row index pair?
                rs1=which(m[rs[i],]==1)
                    # second row
                rs2=which(m[rs[i-1],]==1)

                # which column from the first row of the row pair shall be exchanged?
                x1=sample(length(rs1),1)
                # which column from row 2?
                x2=sample(length(rs2),1)

                # if there's not already a "1" in the to-be-exchanged cells, exchange:
                if ((sum(rs1==rs2[x2])+sum(rs2==rs1[x1]))==0){
                    m[rs[i],]=0
                    m[rs[i],c(rs1[-x1],rs2[x2])]=1
                    m[rs[i-1],]=0
                    m[rs[i-1],c(rs2[-x2],rs1[x1])]=1
                }
            }
        }

        # remove 0 columns from the matrix (should actually not exist)
        m=m[,apply(m,MARGIN=2,sum)!=0]
        # adjust the number of columns in the matrix
        N0=length(m[1,])

```

```
    # normalizing the columns of the matrix to reduce the impact of highly cited papers
    for (j in 1:NO) m[ ,j] = m[ ,j]*log10(n/(sum(m[ ,j])))

    # normalizing the rows of the matrix to 1 to avoid that different numbers
    # of references per article influence the analysis
    for (i in 1:n) m[i, ] = m[i, ]/sqrt(sum((m[i, ])^2))

    # transponse the matrix, so that a document corresponds to a column
    m=t(m)

    # remove 0 rows from the matrix (should actually not exist)
    m=m[apply(m,MARGIN=1,sum)!=0,]

    # core of the calculation: carry out the singular value decomposition
    g=fast.svd(m)

    # determine the sum of the singular values
    # SVSum: sum of singular values
    SVSum=sum(g$d^2)

    # determine the Shannon and Simpson diversity from the singular values
    # ent: Shannon diversity, smp: Simpson diversity
    ent= -sum((g$d^2/SVSum)*log2(g$d^2/SVSum))
    smp=1-sum((g$d^2/SVSum)^2)

    # maximum possible Shannon diversity (entmax)
    entmax=log2(length(g$d))

    # append current diversity values to the diversity vectors and the current year to the year vector
    Svec=c(Svec,smp); Evec=c(Evec,ent); Yvec=c(Yvec,k)
    # append numbers of references per article (Rvec) and numbers of different references per article
    # to the respective vectors
    Rvec=c(Rvec,length(which(m>0))/SampleN); Uvec=c(Uvec,length(m[,1])/SampleN)
    }
  }

# initialize vectors for mean diversities, reference numbers, and reference sizes (sample mean!)
# and associated standard deviations
mSvec=c(); sSvec=c(); mEvec=c(); sEvec=c(); mRvec=c(); mUvec=c(); mVvec=c()

# determine mean and standard deviation for each year
for (j in 1:(length(timespan))) {
  mSvec=c(mSvec,mean(Svec[(((j-1)*SampleS)+1):(j*SampleS)]))
  sSvec=c(sSvec,sd(Svec[(((j-1)*SampleS)+1):(j*SampleS)]))
  mEvec=c(mEvec,mean(Evec[(((j-1)*SampleS)+1):(j*SampleS)]))
```

```
    sEvec=c(sEvec,sd(Evec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mRvec=c(mRvec,mean(Rvec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mUvec=c(mUvec,mean(Uvec[(((j-1)*SampleS)+1):(j*SampleS)]))
    mVvec=c(mVvec,mean(Vvec[(((j-1)*SampleS)+1):(j*SampleS)]))
}

# write resulting data to a text file for later use:
erg=data.frame(mEvec=mEvec,sEvec=sEvec,mRvec=mRvec,mUvec=mUvec,mSvec=mSvec,sSvec=sSvec,mVvec=mVvec,ny=ny,yVvec=yVvec)
write.table(erg, paste(st,"2011-04-20-SelectedJournals-1969-2009-rand-OM/",Sys.Date(),"-timeseriesRef.txt",sep=""))
write.table(data.frame(SampleN=SampleN,SampleS=SampleS,entmax=entmax,tmin=timespan[1],tmax=timespan[length(timespan)]), paste(st,"2011-04-20-SelectedJournals-1969-2009-rand-OM/",Sys.Date(),"-timeseriesRef.para.txt",sep=""))

# save workspace for later use:
save.image(paste(st,"2011-04-20-SelectedJournals-1969-2009-rand-OM/",Sys.Date(),"-LSAnormRef.RData",sep=""))
```

Acknowledgements

Special thanks go to Dr. Frank Havemann, Michael Heinz, and Dr. Oliver Mitesser, who offered me the chance to prepare this thesis and always supported me with new ideas and good advice.

Oliver and I once were colleagues in the field of ecology, while preparing our doctoral theses at the University of Würzburg. Now we both changed profession and became subject librarians, but we still share the fascination and interest in science. Thanks to you, Oliver, I could once again delve into the depths of science, both from an ecologist's and from an information scientist's point of view.