Topic Discovery from Textual Data

Shaheen Syed
This research was funded by the project SAF21, “Social Science Aspects of Fisheries for the 21st Century”. SAF21 is a project financed under the EU Horizon 2020 Marie Skłodowska-Curie (MSC) ITN – ETN program (project 642080).

© 2018 Shaheen Syed

*Topic Discovery from Textual Data*

Topic Discovery from Textual Data

Machine Learning and Natural Language Processing for Knowledge Discovery in the Fisheries Domain

Thema Ontdekking in Tekstuele Data

Machinaal Leren en Natuurlijke Taalverwerking voor Kennisontdekking in het Domein van de Visserij

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 20 maart 2019 des middags te 4.15 uur door

Shaheen Ali Shah Syed

geboren op 9 februari 1985 te Rotterdam
Promotor: Prof. dr. S. Brinkkemper
Copromotor: Dr. M. Spruit
Acknowledgments

This thesis is the result of my three-year PhD journey in which I had the pleasure to meet beautiful minds and souls. Throughout this journey, I also had the chance to work with and learn from some amazing people, for which I am truly grateful. My nearest and dearest have always supported me in the most loving and caring way, something that I truly cherish.

I want to thank my supervisors, Marco Spruit, Sjaak Brinkkemper, and Bruce Edmonds, for giving me the freedom to explore and work on my own ideas and interests. You have been a positive influence and working with all of you has been a real pleasure.

I want to thank my co-authors, Melania Borit, Charlotte Weber, and Lia ní Aodha, for putting up with my crazy way of working. I have learned a lot from each of you, and you have shown me interesting and alternative views of the world that have enriched me personally and professionally.

I want to thank Michaela Aschan for being a great mentor, for providing me with lots of opportunities, and for being an inspirational person. I want to thank Sjaak Brinkkemper for developing a master's program that prepared me for many of the academic challenges in this PhD journey. I want to thank Melania Borit for the nice lunches, for welcoming me, and for helping me during my many visits to Tromsø.

I want to show my gratitude to the EU for funding the project, to all the people involved in writing the SAF21 proposal, to the SAF21 members who I have met, to the UiT BRIDGE group for having me as a guest, to the University of Utrecht for welcoming me, to the Manchester Metropolitan University and the Centre for Policy Modelling for providing me with a work environment, and to all the other people I have had the pleasure to meet and talk to during my PhD.

I want to thank Charlotte Weber who has played many roles throughout this journey and surely will continue to do so in the future. Thank you for being such an amazing and caring person and thank you for showing me how to become a better version of myself. You truly are a unique soul, a blessing to the universe, and I am grateful to
have had the pleasure to meet you.

And last but not least, I want to thank my family for raising me, for making me the person I am today, and for supporting and loving me all these years.

Thank you all,

— Shaheen Syed
## Contents

1 Introduction 1

1.1 Knowledge Discovery Process 3

1.2 Topic Models 5

1.2.1 Latent Dirichlet Allocation 7

1.3 Research Domain 10

1.4 Research Questions 11

1.4.1 Main Research Question (MRQ) 11

1.4.2 Research Questions (RQ) 12

1.5 Research Methods 17

1.5.1 Computational Experiment 18

1.5.2 Quantitative Content Analysis 19

1.5.3 Social Network Analysis 20

1.6 Dissertation Outline 21

2 Full-Text or Abstract? 25

2.1 Introduction 26

2.2 Background 27

2.2.1 Latent Dirichlet Allocation 27

2.2.2 Topic Coherence Measurement 30

2.3 Methodology 32

2.3.1 The Experiment 32

2.3.2 Dataset 32

2.3.3 Creating LDA Models 34
## CONTENTS

2.3.4 Topic Coherence ........................................... 35
2.4 Results ...................................................... 35
  2.4.1 $DS_1$ Dataset ........................................... 39
  2.4.2 $DS_2$ Dataset ........................................... 39
  2.4.3 Human Topic Ranking .................................. 39
2.5 Discussion .................................................. 43
2.6 Conclusion .................................................. 44

3 Exploring Dirichlet Priors .................................. 47
  3.1 Introduction ............................................... 48
  3.2 Background ............................................... 49
    3.2.1 Latent Dirichlet Allocation ......................... 49
    3.2.2 Research Utilizing LDA ............................ 51
    3.2.3 Coherence Scores .................................. 52
  3.3 Methods .................................................. 54
    3.3.1 Dataset .............................................. 54
    3.3.2 Dirichlet Hyperparameters ....................... 55
    3.3.3 Creating LDA Models .............................. 56
    3.3.4 Topic Coherence .................................. 57
    3.3.5 Human Topic Ranking ............................. 57
    3.3.6 Relaxing LDA assumptions ....................... 58
  3.4 Results .................................................. 59
    3.4.1 Topic Coherence .................................. 59
    3.4.2 Human Topic Ranking ............................. 67
  3.5 Discussion and Conclusion ............................. 69

4 Bootstrapping a Semantic Lexicon ....................... 73
  4.1 Introduction ............................................ 74
  4.2 Previous Work .......................................... 75
  4.3 Lexicon Bootstrapping ................................ 76
    4.3.1 Domain and Seed Words ......................... 78
    4.3.2 Building the Corpus ............................ 78
    4.3.3 Chunking ........................................ 79
6.2.4 Pre-processing the Data Set ........................................ 136
6.2.5 Creating LDA Models ............................................. 136
6.2.6 Identifying Subtopics ............................................. 137
6.2.7 Labelling the Topics .............................................. 137
6.2.8 Calculating Sub-Topical Modelling Trends .................... 138

6.3 Results and Discussion ............................................... 138
6.3.1 General Modelling Topics ....................................... 138
6.3.2 Subtopics within Estimation Models ............................ 142
6.3.3 Subtopics within Stock Assessment Models .................... 147

6.4 Conclusions .......................................................... 149

Appendix ............................................................... 150

7 Global Network of Fisheries Science ................................. 155
7.1 Introduction .......................................................... 156
7.2 Results .............................................................. 157
7.2.1 Topology of the Co-Authorship Network ....................... 158
7.2.2 Country-Level Giants ............................................ 159
7.2.3 Institutional Dynamics .......................................... 159
7.2.4 Hidden Collaborative Groups .................................. 161
7.2.5 Country Clusters ................................................ 162
7.2.6 Communities of Authors and their Topical Foci .............. 162
7.3 Discussion ........................................................... 171
7.3.1 A Bourdieusian Perspective .................................... 171
7.3.2 Democratizing Fisheries Science? .............................. 172
7.3.3 Systems of Regionalization ..................................... 173
7.3.4 Collaboration Styles ............................................. 174
7.3.5 The Topical Landscape of Fisheries Science .................. 175
7.4 Limitations and Ways Forward ..................................... 176
7.5 Conclusion .......................................................... 177
7.6 Materials and Methods .............................................. 177
7.6.1 Data Collection .................................................. 178
7.6.2 Social Network Analysis ....................................... 178
7.6.3 Hidden Groups .................................................. 178
Your work is going to fill a large part of your life, and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. If you haven’t found it yet, keep looking. Don’t settle. As with all matters of the heart, you’ll know when you find it.

– Steve Jobs
Chapter 1

Introduction

It is estimated that the world’s data will grow to approximately 160 billion terabytes by 2025 (Reinsel et al., 2017), with most of that data occurring in an unstructured form—typically in the form of text. Approximately 20% of that data will be critical to the continuity of humans’ daily lives, and nearly 10% will be hypercritical with a direct and immediate impact on humans’ health and wellbeing (e.g., medical systems and telemetry). Today, we have reached the point where more data is being produced than can physically be stored (Hilbert and Lopez, 2011). In academia, research output is expected to grow at a rate of 8–9% annually, doubling roughly every nine years (Bornmann and Mutz, 2015). Across different fields and disciplines, problems, challenges, and opportunities arise to ingest all this data, create useful knowledge, and use the results for prediction and understanding (Blei, 2018; Science, 2011).

The surge of large volumes of data makes manual probing of the data slow, expensive, and subjective (Jiawei et al., 2012). New computational tools and algorithms are necessary to support the extraction of useful knowledge in a more structured way (Blei, 2012). Furthermore, to optimally and efficiently explore large volumes of data, it is essential to follow a well-established and accepted systematic process for comprehensive data analysis and knowledge extraction (Krochmal and Husi, 2018). The Knowledge Discovery in Databases (KDD) process is one such approach, which provides a systematic, iterative process to support the extraction of useful knowledge from large volumes of data (Fayyad et al., 1996). This process-oriented approach creates the context for developing and exploring the tools necessary to control the flood of data more systematically and helps to gain a better understanding of the data.

Extracting knowledge from data is particularly challenging for unstructured data, such as text in documents, where the data does not have a clear, semantically overt structure which a computer can easily understand (Manning et al., 2009). The field of natural language processing (NLP) is an area of research and application that examines how computers can be used to understand and manipulate natural language text (Chowd-
CHAPTER 1. INTRODUCTION

hury, 2003). NLP studies how humans interpret and understand natural language, and the field intends to develop appropriate tools and techniques to enable computers to assist in this task. The foundation of NLP is rooted in the fields of computer and information science, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, and psychology (Chowdhury, 2003). From the perspective of natural language understanding—in contrast to natural language generation—applications of NLP include, but are certainly not limited to, the understanding of sentiments and opinions within text (e.g., positive or negative reviews), exploring entities within text (e.g., persons, companies, locations), part-of-speech tagging (e.g., verbs, nouns, adjectives), and automatically uncovering the themes or topics from documents. It is the uncovering of topics that forms the subject of this thesis.

Typically, one would use keyword searches to find documents exhibiting certain themes or topics. However, a common problem with traditional keyword searches is that they have difficulties in detecting the underlying ideas, themes, or topics from documents (Blei and Lafferty, 2009; Srivastava and Sahami, 2009). This is especially true when the topics are hidden or latent, meaning that they are not explicitly mentioned in the document. For instance, a hypothetical document containing the words "blue", "red", and "green" might not explicitly mention the underlying topic of color. Here, a topic is thus a reference to a group of words that one would commonly use to describe something, and such words typically occur within the same linguistic context (DiMaggio et al., 2013). More formally, the group of words tends to co-occur, and this phenomenon is rooted in the distributional hypothesis; namely, words with similar meaning tend to occur in similar contexts (Harris, 1954).

Within the field of NLP, techniques from the field of machine learning—the science of getting computers to act without being explicitly programmed—are commonly applied. One such technique is probabilistic topic modeling, a type of unsupervised machine learning, which can automatically capture latent topics from large collections of documents (Blei, 2012; Blei and Lafferty, 2006, 2007, 2009; Blei et al., 2003; Rosen-Zvi et al., 2004). More concretely, they can discover groups of co-occurring words (the topics) from thousands or millions of documents without the need to manually annotate or label them. The uncovered latent topics can help users to explore, classify, and organize individual documents, as well as a document collection (called a corpus), and can additionally provide an extension to traditional search mechanisms.

The unsupervised nature of topic models makes them ideal candidates to employ on a variety of corpora and in numerous fields (Gatti et al., 2015; Hall et al., 2008; Sun and Yin, 2017; Wang and McCallum, 2006; Westgate et al., 2015). This is primarily since they: (i) require little human intervention and knowledge in the pre-analysis phase, (ii) generate reproducible results without human subjectivity bias, and (iii) can easily scale to thousands or millions of documents (Debortoli et al., 2016; Quinn et al., 2010). However, researchers usually treat topic models as black boxes, without thoroughly exploring their underlying assumptions and parameter values (Chen et al., 2016). This black box approach comes from the inherently complex statistical nature of topic mod-
els, and with improper use, their results may not sustain under scrutiny, may generate replication failure, may produce different outputs on multiple runs, and could be interpreted as questionable (Chuang et al., 2014, 2012; Grimmer and Stewart, 2013). In particular, the variety of topic model parameters and hyper-parameters are commonly set to their default values (Wallach et al., 2009), oftentimes caused by a lack of clear guidelines on how to optimally set them (Chen et al., 2016).

Motivated by the above factors, this thesis aims to provide answers regarding how to optimally and efficiently employ probabilistic topic models to large collections of documents. In this context, we have selected the research domain of fisheries science as our testbed (described further in Section 1.3), and we take fisheries scientific publications as our source of textual data. In particular, for fisheries sustainability, it is crucial that fisheries science appropriately considers the ecological, social, economic, and institutional elements (Boström, 2012; Dahl, 2012; Rindorf et al., 2017). To date, the social, economic, and institutional considerations have been largely neglected (Hicks et al., 2016; Levin et al., 2015), despite international agreements and legislation (Stephenson et al., 2018). Though illuminating, the assessment of considerations has been primarily performed by classical methods from the social sciences. Here, we aim to expand on such analyses by employing techniques from the field of NLP and probabilistic topics models to computationally analyze large collections of scientific data for the assessment of fisheries sustainability.

Our objective here is two-fold. First, this thesis scientifically investigates how to optimally and efficiently apply and interpret topic models to large collections of documents. Specifically, we study how different types of textual data, pre-processing steps, and hyper-parameter settings affect the quality of the derived latent topics. In doing so, we contribute to the methodological analysis and optimization of topic models, providing a starting point for researchers who want to apply topic models with scientific rigorousness to scientific publications. Second, by applying topic models to fisheries science publications, we study the domain through a new computational lens and expand on traditional approaches to assess fisheries sustainability. Moreover, we construct unique datasets comprising thousands of scientific articles and provide a quantitative assessment of fisheries science. For both objectives, we follow the systematic and well-grounded KDD process for comprehensive data analysis and knowledge extraction.

1.1 Knowledge Discovery Process

Within this thesis, our aim is to create new and useful knowledge by exploring latent topics derived from (raw) data. In this pursuit, we follow the well-established knowledge discovery process called Knowledge Discovery in Databases (KDD), originally proposed by Fayyad et al. (1996). Their seminal work provides an iterative systematic process to extract useful knowledge from large collections of data, which
Figure 1.1: The Knowledge Discovery in Databases (KDD) process.

takes the data as the starting point and through a sequence of steps ends with the derived knowledge (see Fig. 1.1). Although the process stems from more than two decades ago, the fundamental principles underlying the process still hold today. In fact, books devoted to analyzing data effectively, efficiently and in a scalable manner for the purpose of knowledge discovery incorporate the same, or very similar, iterative process (Jiawei et al., 2012). However, within the industry, the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000) is widely considered the number-one knowledge discovery process (Spruit and Lytras, 2018). CRISP-DM, although based on KDD (Wirth, 2000), is highly tuned towards the industry, where typically the evaluation phase addresses business needs, and the deployment phase covers customers’ needs. This makes CRISP-DM less suitable for academic purposes, and less suitable for our purpose.

The KDD process starts with obtaining data and selecting only data of interest (target data), on which discovery is to be performed. This is followed by a pre-processing step that might include, but not be limited to, removing noise and outliers from the data, normalizing, handling missing values and so on. The transformation step converts the data into appropriate features (i.e., variables describing the data) or uses dimensionality reduction techniques, such as principal component analysis, to reduce the number of features. In practice, the subsequent data mining step, oftentimes, if not always, dictates how the transformation into appropriate features must commence. The data mining step is the process where the intelligent data mining happens by applying algorithms, such as classification, regression and clustering methods, which result in the discovery of patterns, such as latent topics. Such patterns can then be interpreted and evaluated for their usefulness and correctness, and, if satisfactory, can lead to the creation of knowledge.

The KDD process has been used in many different areas and domains (Krochmal and Husi, 2018), and it has served as the basis for more domain-tailored knowledge discovery processes (Shearer, 2000). It is particularly useful as its aim is to place emphasis
on all steps when deriving knowledge from data, and not specifically on a single process, which is, oftentimes, the data mining process that entails optimizing algorithmic details. To reiterate the authors’ words “most previous work on KDD focused primarily on the data mining step. However, the other steps are equally, if not more, important for a successful application of KDD in practice” (Fayyad et al., 1996). Within the KDD process, the data mining step involves fitting (statistical) models to data to extract patterns and is shown as a single step within the overall sequence of steps. However, data mining can also refer to the entire KDD process, and such definitions are used interchangeably in the literature. In the latter case, the term “data mining” is treated in the broader sense of the words (Jiawei et al., 2012).

The reference to databases in KDD can look confusing, as it might imply that the data must originate from a database. Indeed, databases can hold and accumulate vast amounts of data and tapping into this (raw) data can yield interesting results. However, the original KDD paper defines data as “a set of facts” (Fayyad et al., 1996), which might be stored—and typically is—within a database, but may very well be data originating from elsewhere, such as documents on a disk or content from the web. In this sense, and given that data today can come from a multitude of sources, we interpret the KDD process more as the process of Knowledge Discovery from Data.

### 1.2 Topic Models

As previously mentioned, this thesis explores the research directive of automatically uncovering topics from a corpus in an attempt to improve and better understand the knowledge discovery process. In this pursuit, we employ topic modeling algorithms to automatically uncover topics from documents, and our focus here is on the most popular and highly studied topic model called Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

LDA overcomes the limitations of older topic models, such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999). At the time of writing, the original LDA method proposed by Blei et al. (2003) has over 23,000 citations. The technique has received much attention from machine learning researchers and other scholars, and has been adopted and extended in many ways (Blei and Lafferty, 2006, 2007; Chang and Blei, 2010; Doyle and Elkan, 2009; Reisinger et al., 2010; Rosen-Zvi et al., 2004; Wallach, 2006b; Wang and Blei, 2009; Whye Teh et al., 2004). To provide a sense of the popularity of topic models within the scientific literature, Fig. 1.2 shows how frequently LDA and the two older topic models, LSI (synonymously referred to as Latent Semantic Analysis or LSA) and pLSI (synonymously referred to as probabilistic Latent Semantic Analysis or pLSA), occur within the title, abstract or keywords of publications covered by Scopus in the period 2000 to 2017. Frequencies are obtained by searching for the full name of the topic model enclosed by quotation marks, e.g. “latent Dirichlet allocation” (to disam-
One of the main advantages of LDA is that it produces a set of topics that describe the entire corpus and are individually interpretable, while simultaneously describing how each document exhibits those topics proportionally (Blei and Lafferty, 2009; Griffiths et al., 2007). For example, we can infer that a corpus of 1,000 fisheries science publications contains four topics, and those four topics can be interpreted as topics concerning “fisheries management”, “gear”, “growth”, and “vessels”. Additionally, we can infer how these topics are proportionally present within each document. For example, one document might contain 60% on the topic of “fisheries management”, and 40% on the topic of “vessels”. Technically, all topics are present within each document, but only a few topics make up the most substantial part of the document, with the remaining topics very close to 0%. LDA thus follows the natural assumption that documents exhibit multiple topics in mixed proportions (Blei and Lafferty, 2009), and is able to capture the heterogeneity of topics within documents (Erosheva et al., 2004). In contrast to other topic models (e.g., pLSI), LDA can handle large amounts of data as the model parameters do not grow linearly with the number of documents (Blei et al., 2003; Chen et al., 2016), which prevents overfitting of the model. LDA has also been shown to outperform traditional cluster-based methods for information retrieval purposes (Wei and Croft, 2006) and baseline clustering methods (e.g., K-means) for scientometric research (Yau et al., 2014). More research into the benefits of LDA and the differences between other topic model algorithms is described in Anaya (2011) and Steyvers and Griffiths (2007).

To date, researchers apply LDA on a variety of corpora and in numerous domains, and many toolkits are available that more easily enable this: Mallet (McCallum, 2002), Gensim (Rehurek and Sojka, 2010) and Stanford TMT (Ramage and Rosen, 2009). Knowing how to employ LDA optimally fosters a better understanding of the underlying
implications of the model parameters, pre-processing steps and selected data, and can help future research yield better and scientifically valid results. For these reasons, this thesis takes LDA as the primary technique to uncover latent topics from a set of documents.

1.2.1 Latent Dirichlet Allocation

We will succinctly describe LDA in non-technical language solely for the purpose of understanding the remaining part of this chapter. We do so by presenting the analytical steps involved, the adopted terminology, and the underlying statistical assumptions. Fig. 1.3 displays a graphical diagram of the LDA process where we assume that there are D documents, V words, and K topics. The number of documents D, and all the (distinct) words V within those documents, are defined by the corpus under study. The number of topics, K, is assumed to be known a priori. A formal and mathematical description is given in Chapters 2 and 3, and a non-technical description with concrete examples is additionally given in Chapter 5. We further refer the interested reader to Blei (2012) for an introductory explanation.

Each document includes some words, and the collection of all the words is considered the vocabulary of the corpus (hence the V). In many areas of NLP, documents are represented as bag-of-words (BOW) features, where the words and their frequency within a document are treated as individual document features. A BOW representation of a single document is merely a row (i.e., vector) within a table (i.e., matrix) where each column represents a word (i.e., term or feature) from the vocabulary, and the cells represent the frequency or count of that word within the document. Fig. 1.3 depicts a BOW representation for three hypothetical documents and three hypothetical words. Document d1 thus contains word1 one time, word2 zero times, word3 three times, and so on. The process of converting documents into BOW features is performed in the
CHAPTER 1. INTRODUCTION

pre-processing phase of LDA, and if desired, words with similar meaning (e.g., plural and singular words) can be grouped (stemming or lemmatization), words that have no specific meaning can be filtered out (removing stop words such as “the” and “an”), and high- and low-frequency words can be eliminated (pruning).

Creating the LDA model comprises two steps (see Fig. 1.3): (1) generating documents based on statistical sampling rules, called the generative process; and (2), inferring model parameters from these generated documents, called the inference process. The first step, the generative process, involves the following steps:

1. For each topic (of K topics):
   (a) Sample a distribution over V words (from a Dirichlet distribution).

2. For every document (of D documents):
   (a) Sample a distribution over K topics (from a Dirichlet distribution).
   (b) For every word within this document (document length can be fixed or sampled from an appropriate distribution, e.g., Poisson):
      i. Sample a single topic from the previously sampled topic distribution.
      ii. Sample a word from this topic.

The Dirichlet distribution can be viewed as a “distribution of distributions”, where every point on the distribution (called the simplex, see Fig. 1.4) represents some probability distribution over K discrete classes (where K≤2). The shape of the Dirichlet distribution is defined by a parameter value for each of the K classes. Technically, this is a vector of length K and consists of non-negative values. Fig. 1.4 shows two Dirichlet distributions with three classes (K=3) and corresponding parameter values (labeled \(a_1\), \(a_2\), and \(a_3\)) for each of the three classes. When sampling from the Dirichlet distribution, we obtain some probability mass (between 0 and 1) for each of the K classes, with the restriction that the sum of the probability masses for all classes equals 1. Furthermore, it is important to note that within LDA there are two Dirichlet distributions (for topics in documents (step 1a), and words within topics (step 2a)) from which we sample, and these distributions are considered prior distributions. The chosen parameters of the two Dirichlet distributions affect the smoothing or shape of the distribution (see Fig. 1.4) and the drawn samples will be different with different parameterizations, which in turn affects the output of LDA.

Within the first part, LDA produces a set of generated documents governed by the generative process, the prior Dirichlet distribution, and the statistical sampling process. The second part, the inference process, is directed at uncovering the parameters that most likely would have resulted in those generated documents. This process can be viewed as reverse engineering the generative process. Technically, we want to infer the
hidden variables given the observed documents, also referred to as inferring the posterior probability of the hidden variables given the data. The hidden variables are (i) the word probabilities within each topic, (ii) the topic probabilities within each document, and (iii) the mapping of each word to a topic, called topic assignment. Calculating the posterior probability is intractable and several inference methods exist (Blei and Jordan, 2006; Newman et al., 2007; Porteous et al., 2008; Teh et al., 2006; Wang et al., 2011). Although the discussion of posterior inference techniques is beyond the scope of this thesis, after inference has converged, we obtain topics (represented as topic-word distributions) and how these topics are proportionally present within each of document (represented as document-topic distributions).

In a way, LDA can be seen as a dimensionality reduction technique, as documents are now described as distributions over K topics, in comparison to the original bag-of-words representation of length V. The document-topic distribution provides insights into the topical decomposition of the entire corpus. Additionally, documents with similar topic distributions, measured by, for instance, the Kullback-Leibler divergence (Kullback and Leibler, 1951), or the Hellinger distance (Hellinger, 1909) can be viewed as comparable in the topical content they address. Documents that are comparable can be used for searching (retrieval) or clustering (classification). Until now, we have presented the topics as distributions over V words. To get insights into the semantic meaning of the topic distribution, typically one can sort the probabilities of words within a particular topic in descending order. The top words (usually the top 10) are most informative about the latent meaning of the topic, and typically a domain expert assigns a label to each topic that best captures the semantics of the top words.

An example of two topic-word distributions with the top 10 high probability words is shown in Fig. 1.5. The two latent topics are uncovered from fisheries publica-
1.3 Research Domain

This thesis operates within the domain of fisheries. A fishery can be defined as “the complex of people, their institutions, their harvest and their observations associated with and including a targeted stock or group of stocks (i.e., usually fish), and increasingly, the associated ecosystems that produce said stocks” (Link, 2010). Simply put, fisheries deal with all aspects of harvesting fish, including the people, methods, tools, management, boats, ecosystem, and fish, and how they are interlinked.

The primary reason for the focus of this research domain is the overarching project in which this thesis is embedded: the EU Horizon 2020, Marie Skłodowska-Curie, European Training Network called “Social Science Aspects of Fisheries for the 21st Century”, or SAF21 for short. The project aims to improve the management (i.e., social) understanding of fisheries, where fisheries are recognized to be part of social-ecological systems (Ostrom, 2009); systems where the social (e.g., human interest and actions) and ecological (e.g., fish and ecosystem) dimensions are connected and interlinked.

For fisheries sustainability, it becomes increasingly important to appropriately consider the four elements inherent within social-ecological systems through equal and balanced
CHAPTER 1. INTRODUCTION

These four elements are the ecological, social, economic, and institutional considerations (Boström, 2012; Dahl, 2012; Rindorf et al., 2017). To date, it is argued that typically the ecological considerations are primarily addressed, and as a result, the social, economic, and institutional considerations are relatively neglected (Hicks et al., 2016; Levin et al., 2015).

We additionally strive to provide insights into the prevalence of these dimensions within fisheries, and specifically within fisheries science publications. Fisheries science publications are thus taken as the source of data, i.e., documents or corpus as input for LDA, within this thesis. Additionally, fisheries domain experts are consulted to interpret and evaluate latent topics derived from this data.

1.4 Research Questions

1.4.1 Main Research Question (MRQ)

A large body of scientific work aimed at uncovering latent topical structures from textual data has focused on the theoretical aspects of the topic model’s algorithmic complexity. Such aspects include, but are not limited to, model run time, model complexity, model inference, and other forms of optimizing algorithmic details. Although these are highly relevant and important endeavors, little research has taken the solution-oriented, applied data science approach as its vantage point (Spruit and Lytras, 2018); that is, how to optimally and efficiently apply topic models to large collections of documents and interpret them to gain a better understanding of the underlying topical content.

Indeed, much research has applied topic models in a variety of contexts and research domains, such as transportation research (Sun and Yin, 2017), computer science (Hall et al., 2008; Wang et al., 2011; Wang and McCallum, 2006), conservation science (Westgate et al., 2015) and the fields of operations research and management science (Gatti et al., 2015). Although each uncovers interesting patterns and relevant knowledge from documents, they often lack foundational theory for data selection and model (hyper-) parameterization (Chen et al., 2016), which form the basis of any derived output. Especially when the underlying data represents knowledge itself—as is the case with scientific literature—and when it forms the basis of a new knowledge discovery process, making theoretically justified choices to optimally utilize the potential of topic models becomes extremely important. Furthermore, within an applied data science context, and from a domain-specific perspective, there is space to not only interpret the latent topics in isolation, but to interpret them in a broader, domain-related context. We hypothesize that data selection and model parameterization significantly affect the derived underlying latent topical structure from a set of documents, and, subsequently, affect any derived knowledge in this regard. We further hypothesize that we
can create new (fisheries) domain knowledge from the interpretation of latent topics and explore them beyond the raw topic-word and document-topic distributions. Thus, this dissertation poses the following main research question (MRQ):

**MRQ — How can we improve the knowledge discovery process from textual data through latent topical perspectives?**

In this thesis, we take the entire KDD process, from raw data to useful knowledge, as a blueprint for an appropriate knowledge discovery process. The ultimate objective of a KDD process is to turn raw data, with implicitly captured knowledge, into patterns that explicitly expose this knowledge. Each step in a KDD process is considered equally important and together they make up the knowledge discovery process (see Section 1.1). Since our main research question is to improve the knowledge discovery process of topical perspectives from textual data, it is imperative that the intricacies of each KDD step are dealt with appropriately, that is, with sufficient scientific rigorousness.

### 1.4.2 Research Questions (RQ)

To find answers and support our main research question, we formulated six research questions (RQ1–6) with the aim to improve and better understand the knowledge discovery process. We provide a list of the six research questions below, and each is treated in more detail in subsequent sections of this chapter:

1. What types of textual data result in high-quality latent topics?
2. How does the hyper-parameterization of a topic model algorithm affect the quality of latent topics?
3. Can we assess the quality of latent topics using a semi-automatically constructed list of semantically related words?
4. How can we construct knowledge from latent topics derived from large collections of documents?
5. How can we construct knowledge from latent sub-topics derived from large collections of documents?
6. How can we utilize knowledge derived from latent topics for a subsequent knowledge discovery process?

The mapping of the six research questions to existing KDD steps is shown in Fig. 1.6. The six research questions can be divided into two parts. The first consists of RQ1, RQ2, and RQ3, to scientifically study the effects several KDD steps have on the derived patterns (the latent topics). RQ1 studies the KDD data selection and related pre-processing
steps, RQ2 studies the settings of hyper-parameters within the KDD data mining step, and RQ3 studies the evaluation of the patterns within the interpretation and evaluation step. Thus, RQ1–RQ3 aim to optimize the choices made in a number of KDD steps to ensure that such choices are based on proper scientific rationale, rather than relying, for instance, on default choices (Spruit and Lytras, 2018; Wallach et al., 2009). The second part consists of RQ4, RQ5, and RQ6, which aim to improve the understanding of the derived patterns and the translation into useful knowledge. Below, each RQ is discussed in more detail.

**RQ1 — What types of textual data result in high-quality latent topics?**

The first step of the KDD process starts with the data, and the selection of appropriate or relevant data thereof. A known phrase amongst computer scientists is the phrase “garbage in, garbage out”. It means that the quality of data you feed into a system affects the quality of the results you obtain from that system. As a first step within the KDD process, the data selection process can be extremely important as it could affect any subsequent step or process. This thesis operates within the domain of fisheries and uses scientific literature, in the form of scientific peer-reviewed articles, as the basis for data. Logically, that leaves the choice for two types of variants of this data: the abstracts of the scientific articles, and the full-text of the articles. Although any combination of article meta-data can be viewed as a type representation, from an article data retrieval perspective, abstracts and full-text data are usually treated as two types. Abstracts are easily accessible, or at least more easily, than obtaining full-text data for scientific articles. When employing probabilistic topic models, many studies use abstract data as the source of textual data (Gatti et al., 2015; Grimmer and Stewart, 2013; Sun and Yin,
2017; Westgate et al., 2015), and a smaller number of studies use full-text data (Alston and Pardey, 2016; Hall et al., 2008; Wang and McCallum, 2006). It can be argued that the abstracts should contain most parts of the research underpinning the article (or at least the essential parts), and, as such, should be sufficient to distill the main topical content of the article (Lin, 2009).

This research question aims to (i) explore the effects of using abstracts and full-text documents on the quality of derived latent topics, and (ii) explore what role related pre-processing steps have within this process. Until now, no study has examined the effects of using abstracts and full-text data when uncovering latent topics. As a result, researchers using abstracts or full-text publications oftentimes do so without any scientific rational, and the choice for either is typically not argued for. There are various reasons why this might be the case: one might only have access to abstract data (i.e. availability), one may want to keep the computational time to a minimum (i.e. feasibility), or one may want to reduce the pre-processing steps that are often necessary when dealing with full-text articles (i.e. simplicity). These pre-processing steps could include scraping the publishers’ repositories; converting PDF to plain-text, either directly or with the aid of optical character recognition (OCR) software; or an increased boilerplate cleaning phase. Since the data is a crucial determinant within any knowledge discovery process, and since our main aim is to improve any derived topical content from it, we study the empirical effects abstracts and full-text have on the derived latent topics. Fig. 1.6 highlights where RQ1 takes place, and spans, mainly, the data and data selection phase, while also considering aspects of the pre-processing phase.

RQ2 — How does the hyper-parameterization of a topic model algorithm affect the quality of latent topics?

Many topic model algorithms allow for tuning and tweaking of a variety of hyperparameters that will affect the model’s output. This is certainly the case for LDA. Since LDA is a Bayesian probabilistic topic model, one of its advantages is that it incorporates prior knowledge into the model before topic discovery starts (see section 1.2.1). Although Bayesian statistics is often heralded for its ability to use prior knowledge in the statistical equation, it can simultaneously be seen as a disadvantage, as selecting and choosing an appropriate prior can be a difficult task in itself. For LDA, this prior knowledge is encoded in the two Dirichlet priors on (i) the distribution of topics within documents, and (ii) the distribution of words within topics. Effectively, such priors can be set to be symmetrical or asymmetrical, meaning that (i) the probability distribution of topics is evenly (symmetrically) or unevenly (asymmetrically) distributed within a document, and (ii) the probability distributions of words within a topic is evenly (symmetrically) or unevenly (asymmetrically) distributed. One might surmise that, for instance, the topics within a corpus are not all equally present, such as is the case when one topic is more popular or gains more attention than another. In this case, a symmetrical prior distribution, thus an evenly distributed prior for topics in documents, can be a false assumption. The symmetrical Dirichlet prior, which is almost always the
default choice for researchers to use when using LDA (Wallach et al., 2009), can thus, possibly, lead to incorrect or less optimal latent topics.

Little research has been performed on the effects of Dirichlet priors on the discovery of latent topics; a single study explores these effects for documents related to news content and patents (Wallach, 2006b). However, studies exploring the effects of priors on scientific articles have not been performed. RQ2 thus aims to (i) explore the empirical effects Dirichlet priors have on the quality of the uncovered latent topics, and (ii) simultaneously consider the effects on different parts of the scientific article (abstract and full-text). Fig. 1.6 shows where RQ2 fits into the overall KDD process; mainly in the data mining phase. Furthermore, incorporating appropriate prior knowledge within the KDD process—here through a statistical Bayesian prior—“ensures that useful knowledge is derived from the data” (Fayyad et al., 1996). With the main research question targeted at improving the understanding of latent topics, RQ2 is an important question to help achieve this goal.

**RQ3 — Can we assess the quality of latent topics using a semi-automatically constructed list of semantically related words?**

LDA outputs a single topic as a discrete probability distribution over all the words in the vocabulary. Thus, each word within a topic has a certain probability assigned to it (this is true for each topic), and when sorted by probability value, the words with the highest probabilities reveal the semantic meaning of that topic (see Section 1.2.1). The use of coherence scores is a typical approach to quantify the quality of the words within topics (Aletras and Stevenson, 2013; Newman et al., 2010b; Röder et al., 2015; Stevens et al., 2012), and is also the approach adopted in this thesis. To assess the quality of topics, the gold standard approach is human topic ranking (Lau et al., 2011). However, the human ranking approach is, oftentimes, too expensive or time consuming. Coherence scores have been shown to correlate well with human ranking data (Chang et al., 2009), and can thus be considered an appropriate and adequate approach when human ranking is not feasible.

In addition to coherence scores, RQ3 investigates whether a list of semantically related words, called “semantic lexicons”, can be obtained to assess the quality of the words that constitute a latent topic (Igo and Riloff, 2009; Qadir et al., 2015; Thelen and Riloff, 2002; Ziering et al., 2013b). Ideally, such lists should be constructed from a separate, unrelated collection of textual data, such as content from the web. The semi-automatic process of creating a semantic lexicon is based on a bootstrapping process, which starts with an initial set of words (called “seed words”) and follows with an automatic process to find semantic related words. The semi-automatic bootstrapping process is necessary to steer the process into a specific context or topic, as determined by the choice of seed words. In doing so, we aim to investigate whether semantic lexicons, which consist of semantically related words, can be used to assess the quality of latent topics. RQ3 thus aims to assess the quality of latent topics with a reference list of words, which can then help in “finding understandable patterns that can be interpreted as useful or interesting knowledge” (Fayyad et al., 1996). With respect to the KDD process, RQ3
can be mapped onto the evaluation phase, as shown in Fig. 1.6.

Up to this point, RQ1, RQ2, and RQ3 all aim to improve the quality of the derived knowledge output intrinsic within the KDD process. The subsequent three research questions are aimed at the interpretation of the derived patterns, the latent topics, and how such patterns can be turned into usable and, within the context of this thesis, also re-usable knowledge. We take the lessons learned from RQ1, RQ2 and, to a lesser extent, RQ3 as groundwork for a large-scale topic model analysis within the domain of fisheries science through answering RQ4–6.

**RQ4 — How can we construct knowledge from latent topics derived from large collections of documents?**

The primary result of topic model analysis is a set of topics and, for each topic, a list of words with associated probabilities. This new representation, where documents are now represented as distributions over topics, forms the basis for the evaluation step and, ultimately, the knowledge construction step within the KDD process (Fig. 1.6). This research question aims to interpret the latent topics, which are derived from the data mining phase, and aims to extract useful, relevant and understandable knowledge from them. Several studies have already explored the latent topical contents from a collection of scientific publications, and we certainly do not claim to be pioneers in this respect. However, we are the first to perform such (large-scale) analysis within the domain of fisheries, and the first to use scientific rationale for the steps leading up to the evaluation phase. Also, we aim to go beyond exploring the latent structures of the documents and interpret the results in a broader context. That is, we aim to construct new knowledge and shed light on the ecological, social, economic, and institutional considerations within fisheries for increased fisheries sustainability (Boström, 2012; Dahl, 2012; Rindorf et al., 2017; Stephenson et al., 2018). Thus, RQ4 is aimed at the interpretation of latent topics for the construction of new knowledge, which is the final step in the KDD process (Fig. 1.6).

**RQ5 — How can we construct knowledge from latent sub-topics derived from large collections of documents?**

Within the KDD process, there is a possibility to utilize the results or outcomes of one step, and to return to a previous step to exploit this result. This process is indicated by the dashed line in Fig. 1.6 (below each consecutive step). RQ5 uses this possibility in the penultimate phase (evaluation and interpretation), which causes the output of that phase to be the input of a new KDD process. More concretely, we explore how we can utilize the latent topics of one KDD process to then re-run the KDD process with a subset of the original document collection. LDA lends itself for such an analysis, as the latent topics can be used to select a subset of the original set of documents (which then only address a single topic, for instance). In doing so, we extract sub-topics from a main topic. The distinction between a main topic and a sub-topic can be made more explicit with an example. Imagine a collection of news articles. When performing a topic model analysis, we might find (latent) topics related to sports, politics, and the
When isolating the articles covering (mainly) the sports topic, we can then take the collection of sports articles as our starting point. A new topic model analysis on that subset of articles could then reveal specific types of sports, such as football and basketball. Within the context of fisheries science, a main topic can be, for instance, fisheries management, with subtopics covering aspects of the management process, such as the management of recreational fisheries.

RQ5, similar to RQ4, aims to construct new knowledge, but now on a sub-topical level. It provides a knowledge discovery process on a lower level (the sub-topics), in contrast to the higher level (the main topics) analysis performed in RQ4.

**RQ6 — How can we utilize knowledge derived from latent topics for a subsequent knowledge discovery process?**

The traditional KDD process has knowledge itself as its final output. Though it looks like this is the final step, this knowledge may very well be the start (or any other step) of a new (related or unrelated) KDD process. Extended KDD models capture this step more explicitly (Quick and Choo, 2014). RQ4 and RQ5 explore the latent topics extracted from fisheries science publications with the aim to explore and better understand the corpus, and shed light on particular domain-specific challenges, such as fisheries sustainability. Such analysis reveals interesting results, but we aim to combine these results with an additional analysis. RQ6 explores how we can use the discovery of latent topics (and the knowledge it contains) to enrich a new knowledge discovery process. Within the domain of fisheries, and within the context of the scientific articles, we aim to use latent topics to gain a better understanding of the knowledge production of fisheries science. More concretely, we utilize techniques from the field of social network analysis, particularly community detection algorithms, to explore the spatial, temporal and, by enriching it with a topic model analysis, topical foci. In doing so, RQ6 aims to explore whether new knowledge can be constructed by combining topic model analysis with a new knowledge discovery process.

### 1.5 Research Methods

Within this thesis, different research methods are used to formulate answers to the research questions described in Section 1.4. We will briefly describe the three major ones: computational experiments (used for RQ1–3), content analysis (used for RQ4–6), and social network analysis (used for RQ6).
1.5.1 Computational Experiment

A computational experiment concerns itself with the theoretical analysis and empirical testing of a computational method, such as approximation or optimization algorithms. The experiment is a set of tests which runs under controlled definitions for a specific purpose (Barr et al., 1995). Such a purpose can include a demonstration of known truths, validating hypotheses, and examining the performance of something new. Typically, the effects and influences of controllable variables (factors) within an experiment are measured and studied on some phenomena. Within computational experiments, the studied phenomena oftentimes entail the algorithm’s performance (examples described later). When new computational methods are presented, their contributions should be evaluated scientifically and reported on in an objective manner, which is not always done (Barr et al., 1995). A proper experiment includes the following seven steps (Montgomery, 2012):

1. Recognition of and statement of the problem
2. Selection of the response variables
3. Choice of factors, levels, and ranges
4. Choice of experimental design
5. Performing the experiment
6. Statistical analysis of the data
7. Conclusions and recommendations

**Recognition of and statement of the problem:** The first step defines the purpose of the experiment and should be determined before performing the actual experiment. The objective is a statement of the questions to be answered and the reasons the experiment should be conducted. For computational experiments, two types can be discerned: (i) comparing different algorithms addressing the same class of problems, or (ii) characterizing the algorithm in isolation. The latter type is used within RQ1 and RQ2, with the objective to gain an understanding of the behavior of the algorithm (LDA) and the factors that influence its behavior (data, pre-processing, and priors). In both cases, the experiment is performed by singling out factors while fixing the remaining code, which is a prerequisite of a well-designed experiment (Barr et al., 1995). RQ3 fits into the former type, exploring different algorithms with the aim of constructing a high-quality semantic lexicon.

**Selection of the response variables:** The second step defines how to measure the performance of the algorithm under study. Here, performance is the response variable and can be determined or characterized in a number of ways. Typically, performance addresses the quality of the output, the time it takes to find a good or optimal solution, the robustness of the algorithm, or the tradeoff between feasibility and quality. Within
RQ1 and RQ2, the experiments are designed to compare the quality of the output, which are the latent topics derived from the data (documents) under study, and for RQ3, the quality of the semantic lexicon. Additionally, robustness is addressed by experimenting with different datasets.

Choice of factors, levels, and ranges: The third step—often performed simultaneously with the second step—deals with the selection of factors that affect the performance of the algorithm within the computational experiment. Such factors can take on many forms, but within the context of RQ1–3, the factors include data types (abstract and full-text), data sources (journals and web), parametric distributions (Dirichlet priors), model parameters (number of topics) and, to a lesser extent, data representation as a result of different pre-processing steps.

Choice of experimental design and execution: The fourth and fifth steps design and execute the experiment to ensure that appropriate test results can be collected and are suitable for statistical analysis (Mason et al., 2003). Additionally, the design should achieve the experimental goals and measure the response variable without bias. The variation and combination of factors (with their values) are explored with a grid search, effectively experimenting with a structured range of values, to allow for a wide array of measurements. By using a grid search, the aim is to find the combination of factors that result in high-quality latent topics, or algorithmic variations that lead to high-quality semantic lexicons. Additionally, random initializations are used to balance out some effects of uncontrollable factors.

Statistical analysis of the data: Within the sixth step, statistical methods are used to analyze the data so that the results and conclusions are objective in nature. The statistical tests cannot prove causality; instead, they allow for an indication of the strength of the relationships between the factors and the performance measure. Typically, the evaluation of the algorithm is done through well-studied and accepted performance metrics, such as error or loss functions.

Conclusions and recommendations: The seventh, and final step, includes the derivation of practical conclusions. Besides statistical tests, graphical methods, such as visualizations, are used to present the results. Within this thesis, results obtained from RQ1–3 feed into RQ4–6, which are focused on the interpretation of latent topics for knowledge discovery. Moreover, the computational experiments will serve as scientifically justified choices within a KDD process to convert latent topics into useful knowledge for the fisheries domain.

1.5.2 Quantitative Content Analysis

Several definitions of content analysis exist, but within the context of this thesis, we adopt the definition proposed by (Neuendorf, 2016): “Content analysis is a summarizing, quantitative analysis of messages that follows the standard of the scientific method
and is not limited as to types of variables that may be measured or the context in which the messages are created or presented.” Simply put, content analysis is a research methodology to study and make inferences from text. The scientific method guards for threats to reliability, validity, generalizability, and replicability. The goal of (any) quantitative analysis is to produce counts or measurements of categories (such as themes or topics) or variables (such as individual words) (Fink, 2009). Within the context of this thesis, the proposed definition also allows for the measurement of latent variables (elements that are not explicitly present), in contrast to definitions where content is strictly bound to manifest content (elements that are physically present and countable) (Berelson, 1952). The summarization aspect of the definition deals with summarizing results (nomothetic approach), instead of reporting on all the details within the text (idiographic approach). Within nomothetic approaches, conclusions are broadly based, generalizable, objective, summarizing, and inflexible (Neuendorf, 2016).

A downside of (traditional) content analysis is the manual coding process. To provide answers to RQ4–6, we apply quantitative content analysis to fisheries science publications through probabilistic topic models (LDA). In doing so, we overcome the need for manual coding, and commonly this type of content analysis is referred to as “computer-assisted content analysis” (Chuang et al., 2014). During this process, we utilize results obtained from RQ1–3 to increase reliability and validity of the results.

1.5.3 Social Network Analysis

Within a social network, people are typically represented as circles (called nodes) and are connected through a type of relationship (called an edge). Such relationships might entail a friendship (online or physical), a work colleague or, within the context of this thesis, a scientific collaboration (co-authorship). The language to represent social networks, the constellation of all nodes and edges, is the formal language of graph theory (West, 2000). In short, social network analysis (SNA) can be described as the “study of human relationships by means of graph theory” (Tsvetovat and Kouznetsov, 2011). More formally, SNA comprises a broad approach to sociological analysis and a set of methodological techniques that aim to describe and explore the patterns apparent in the social relationships that individuals and groups form with each other (Scott, 2017). The patterns are not limited to those derived from visualization only, but can include computational patterns as well, such as detecting communities from large networks (Clauset et al., 2004; Girvan and Newman, 2002; Newman, 2012a). SNA has also shown to allow for a variety of extensions, for example, by modeling knowledge velocity and viscosity within networks (Helms and Buijsrogge, 2005).

To provide answers to RQ6, we utilize SNA to detect community structures found within a large social network created from authors (nodes) and their collaborations through co-authorship (edges).
1.6 Dissertation Outline

This thesis consists of eight chapters, of which Chapters 2–7 discuss the six research questions described in Section 1.4. Before outlining each chapter in particular, we present an overview in Table 1.1 of Chapters 2–7 along with the research question each addresses, the corresponding publication, the employed research method, and a summary of the dataset that we constructed for the domain of fisheries science.

Table 1.1: High-level overview of Chapters 2–7 in relation to the research questions, the corresponding publications, the employed research methods, and descriptions of the datasets used.

<table>
<thead>
<tr>
<th>Ch.</th>
<th>RQ</th>
<th>Publication</th>
<th>Research Method</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>RQ6</td>
<td>S. Syed, L. ni Aodha, C. Scougal, and M. Spruit. Mapping the global network of fisheries science collaboration. Reinforcing or broad-based structures of knowledge production? (submitted for publication). 2018b</td>
<td>Quantitative content analysis; Social network analysis</td>
<td>– 73,240 abstracts – 106,137 authors – 100,175 affiliations</td>
</tr>
</tbody>
</table>
Chapter 1 — Introduction

The introduction places this thesis into context and highlights the relevance of this work. We start with an introduction of the challenges of unstructured data, a description of the Knowledge Discovery in Databases (KDD) process, the technique of topic modeling and specifically Latent Dirichlet Allocation, and a brief description of the fisheries domain in which this thesis is embedded. It then provides an overview of the research questions and employed research methods.

Chapter 2 — Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation

This chapter answers RQ1, which constitutes the basis of the data selection phase within a KDD process. Through computational experiments, we examine the effects that abstract and full-text data have upon the quality of the derived latent topics. Different datasets in terms of size and scope are analyzed. The quality of the derived latent topics is measured by coherence scores, a quantitative metric that correlates well with the qualitative human interpretation, which is considered the gold-standard in topic modeling. We also evaluate the quality of topics using a domain expert evaluation. Furthermore, we provide guidelines for the data pre-processing step that help increase the quality of latent topics, which is a crucial second step within the KDD process.


Chapter 3 — Exploring Symmetrical and Asymmetrical Dirichlet Priors for Latent Dirichlet Allocation

This chapter answers RQ2 and deals with hyper-parameter optimization within the KDD data mining phase. Through computational experiments, we examine the effects of symmetrical and asymmetrical Dirichlet priors on the quality of the derived latent topics, concerning both the topics within documents, as well as the words within topics. As a result, a total of four different combinations of Dirichlet priors are studied concurrently, with a randomized and grid-search approach to allow for a wide array of experimental results. We furthermore evaluate the quality of derived latent topics using domain expert evaluation. Additionally, by conducting computational experiments on different datasets (in size and scope), we aim to enhance generalizability.


A short version of this chapter was published as: S. Syed and M. Spruit. Select-
Chapter 4 — Bootstrapping a Semantic Lexicon on Verb Similarities

This chapter answers RQ3 and examines the construction of a semantic lexicon—a list of semantically related words. We conduct computational experiments by comparing different algorithms with the aim of creating semantic lexicons from a list of initial seed words. The semantic lexicon can assist as a validation and assessment process for the quality of latent topics, which can be mapped to the evaluation step within the KDD process. Here, we aim to develop optimized algorithms that facilitate such evaluation, rather than performing the actual validation step.


Chapter 5 — Narrow Lenses for Capturing the Complexity of Fisheries: A Topic Analysis of Fisheries Science from 1990 to 2016

This chapter answers RQ4 and deals with the knowledge discovery process concerning the obtained latent topics (patterns) from fisheries science publications. We apply quantitative content analysis to understand the knowledge captured in latent topics from 46,582 full-text articles from 21 fisheries journals over a period of 26 years (1990–2016). In doing so, we intend to understand the topical and temporal dynamics found within 26 years of scientific output. The derived knowledge from latent topics will aid a better understanding of fisheries sustainability by quantifying the underlying dimensions (e.g., social, ecological, economic and institutional) within fisheries science.


Chapter 6 — Using Machine Learning to Uncover Latent Research Topics in Fishery Models

This chapter answers RQ5 and deals with the knowledge discovery process concerning the obtained latent sub-topics; this is in contrast to Chapter 5, which deals with the broader and more general topics. Quantitative content analysis is employed to examine 22,236 full-text publications from 13 fisheries journals over a time span of 26 years (1990–2016). The sub-topics are derived through a re-iteration of the KDD process, where latent topics are used to produce latent sub-topics. In doing so, we strive to
improve the understanding of sub-topics within the domain of fisheries science and, more specifically, the science regarding fisheries models. Similar to Chapter 5, we additionally aim to quantify fisheries dimensions to better understand the aspects of fisheries sustainability, although our focus here is specifically on fisheries models.


**Chapter 7 — Mapping the Global Network of Fisheries Science Collaboration**

This chapter answers RQ6 by utilizing the knowledge output from one KDD process and, through a feedback loop, uses that knowledge to enrich a new KDD process. Within the first KDD process, we employ quantitative content analysis through probabilistic topic models to analyze 73,240 fisheries publications from 50 fisheries journals over a time span of 17 years (2000–2017). The first step uncovers the latent topical structure of the constructed corpus, which provides an overview of the latent topics found within all the documents, and the topical distribution of each individual document. In the second step, we employ techniques from the field of social network analysis to construct a co-author network of 106,137 authors from 100,175 different affiliations. We use community detection algorithms to detect hidden communities (sub-networks) within the larger co-author network. Besides studying spatial (through affiliation data) and temporal (through publication years) characteristics of the network, we enrich the new knowledge discovery process with a topic model analysis.

**Submitted for publication:** S. Syed, L. ni Aodha, C. Scougal, and M. Spruit. Mapping the global network of fisheries science collaboration. Reinforcing or broad-based structures of knowledge production? (submitted for publication). 2018b

**Chapter 8 — Conclusions**

In this final chapter of this thesis, we answer our research questions based on the results obtained from the previous chapters. We discuss the results and findings, the implications, the limitations, we provide directions for future research, and highlight some personal reflections.
Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation

This paper assesses topic coherence and human topic ranking of uncovered latent topics from scientific publications when utilizing the topic model latent Dirichlet allocation (LDA) on abstract and full-text data. The coherence of a topic, used as a proxy for topic quality, is based on the distributional hypothesis that states that words with similar meaning tend to co-occur within a similar context. Although LDA has gained much attention from machine-learning researchers, most notably with its adaptations and extensions, little is known about the effects of different types of textual data on generated topics. Our research is the first to explore these practical effects and shows that document frequency, document word length, and vocabulary size have mixed practical effects on topic coherence and human topic ranking of LDA topics. We furthermore show that large document collections are less affected by incorrect or noise terms being part of the topic-word distributions, causing topics to be more coherent and ranked higher. Differences between abstract and full-text data are more apparent within small document collections, with differences as large as 90% high-quality topics for full-text data, compared to 50% high-quality topics for abstract data.

This work was originally published as:

CHAPTER 2. FULL-TEXT OR ABSTRACT?

2.1 Introduction

There is an ever-growing amount of scientific literature with which scientists must grapple and which threatens to overwhelm their capacity to stay up to date with new research (Larsen and von Ins, 2010). As a consequence, increased availability of tools and algorithms is necessary to match the ever-growing scientific output (Boyack and Klavans, 2014). These tools and algorithms could aid in exploring large document collections in alternative and structured new ways in contrast to traditional searches. This is especially important as the topics within articles, the main ideas within articles that can be shared among similar articles, cannot always be detected through traditional keyword searches (Srivastava and Sahami, 2009).

Topic models are machine-learning algorithms to uncover hidden or latent thematic structures (i.e. topics) from large collections of documents (Deerwester et al., 1990; Hofmann, 1999; Blei et al., 2003; Blei and Lafferty, 2009). The latent thematic structures automatically emerge from the statistical properties of the documents and, as such, no prior labeling or annotation is necessary. In turn, the thematic structures can be used to automatically categorize or summarize documents up to a scale that would be impossible to do manually. Topic modeling approaches have proved to be very helpful in elucidating the key ideas within a set of documents (Griffiths and Steyvers, 2004; Grimmer and Stewart, 2013; Rusch et al., 2013), and they do so with greater speed and a quantitative rigor that would otherwise be possible only through a traditional narrative review (Grimmer and Stewart, 2013).

One of the most popular and highly researched topic models is latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is a generative probabilistic topic model that overcomes the limitations of other well-known topic model algorithms such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999). LDA models documents as multinomial distributions over $K$ latent topics and each topic is modeled as a multinomial distribution over the fixed vocabulary $V$. As such, LDA captures the heterogeneity of research topics or ideas within scientific publications and can be viewed as a mixed membership model (Erosheva et al., 2004).

Utilizing LDA to uncover latent topics from textual data has been successfully applied in several research domains. Griffiths and Steyvers (2004) performed LDA on 28,154 abstracts of the journal *Proceedings of the National Academy of Sciences* (PNAS) to uncover topics and to illustrate their relation to the journal’s categorization scheme. Gatti et al. (2015) used LDA on 80,757 abstracts from 37 primary journals from the fields of operations research and management science (OR/MS) to gain insight into the historical and current publication trends. A similar approach was performed within the field of transportation research on 17,163 abstracts from 22 leading transportation journals (Sun and Yin, 2017) and within the field of conservation science on 9,834 abstracts (Westgate et al., 2015). Besides being performed on abstract data, LDA has also been applied to 12,500 full-text research articles within the field of computational lin-
guistics (Hall et al., 2008), 2,326 articles from Neural Information Processing Systems papers (NIPS) (Wang and McCallum, 2006), and 1,060 articles within agricultural and resource economics (Alston and Pardey, 2016).

However, the reason for choosing abstract data over full-text data, or vice versa, when using LDA has not been argued for. Although some researchers (e.g. (Gatti et al., 2015)) mention that abstract data is likely to contain a high density of words, thus making it suitable for LDA, others simply mention the dataset without explaining the rationale for the choice. There are various reasons why this might be the case: one might simply only have access to abstract data (i.e. availability), one may want to keep the computational time to a minimum (i.e. feasibility), or one may want to reduce the pre-processing steps that are often necessary when dealing with full-text articles (i.e. simplicity). These pre-processing steps could include scraping the publishers’ repositories; converting PDF to plain-text, either direct or with the aid of optical character recognition (OCR) software; or an increased boilerplate cleaning phase. However, a more scientific rationale is required to aid in the choice of abstract or full-text data when uncovering latent topics with LDA.

This research is the first to explore the practical effects of choosing abstract or full-text data when uncovering latent topics with LDA. In particular, it shows the practical effects when revealing latent semantic structures from documents concerning scientific research publications. The differences between topics are calculated with a topic coherence measure (Aletras and Stevenson, 2013; Röder et al., 2015; Stevens et al., 2012; O’Callaghan et al., 2015) that shows, in contrast to the likelihood of held-out data, a higher correlation with human topic ranking data, the gold standard for topic interpretability. The underlying idea of topic coherence is rooted in the distributional hypothesis of linguistics (Harris, 1954)—namely, words with similar meanings tend to occur in similar contexts. Additionally, we use the knowledge of a domain expert to rank topics, thus providing, along with topic coherence, a comparison of topic quality from a human perspective.

2.2 Background

2.2.1 Latent Dirichlet Allocation

LDA is a generative probabilistic topic model that aims to uncover latent or hidden thematic structures from a corpus $D$. The latent thematic structure, expressed as topics and topic proportions per document, is represented by hidden variables that LDA posits onto the corpus. The generative nature of LDA describes an imaginary random process based on probabilistic sampling rules from which we assume that the documents come from. However, we only observe the words within documents and need to infer the hidden structure, that is, the topics and topic proportions per document, by applying
statistical inference techniques. This process aims to answer the question: Which hidden structure or topic model is most likely to have generated these documents? In doing so, we obtain the posterior distribution that captures the hidden structure given the observed documents. The generative process is defined as follows:

1. For every topic \( k = \{1, \ldots, K\} \)
   
   (a) draw a distribution over the vocabulary \( V \), \( \beta_k \sim \text{Dir}(\eta) \)

2. For every document \( d \)
   
   (a) draw a distribution over topics, \( \theta_d \sim \text{Dir}(\alpha) \) (i.e. per-document topic proportion)
   
   (b) for each word \( w \) within document \( d \)
      
      i. draw a topic assignment, \( z_{d,n} \sim \text{Mult}(\theta_d) \), where \( z_{d,n} \in \{1, \ldots, K\} \) (i.e. per-word topic assignment)
      
      ii. draw a word \( w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}}) \), where \( w_{d,n} \in \{1, \ldots, V\} \)

Each topic \( \beta_k \) is a multinomial distribution over the vocabulary \( V \) and comes from a Dirichlet distribution \( \beta_k \sim \text{Dir}(\eta) \). Additionally, every document is represented as a distribution over \( K \) topics and come from a Dirichlet distribution \( \theta_d \sim \text{Dir}(\alpha) \). The Dirichlet parameter \( \alpha \) denotes the smoothing of topics within documents, and \( \eta \) denotes the smoothing of words within topics. The joint distribution of all the hidden variables \( \beta_K \) (topics), \( \theta_D \) (per-document topic proportions), \( z_D \) (word topic assignments), and observed variables \( w_D \) (words in documents) is expressed by (2.1):

\[
p(\beta_K, \theta_D, z_D, w_D|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{d,k}) \tag{2.1}
\]

Figure 2.1 shows the LDA probabilistic graphical model in plate notation (Buntine, 1994), where the unshaded nodes represent the hidden random variables, the shaded nodes the observed random variables, and the edges the conditional dependencies between them. The rectangles, called plates, represent replication. The graphical model is equivalent to the joint probability of all the hidden and observed variables expressed in (2.1). We have \( K \) topics \( \beta_K \) (K-plate) as distributions over words depending on the Dirichlet parameter \( \eta \), i.e. \( \prod_{k=1}^{K} p(\beta_k|\eta) \). For all \( D \) documents (D-plate) we have a per-document topic proportion \( \theta_d \) depending on the Dirichlet parameter \( \alpha \), i.e. \( \prod_{d=1}^{D} p(\theta_d|\alpha) \). Finally, for all \( N \) words (N-plate) of a document \( d \in D \), we find that the per-word topic assignment \( z_{d,n} \) depends on the previously drawn per-document topic proportion \( \theta_d \), and the drawn word \( w_{d,n} \) depends on the per-word topic assignment \( z_{d,n} \) and all the topics \( \beta_{d,k} \), i.e. \( \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{d,k}) \) [we retrieve the probability of \( w_{d,n} \) (row) from \( z_{d,n} \) (column) within the \( K \times V \) topic matrix].
Figure 2.1: LDA represented as a graphical model in which the nodes denote the random variables and the edges the dependencies between them. Unshaded nodes are unobserved or hidden variables and the shaded nodes represent the observed random variables. The boxes, called plates, indicate replication.

The per-word topic assignment, the per-document topic distribution, and the topics are the latent variables and are not observed. We would have to condition on the only observed variable, the words within the documents, to infer the hidden structure with statistical inference. This can be viewed as a reversal of the generative process. The conditional probability, also known as the posterior, is expressed by (2.2):

\[
p(\beta_K, \theta_D, z_D | w_D) = \frac{p(\beta_K, \theta_D, z_D, w_D)}{p(w_D)} \tag{2.2}
\]

Unfortunately, computation of the posterior is intractable due to the denominator (Blei et al., 2003). The marginal probability \(p(w_D)\) is the sum of the joint distribution over all instantiations of the hidden structure and is exponentially large (Blei, 2012). Although the posterior cannot be computed exactly, a close enough approximation to the true posterior can be achieved with statistical posterior inference. Mainly two types of inference techniques can be discerned: sampling-based algorithms (e.g. (Newman et al., 2007; Porteous et al., 2008)) and variational-based algorithms (e.g. (Blei and Jordan, 2006; Teh et al., 2006; Wang et al., 2011)). It is important to note that both variational and sampling-based algorithms provide similarly accurate results (Asuncion et al., 2012).
2.2.2 Topic Coherence Measurement

After approximating LDA’s posterior distribution, the $K$ topics are represented as multinomial distributions over $V$. Each topic distribution contains every word but assigns a different probability to each of the words. The words within topics with high probability are words that tend to co-occur more frequently. These high-probability words, usually the top 10 or top 15, are used to interpret and semantically label the topics. However, LDA outputs as many topics as are defined by $K$: a low $K$ results in too few or very broad topics, whereas a high $K$ results in uninterpretable topics or topics that ideally should have been merged. Choosing the right value of $K$ is thus an important task in topic modeling algorithms, including LDA.

Measures such as the predictive likelihood of held-out data (Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan and Mimno, 2009) have been proposed to evaluate the quality of generated topics. However, such a measure correlates negatively with human interpretability (Chang et al., 2009), making topics with high predictive likelihood less coherent from a human perspective. This is especially important when generated topics are used for browsing document collections by users or understanding the trends and development within a particular research field. As a result, researchers have proposed topic coherence measures, which are a qualitative approach to automatically uncover the coherence of a topic (Aletras and Stevenson, 2013; Newman et al., 2010a), and the underlying idea is rooted in the distributional hypothesis of linguistics (Harris, 1954); words with similar meanings tend to occur in similar contexts. The topics are considered to be coherent if all or most of the words, for example, the topic’s top $N$ words, are related. The computational challenge is to obtain a measure that correlates highly with human topic ranking data, such as topic ranking data obtained by word and topic intrusion tests (Chang et al., 2009). Human topic ranking data are often considered to be the gold standard, and consequently a measure that correlates well is a good indicator of topic interpretability. A recent study by Röder et al. (Röder et al., 2015) systematically and empirically explored the multitude of topic coherence measures and their correlation with available human topic ranking data. Additionally, new coherence measures obtained by combining existing elementary elements were explored as well. Their systematic approach revealed a new unexplored coherence measure, which they labeled $C_V$, to achieve the highest correlation with all available human topic ranking data. As a result, this study adopts the $C_V$ coherence measure for topic coherence calculations. $C_V$ is based on four parts: (i) segmentation of the data into word pairs, (ii) calculation of word or word pair probabilities, (iii) calculation of a confirmation measure that quantifies how strongly a word set supports another word set, and finally (iv) aggregation of individual confirmation measures into an overall coherence score.

(i) Data segmentation pairs each of the topic’s top-$N$ words with every other top-$N$ word. Let $W$ be the set of a topic’s top-$N$ most probable words $W = \{W_1, ..., W_N\}$, $S_i$ a segmented pair of each word $W' \in W$ paired with all other words $W^* \in W$, and $S$ the set of all pairs defined as $S = \{(W', W^*) \mid W' = \{w_i\}; w_i \in W; W^* = W\}$. For example, if $W = \{w_1, w_2, w_3\}$, then a pair $S_i = (W' = w_1), (W^* = w_1, w_2, w_3)$. Such segmentation
measures the extent to which the subset $W^*$ supports, or conversely undermines, the subset $W'$ (Douven and Meijs, 2007).

(ii) Probabilities of single words $p(w_i)$ or the joint probability of two words $p(w_i, w_j)$ can be estimated by Boolean document calculation, that is, the number of documents in which $(w_i)$ or $(w_i, w_j)$ occurs, divided by the total number of documents. The Boolean document calculation, however, ignores the frequencies and distances of words. $C_V$ incorporates a Boolean sliding window calculation in which a new virtual document is created for every window of size $s$ when sliding over the document at a rate of one word token per step. For example, document $d_1$ with words $w$ results in virtual documents $d'_1 = \{w_1, \ldots, w_s\}$ and $d'_2 = \{w_2, \ldots, w_{s+1}\}$, and so on. The probabilities $p(w_i)$ and $p(w_i, w_j)$ are subsequently calculated from the total number of virtual documents. In contrast to Boolean document calculation, the Boolean sliding window calculation tries to capture the word token proximity to some degree.

(iii) For every $S_i = (W', W^*)$, we calculate a confirmation measure $\phi$ that calculates how strongly $W^*$ supports $W'$ and is based on the similarity of $W'$ and $W^*$ in relation to all the words in $W$. To calculate this similarity, $W'$ and $W^*$ are represented as context vectors (Aletras and Stevenson, 2013) as a means to capture the semantic support of all the words in $W$. These vectors $\vec{v}(W')$ and $\vec{v}(W^*)$ are created by pairing them to all words in $W$, as exemplified in (2.3). The agreement between individual words $w_i$ and $w_j$ is calculated via normalized pointwise mutual information (NPMI), as shown in (2.4). NPMI, in contrast to pointwise mutual information (PMI), shows a higher correlation with human topic ranking data (Bouma, 2009). Additionally, $\epsilon$ is used to account for the logarithm of zero and $\gamma$ to place more weight on higher NPMI values. The confirmation measure $\phi$ of a pair $S_i$ is obtained by calculating the cosine vector similarity of all context vectors $\phi_{S_i}(\vec{u}, \vec{v})$ within $S_i$, with $\vec{v}(W') \in \vec{u}$ and $\vec{v}(W^*) \in \vec{v}$ as expressed in (2.5).

\[
\tilde{v}(W') = \left\{ \sum_{w_j \in W'} \text{NPMI}(w_i, w_j)^\gamma \right\}_{j=1,\ldots,|W|}
\]  

\[
\text{NPMI}(w_i, w_j)^\gamma = \frac{\log \frac{p(w_i, w_j) + \epsilon}{p(w_i) p(w_j)}}{-\log (p(w_i, w_j) + \epsilon)}
\]  

\[
\phi_{S_i}(\vec{u}, \vec{v}) = \frac{\sum_{i=1}^{|W|} u_i \cdot v_i}{||\vec{u}||_2 \cdot ||\vec{v}||_2}
\]  

(iv) The final coherence score is the arithmetic mean of all confirmation measures $\phi$. 

31
2.3 Methodology

2.3.1 The Experiment

This paper explores the effects of uncovered latent topics and their topic coherence score, a proxy for topic quality when applying LDA on abstract and full-text data. Besides topic coherence, we explore the effects of human topic ranking—often considered the gold standard for topic interpretability—on topics uncovered from abstract and full-text data. In doing so, we explore the practical effects that types of documents, and more specifically, word length, vocabulary size, and document frequency have on the coherence and interpretability of LDA topics.

2.3.2 Dataset

Two datasets were created that contain abstract and full-text data: $DS_1$ contains 4,417 research articles (1996 to 2016) from the journal *Canadian Journal of Fisheries and Aquatic Sciences*, and $DS_2$ contains 15,004 research articles (2000 to 2016) from 12 top-tier fisheries journals: *Canadian Journal of Fisheries and Aquatic Sciences*, *Fish and Fisheries*, *Fisheries Management and Ecology*, *Fisheries Oceanography*, *Fisheries Research*, *Fisheries Management and Ecology*, *Fisheries Oceanography*, *Fisheries Research*, *Fishery Bulletin*, *Marine and Coastal Fisheries*, *North American Journal of Fisheries Management*, *Reviews in Fish Biology and Fisheries*, *Reviews in Fisheries Science*, and *Transactions of the American Fisheries Society*. Note that $DS_1 \subset DS_2$ for $Y = 2000$ to 2016. Regular expressions were used to extract abstracts from full-text articles, as the downloaded articles appeared in full-text.

The $DS_1$ dataset relates to studies where a single scientific journal was analyzed from a domain-specific journal (e.g. (Wang and McCallum, 2006)), and $DS_2$ to studies where LDA was used to uncover topics from a multitude of related domain-specific journals (e.g. (Gatti et al., 2015; Sun and Yin, 2017; Hall et al., 2008) ). The two datasets allow for comparison of not only abstract and full-text data but also on corpus size (i.e. the number of scientific publications). An overview of $DS_1$ and $DS_2$ is given in Table 2.1, and histograms of token and vocabulary (i.e. distinct words) frequencies are displayed in Fig. 2.2.

The choice of these journals was based on two factors: (i) they are domain-specific journals but employ a broad scope of research topics from the field of fisheries, and (ii) a fisheries domain expert was available to manually label and rank the topics as an alternative means of assessing the quality of topics. Furthermore, a domain-specific journal might increase the generalizability to other domain-specific journals (e.g. journals in the domain of social psychology or resource economics) compared to a more general or broadly oriented journal such as *Nature*, *Science*, or *PLOS ONE*. The domain of fisheries includes a multitude of knowledge production approaches, from mono-
Figure 2.2: Histograms of token and vocabulary frequencies for DS$_1$ and DS$_2$ for both abstract and full-text data. (b) and (d) contain very long tails for the number of tokens (up to 18,000).
Table 2.1: Overview of the $DS_1$ and $DS_2$ datasets where $J =$ num. journals; $Y =$ time range; $D =$ num. documents; $N_d =$ mean document length; $N =$ num. tokens; $V =$ vocabulary size.

<table>
<thead>
<tr>
<th></th>
<th>$DS_1$</th>
<th></th>
<th>$DS_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>Full-text</td>
<td>Abstract</td>
</tr>
<tr>
<td>$J$</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>$D$</td>
<td>4,417</td>
<td>4,417</td>
<td>15,004</td>
</tr>
<tr>
<td>$N_d$</td>
<td>108.94</td>
<td>3,855.36</td>
<td>123.7</td>
</tr>
<tr>
<td>$N$</td>
<td>481,168</td>
<td>17,029,133</td>
<td>1,856,700</td>
</tr>
<tr>
<td>$V$</td>
<td>14,643</td>
<td>142,852</td>
<td>25,781</td>
</tr>
</tbody>
</table>

transdisciplinary. Biologists, oceanographers, mathematicians, computer scientists, anthropologists, sociologists, political scientists, economists, and researchers from many other more disciplines contribute to the body of knowledge of fisheries, together with non-academic participants such as decision makers and stakeholders. Within the domain of fisheries, research into text analytics techniques has only been applied in a number of cases (e.g. (Jarić et al., 2012; Syed et al., 2016)).

All research articles were downloaded from the journals’ repository and converted from PDF to plain text. Full-text data and abstract data were tokenized, and single-character words, numbers, and punctuation marks were removed. Furthermore, we removed all single-occurrence words, words that occurred in more than 90% of the documents, and words that belonged to a standard English stop word list ($n = 153$). Apart from grouping lowercase and uppercase words, no normalization method such as stemming or lemmatization was applied to reduce inflectional and derivational forms of words to a common base form; stemming algorithms can be overly aggressive and could result in unrecognizable words that reduce the interpretability when labeling the topics. Stemming might also lead to another problem, namely that it cannot be deduced whether a stemmed word comes from a verb or a noun (Evangelopoulos et al., 2012).

### 2.3.3 Creating LDA Models

For both datasets, and for both abstract and full-text data, we created 40 different LDA models by varying the $K$ parameter (i.e. the number of topics) from 1 to 40 and repeating this process three times ($4 \times 120$ LDA models in total). The Dirichlet parameters are set to be symmetrical for the smoothing of words within topics $\eta = \frac{1}{V}$ and topics within documents $\alpha = \frac{1}{K}$. By keeping $\alpha < 1$, the modes of the Dirichlet distribution are close to the corners, thus favoring just a few topics for every document and leaving
the larger part of topic proportions very close to zero. The LDA models are created using the Python library *Gensim* (Rehurek and Sojka, 2010). Gensim uses variational inference called online LDA (Hoffman et al., 2010) to approximate the posterior. The convergence iteration parameter for the expectation step (i.e. E-step) is set to 100; the part where per-document parameters are fit for the variational distributions [see Algorithm 2 in (Hoffman et al., 2010)].

### 2.3.4 Topic Coherence

For every LDA model created (480 in total), we calculated the $C_V$ coherence score as explained in Section 2.2.2. Segmentation of top pairs is obtained by pairing every word from the top 15 words with every other word from the top 15 words. In some cases, coherence calculations are based on the top 10 most probable words. However, as no stemming or lemmatization was applied, several words with the same base form were among the top 10 words (e.g. *sample*, *sampling*), so analyzing the top 10 words would effectively mean analyzing less than 10 distinct words. To avoid logarithms of zero when calculating coherence scores, $\epsilon$ is set to a very small number, $10^{-12}$, as proposed by Stevens et al. (2012). We furthermore set $\gamma = 1$ to place equal weights on all NPMI values as researched by Röder et al. (2015) and have shown the highest correlation with all topic ranking data, in contrast to Aletras and Stevenson (Aletras and Stevenson, 2013), where $\gamma = 2$ shows better results. To capture word proximity when calculating word or word pair probabilities, the Boolean sliding window for Boolean document calculation is set to $s = 110$ (Röder et al., 2015).

The LDA model with the optimal coherence score, obtained with an elbow method (the point with maximum absolute second derivative), was additionally analyzed by a fisheries domain expert. The domain expert is affiliated with the leading competence institution for fishery and aquaculture in Norway. The analysis consisted of an inspection of the top 15 most probable words for each topic, together with an inspection of the document titles and content. Additionally, the domain expert rated the topics (high, medium, low) by assessing the coherence of the top 15 words and the presence of incorrect terms (i.e. words) within each topic. High-quality topics contain no incorrect terms, medium-quality topics contain one or two, and low-quality topics contain three or more. An incorrect term is defined as a word that has no semantic relationship with the topic’s top 15 words. The domain expert attached a label to each topic that best captured the semantics of the top 15 words.

### 2.4 Results

Fig. 2.3 shows the obtained $C_V$ coherence scores for all 480 LDA models created, with Fig. 2.3a and Fig. 2.3b displaying the results for the $DS_1$ and $DS_2$ datasets, respectively.
CHAPTER 2. FULL-TEXT OR ABSTRACT?

Figure 2.3: Calculated $C_V$ topic coherence score for LDA models with $K = \{1, \ldots, 40\}$ for (a) $DS_1$ and (b) $DS_2$. The coherence score is the mean score for all 3 runs. Scores for $DS_1$ with 4,417 documents shows that full-text data achieves a higher topic coherence score for all $k$-values. In contrast, $DS_2$ with 15,004 documents show similar coherence scores. Individual lines for each run are not shown for clarity.

The lines represent the mean coherence scores from 3 runs where the number of topics was varied from 1 to 40. A visual inspection of Fig. 2.3a shows that LDA models created with full-text data from the $DS_1$ dataset achieve higher mean coherence scores among all values of $K$, a result that is not visible for $DS_2$ (Fig. 2.3b).

Table 2.2 displays the actual coherence score values for uncovered topics from abstract and full-text data for both datasets. It shows the mean $C_V$ coherence score ($\bar{X}$), the standard deviation ($s$), and the difference between mean values ($\bar{X}_2 - \bar{X}_1$) calculated from all three runs for $K = \{2, \ldots, 40\}$. Positive differences between mean values indicate a higher achieved coherence score for full-text data. We furthermore calculate the significance ($p < 0.05$, $p < 0.01$, and $p < 0.001$) between $\bar{X}_1$ and $\bar{X}_2$ with an independent two-sample t-test as Levene’s test for homoscedasticity assumes equal variances for all $K$ values.
Table 2.2: Calculated coherence score for abstract and full-text data for both datasets. 
\( \bar{X} = \) Mean coherence score, \( s = \) Standard deviation coherence score, \( \bar{X}_2 - \bar{X}_1 = \) Difference in mean coherence scores, \( t = \) Calculated t-statistic, \( p = \) Two-tailed p-value, \( K = \) number of topics

<table>
<thead>
<tr>
<th>Dataset DS(_1) (4,417 documents)</th>
<th>Abstract(_1)</th>
<th>Full-text(_2)</th>
<th>Statistics (t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K )</td>
<td>( \bar{X}_1 )</td>
<td>( s_1 )</td>
<td>( \bar{X}_2 )</td>
</tr>
<tr>
<td>2</td>
<td>0.392</td>
<td>0.040</td>
<td>0.547</td>
</tr>
<tr>
<td>3</td>
<td>0.454</td>
<td>0.032</td>
<td>0.536</td>
</tr>
<tr>
<td>4</td>
<td>0.433</td>
<td>0.027</td>
<td>0.556</td>
</tr>
<tr>
<td>5</td>
<td>0.454</td>
<td>0.028</td>
<td>0.575</td>
</tr>
<tr>
<td>6</td>
<td>0.479</td>
<td>0.044</td>
<td>0.572</td>
</tr>
<tr>
<td>7</td>
<td>0.503</td>
<td>0.009</td>
<td>0.560</td>
</tr>
<tr>
<td>8</td>
<td>0.509</td>
<td>0.024</td>
<td>0.567</td>
</tr>
<tr>
<td>9</td>
<td>0.492</td>
<td>0.016</td>
<td>0.576</td>
</tr>
<tr>
<td>10</td>
<td>0.475</td>
<td>0.008</td>
<td>0.566</td>
</tr>
<tr>
<td>11</td>
<td>0.473</td>
<td>0.015</td>
<td>0.578</td>
</tr>
<tr>
<td>12</td>
<td>0.491</td>
<td>0.010</td>
<td>0.572</td>
</tr>
<tr>
<td>13</td>
<td>0.484</td>
<td>0.010</td>
<td>0.591</td>
</tr>
<tr>
<td>14</td>
<td>0.515</td>
<td>0.014</td>
<td>0.568</td>
</tr>
<tr>
<td>15</td>
<td>0.475</td>
<td>0.022</td>
<td>0.583</td>
</tr>
<tr>
<td>16</td>
<td>0.485</td>
<td>0.021</td>
<td>0.585</td>
</tr>
<tr>
<td>17</td>
<td>0.489</td>
<td>0.015</td>
<td>0.590</td>
</tr>
<tr>
<td>18</td>
<td>0.506</td>
<td>0.035</td>
<td>0.592</td>
</tr>
<tr>
<td>19</td>
<td>0.493</td>
<td>0.009</td>
<td>0.589</td>
</tr>
<tr>
<td>20</td>
<td>0.493</td>
<td>0.007</td>
<td>0.584</td>
</tr>
<tr>
<td>21</td>
<td>0.504</td>
<td>0.020</td>
<td>0.579</td>
</tr>
<tr>
<td>22</td>
<td>0.497</td>
<td>0.012</td>
<td>0.576</td>
</tr>
<tr>
<td>23</td>
<td>0.486</td>
<td>0.009</td>
<td>0.572</td>
</tr>
<tr>
<td>24</td>
<td>0.489</td>
<td>0.001</td>
<td>0.584</td>
</tr>
<tr>
<td>25</td>
<td>0.471</td>
<td>0.006</td>
<td>0.567</td>
</tr>
<tr>
<td>26</td>
<td>0.490</td>
<td>0.016</td>
<td>0.589</td>
</tr>
<tr>
<td>27</td>
<td>0.482</td>
<td>0.013</td>
<td>0.573</td>
</tr>
<tr>
<td>28</td>
<td>0.488</td>
<td>0.009</td>
<td>0.585</td>
</tr>
<tr>
<td>29</td>
<td>0.500</td>
<td>0.017</td>
<td>0.590</td>
</tr>
<tr>
<td>30</td>
<td>0.475</td>
<td>0.010</td>
<td>0.583</td>
</tr>
<tr>
<td>31</td>
<td>0.478</td>
<td>0.010</td>
<td>0.584</td>
</tr>
<tr>
<td>32</td>
<td>0.488</td>
<td>0.007</td>
<td>0.588</td>
</tr>
<tr>
<td>33</td>
<td>0.484</td>
<td>0.013</td>
<td>0.581</td>
</tr>
<tr>
<td>34</td>
<td>0.488</td>
<td>0.002</td>
<td>0.594</td>
</tr>
<tr>
<td>35</td>
<td>0.502</td>
<td>0.011</td>
<td>0.584</td>
</tr>
<tr>
<td>36</td>
<td>0.481</td>
<td>0.002</td>
<td>0.578</td>
</tr>
<tr>
<td>37</td>
<td>0.491</td>
<td>0.015</td>
<td>0.591</td>
</tr>
<tr>
<td>38</td>
<td>0.476</td>
<td>0.008</td>
<td>0.580</td>
</tr>
<tr>
<td>39</td>
<td>0.483</td>
<td>0.024</td>
<td>0.576</td>
</tr>
<tr>
<td>40</td>
<td>0.494</td>
<td>0.007</td>
<td>0.586</td>
</tr>
</tbody>
</table>

*\( p < 0.05 \), **\( p < 0.01 \), ***\( p < 0.001 \)
### Table 2.2: continued.

<table>
<thead>
<tr>
<th>K</th>
<th>Dataset $D_{S2}$ (15,004 documents)</th>
<th>Abstract</th>
<th>Full-text</th>
<th>Statistics (t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{X}_1$</td>
<td>$s_1$</td>
<td>$\bar{X}_2$</td>
<td>$s_2$</td>
</tr>
<tr>
<td>2</td>
<td>0.448</td>
<td>0.016</td>
<td>0.490</td>
<td>0.004</td>
</tr>
<tr>
<td>3</td>
<td>0.434</td>
<td>0.016</td>
<td>0.517</td>
<td>0.024</td>
</tr>
<tr>
<td>4</td>
<td>0.482</td>
<td>0.020</td>
<td>0.522</td>
<td>0.014</td>
</tr>
<tr>
<td>5</td>
<td>0.484</td>
<td>0.016</td>
<td>0.520</td>
<td>0.016</td>
</tr>
<tr>
<td>6</td>
<td>0.488</td>
<td>0.017</td>
<td>0.543</td>
<td>0.010</td>
</tr>
<tr>
<td>7</td>
<td>0.507</td>
<td>0.029</td>
<td>0.529</td>
<td>0.002</td>
</tr>
<tr>
<td>8</td>
<td>0.496</td>
<td>0.010</td>
<td>0.518</td>
<td>0.019</td>
</tr>
<tr>
<td>9</td>
<td>0.527</td>
<td>0.015</td>
<td>0.531</td>
<td>0.007</td>
</tr>
<tr>
<td>10</td>
<td>0.536</td>
<td>0.007</td>
<td>0.538</td>
<td>0.013</td>
</tr>
<tr>
<td>11</td>
<td>0.539</td>
<td>0.010</td>
<td>0.536</td>
<td>0.011</td>
</tr>
<tr>
<td>12</td>
<td>0.550</td>
<td>0.013</td>
<td>0.545</td>
<td>0.006</td>
</tr>
<tr>
<td>13</td>
<td>0.538</td>
<td>0.007</td>
<td>0.533</td>
<td>0.003</td>
</tr>
<tr>
<td>14</td>
<td>0.536</td>
<td>0.014</td>
<td>0.548</td>
<td>0.003</td>
</tr>
<tr>
<td>15</td>
<td>0.558</td>
<td>0.017</td>
<td>0.555</td>
<td>0.008</td>
</tr>
<tr>
<td>16</td>
<td>0.542</td>
<td>0.007</td>
<td>0.561</td>
<td>0.010</td>
</tr>
<tr>
<td>17</td>
<td>0.562</td>
<td>0.022</td>
<td>0.557</td>
<td>0.009</td>
</tr>
<tr>
<td>18</td>
<td>0.558</td>
<td>0.015</td>
<td>0.550</td>
<td>0.005</td>
</tr>
<tr>
<td>19</td>
<td>0.543</td>
<td>0.017</td>
<td>0.553</td>
<td>0.011</td>
</tr>
<tr>
<td>20</td>
<td>0.550</td>
<td>0.019</td>
<td>0.561</td>
<td>0.006</td>
</tr>
<tr>
<td>21</td>
<td>0.569</td>
<td>0.014</td>
<td>0.560</td>
<td>0.014</td>
</tr>
<tr>
<td>22</td>
<td>0.559</td>
<td>0.016</td>
<td>0.564</td>
<td>0.006</td>
</tr>
<tr>
<td>23</td>
<td>0.562</td>
<td>0.006</td>
<td>0.562</td>
<td>0.012</td>
</tr>
<tr>
<td>24</td>
<td>0.552</td>
<td>0.008</td>
<td>0.564</td>
<td>0.006</td>
</tr>
<tr>
<td>25</td>
<td>0.548</td>
<td>0.006</td>
<td>0.564</td>
<td>0.011</td>
</tr>
<tr>
<td>26</td>
<td>0.554</td>
<td>0.018</td>
<td>0.564</td>
<td>0.011</td>
</tr>
<tr>
<td>27</td>
<td>0.553</td>
<td>0.010</td>
<td>0.561</td>
<td>0.010</td>
</tr>
<tr>
<td>28</td>
<td>0.552</td>
<td>0.004</td>
<td>0.567</td>
<td>0.014</td>
</tr>
<tr>
<td>29</td>
<td>0.543</td>
<td>0.015</td>
<td>0.560</td>
<td>0.003</td>
</tr>
<tr>
<td>30</td>
<td>0.558</td>
<td>0.007</td>
<td>0.557</td>
<td>0.012</td>
</tr>
<tr>
<td>31</td>
<td>0.557</td>
<td>0.014</td>
<td>0.568</td>
<td>0.006</td>
</tr>
<tr>
<td>32</td>
<td>0.553</td>
<td>0.002</td>
<td>0.557</td>
<td>0.003</td>
</tr>
<tr>
<td>33</td>
<td>0.541</td>
<td>0.009</td>
<td>0.564</td>
<td>0.004</td>
</tr>
<tr>
<td>34</td>
<td>0.554</td>
<td>0.010</td>
<td>0.565</td>
<td>0.013</td>
</tr>
<tr>
<td>35</td>
<td>0.550</td>
<td>0.002</td>
<td>0.568</td>
<td>0.014</td>
</tr>
<tr>
<td>36</td>
<td>0.550</td>
<td>0.016</td>
<td>0.573</td>
<td>0.010</td>
</tr>
<tr>
<td>37</td>
<td>0.545</td>
<td>0.009</td>
<td>0.576</td>
<td>0.005</td>
</tr>
<tr>
<td>38</td>
<td>0.550</td>
<td>0.008</td>
<td>0.565</td>
<td>0.003</td>
</tr>
<tr>
<td>39</td>
<td>0.546</td>
<td>0.019</td>
<td>0.577</td>
<td>0.005</td>
</tr>
<tr>
<td>40</td>
<td>0.569</td>
<td>0.016</td>
<td>0.574</td>
<td>0.009</td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
2.4.1 DS1 Dataset

Although every $\tilde{X}_2$ (full-text) outperforms $\tilde{X}_1$ (abstract), not all differences are statistically significant. For $k = 3$ and $k = 6$, the differences between coherence scores are not significant but are still very close to the 5% significance threshold. The largest difference between mean values is achieved at $k = 2$ (two-topic LDA model), although it is only significant at $p < 0.05$. Looking at all $K$ values, three runs achieve $p < 0.05$ significance, 18 achieve $p < 0.01$ significance, and 16 achieve $p < 0.001$ significance. The choice of full-text data results in overall topics with higher coherence for all $K$ values, and these differences are significant for all but two LDA models. The abstract data achieved the optimal coherence score (via elbow method) at $k = 14$, and the full-text data achieved this at $k = 13$.

2.4.2 DS2 Dataset

The $DS_2$ dataset with 15,004 research articles from 12 top-tier fisheries journals show that only 5 LDA models are significantly different at the 5% significance threshold; $k = 2, 3, 6, 33,$ and $37$. Looking at the actual coherence scores, most LDA models show a slightly higher coherence score (shown in bold) for full-text data compared to abstract data. However, the large difference in coherence scores and significance levels are not similar to the $DS_1$ dataset. The LDA model with the optimal coherence score for abstract data is at $k = 17$, and for full-text data at $k = 16$.

2.4.3 Human Topic Ranking

Table 2.3 shows the results of the human topic ranking by a fisheries domain expert. For an equal comparison, the LDA models with optimal coherence scores were ranked and compared. The LDA model from $DS_1$ abstract data ($k = 14$) contains 50% high-quality topics, 36% medium-quality topics, and 14% low-quality topics. In contrast, the LDA model from full-text data ($k = 13$) contains 92% high-quality, 8% medium-quality, and no low-quality topics. $DS_2$ abstract and full-text data show similar ranking scores; almost 90% high-quality topics with just two topics ranked as medium-quality. Table 2.4 provides an example of high- medium- and low-quality topics, the top 15 words, and the incorrect terms that caused the topics to be ranked lower for $DS_1$. A two-dimensional inter-topic distance map for the LDA models is displayed in Fig. 2.4. This two-dimensional representation is obtained by computing the distance between topics (Chuang et al., 2012) and applying multidimensional scaling (Sievert and Shirley, 2014). It displays the similarity between topics with respect to their probability distribution over words. Furthermore, it shows the topic label that best captures the semantics of the top 15 words. The color coding indicates the quality of the topics based on human interpretation (see Section 2.3.4 for ranking method). It shows that,
Table 2.3: Manual topic ranking for $DS_1$ and $DS_2$ datasets for abstract and full-text. $H$ = high-quality, $M$ = medium-quality, and $L$ = low-quality topics.

<table>
<thead>
<tr>
<th></th>
<th>$DS_1$ Abstract</th>
<th>$DS_1$ Full-text</th>
<th>$DS_2$ Abstract</th>
<th>$DS_2$ Full-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>7/14 (50.0%)</td>
<td>12/13 (92.3%)</td>
<td>15/17 (88.2%)</td>
<td>14/16 (87.5%)</td>
</tr>
<tr>
<td>$M$</td>
<td>5/14 (35.7%)</td>
<td>1/13 (7.7%)</td>
<td>2/17 (11.8%)</td>
<td>2/16 (12.5%)</td>
</tr>
<tr>
<td>$L$</td>
<td>2/14 (14.3%)</td>
<td>0/13 (0%)</td>
<td>0/17 (0%)</td>
<td>0/16 (0%)</td>
</tr>
</tbody>
</table>

Overall, more high-quality topics are obtained from full-text data than from the abstract counterpart for $DS_1$, and similar topic rankings are achieved for $DS_2$. 
Figure 2.4: Inter-topic distance map showing a two-dimensional representation (via multi-dimensional scaling) of the latent topics. The distance between the nodes represents the topic similarity with respect to the distributions of words. The surface of the nodes represents the prevalence of the topic within the corpus. Color coding is used to display the topic ranking: green = high-quality topic, orange = medium-quality topic, and red = low-quality topic.
Table 2.4: A selection of topics from $DS_1$ with the 15 most probable words, topic label, and ranking data. Text in bold indicates incorrect terms.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Label</th>
<th>Top 15 words</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>fish distribution</td>
<td>fishing, distribution, data, species, areas, catch, abundance, spatial, habitat, model, fishery, effort, fish, water, sea</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Population models</td>
<td>model, data, mortality, stock, fish, population, fishing, models, recruitment, cod, estimates, <strong>using</strong>, size, rates, <strong>used</strong></td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Population genetics</td>
<td>genetic, populations, population, <strong>among</strong>, lake, fish, loci, microsatellite, <strong>two</strong>, structure, diversity, <strong>within</strong>, samples, species, river</td>
<td>Low</td>
</tr>
<tr>
<td>Full-text</td>
<td>Salmon population dynamics</td>
<td>salmon, trout, prey, growth, atlantic, temperature, water, rate, juvenile, salmo, feeding, wild, density, food, populations</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Population genetics</td>
<td>genetic, populations, population, river, samples, loci, salmon, <strong>among</strong>, dna, atlantic, sample, sea, microsatellite, structure, alleles</td>
<td>Medium</td>
</tr>
</tbody>
</table>
2.5 Discussion

The coherence of a topic is based on the topic’s top 15 words and shows how strongly pairs of these top 15 words support each other within the corpus. Such an approach, drawing on the philosophical premise that a set of statements or facts is said to be coherent if its statements or facts support each other, informs us about the understandability and interpretability of topics from a human perspective. The LDA models obtained from DS1 full-text data, compared to DS1 abstract data, show a higher coherence overall, with the test statistics showing that these differences are significant for all but two LDA models. On the other hand, such significant differences are not present within the DS2 dataset when comparing abstract and full-text data, although full-text data achieved more topics with a higher coherence score.

Additionally, topic ranking by a fisheries domain expert shows similar, or even greater, improvements in results for the DS1 full-text data; topics uncovered from full-text data contain 92% high-quality topics compared to 50% high-quality topics from abstract data. The quality of topics from a human perspective was lowered by the inclusion of incorrect terms in the top 15 words. Such terms, however, are not related to the biological, ecological, or socio-ecological meanings of those topics but can be seen as noise terms: using, used, use, within, total, two, among, and within. There is little to no specific semantic meaning behind these terms, and although they are important in written text, they are less important when uncovering latent semantic structures (i.e. topics) from documents. This issue may be potentially rectified by a part-of-speech (POS) tagger to eliminate the verbs or prepositions that crop up as noise among the topic’s top words. However, one should proceed carefully in cases where verbs are important cues for understanding the semantics of the top words. For example, Table 2.4 shows that fishing and feeding are among the top 15 words, and in these cases, the verbs are important terms that are necessary for the understanding of the semantic context. In such cases, one might proceed with a domain-specific stop word list to prevent such terms from becoming part of the topic-word distribution. A lower ranked topic caused by noise terms is not as apparent for full-text data, nor does it seem to hold for abstract data from the DS2 dataset. Such noise terms seem less of an issue when document frequency, word length, or vocabulary size increases.

Also worth noting is an increased level of detailed topics within DS1 full-text data (Fig. 2.4b) compared to DS1 abstract data (Fig. 2.4a). For example, the topics salmon population dynamics and salmon reproduction were uncovered from full-text data, where the single topic salmon was uncovered from abstract data. Similarly, the topics dealing with lakes are split into three topics (lake nutrients and algae, lake sediments, and lake ecology) from full-text data, compared to two (lake tropic level interaction and lake nutrients and algae) from the abstract data. Lastly, the topics dealing with models were split into three (estimation models, stock assessment models, and reproduction models) from full-text data in contrast to an overarching population model topic from abstract data. Such a clear difference between low and high granularity topics is not present
within the $DS_2$ dataset. Although the differences in word length and vocabulary size exists, similarly to $DS_1$, it seems that a higher number of documents makes up for these differences in granularity. A comparison between other LDA models (not presented) shows similar granularity between abstract and full-text for $DS_2$. Although the article’s abstract aims to provide a complete but succinct description of the whole paper, it is often restricted by a limitation on the number of words. Such word limitation, with a relatively small number of documents, has practical effects on the level of detail (i.e. granularity) of uncovered LDA topics.

Besides topic coherence, topic ranking, and the level of detail, Fig. 2.4 shows a number of uncovered topics that are present in abstract data but absent in full-text data. Within $DS_1$ for example, the topics temperature effects, cod genetics, management, and fish abundance were not found within full-text data, and neither were related topics showing semantic resemblance to these absent topics. Although we identified similar and detailed topics, there remains an inconsistency between some uncovered topics from both datasets. Knowing that abstracts were retrieved from full-text articles and are thus, in essence, a subset of the full-text data, one might question why these differences exist. One reason might be that manual topic labeling is limited to the subjectivity inherent in human interpretation and an analysis of the topics by others could yield opposite results, explaining away any differences between the two datasets. On the other hand, topic labeling is usually performed by inspection of the topic words with the highest probabilities (top 10 or 15). Such an approach might up-weight terms that have high probability under all topics. Other approaches to identify the terms that best describe a topic exist (e.g. (Blei and Lafferty, 2009; Tang and Maclennan, 2005)) and could yield different results. Finally, abstract data, being restricted by the limited number of words, fail to adequately convey the heterogeneity of research ideas or topics that are part of a document. Uncovered latent topics might thus not completely resemble the document collection and, as a result, provide a limited or even incorrect view of the underlying thematic structure.

2.6 Conclusion

In this paper, we presented a comparison between topic coherence scores and human topic ranking when creating LDA topics from abstract and full-text data. Two datasets were compared, $DS_1$ consisting of a single fisheries journal with 4,417 scientific research articles that span 20 years of scientific output, and $DS_2$ consisting of 12 fisheries journals, 15,004 articles, and span 16 years of research. The two types of data, abstract and full-text, combined with two different datasets, a single journal and a set of journals, allow for comparison on a variety of characteristics, such as document length, document frequency, and vocabulary size. Topics were statistically compared by adopting the $C_V$ coherence measure that shows the highest correlation with all available human topic-ranking data. Furthermore, the LDA models with the optimal coherence scores were manually inspected and ranked by a fisheries domain expert.
Our results show that uncovering LDA models from a single journal with, relatively speaking, a low number of documents are very prone to noise terms that crop up into the topic’s top words—the words that are often used to capture the semantics of the topic—for abstract data. Such noise terms require special attention when dealing with abstract data with, e.g. an increased cleaning phase, POS filtering, or a domain-specific stop word list. Our results show that full-text data seem less affected by such words, thus increasing the coherence and manual topic ranking. On the other hand, increasing the number of document (e.g. $DS_2$) results in fewer noise terms, thus an improvement in coherence and human topic ranking for both abstract and full-text data. Furthermore, on a small dataset (e.g. $DS_1$) abstract topic distributions capture more broad topics, with full-text topics achieving more fine-grained results. These differences in detail are not present for bigger datasets containing a higher number of documents, regardless the choice for abstract or full-text data.

We identified a number of topics that were uncovered from abstract data but were absent among the topics uncovered from full-text data. A detailed analysis of the reasons behind these differences would yield interesting results and would be a possible direction for future research.
Chapter 3

Exploring Symmetrical and Asymmetrical Dirichlet Priors for Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) has gained much attention from researchers and is increasingly being applied to uncover underlying semantic structures from a variety of corpora. However, nearly all researchers use symmetrical Dirichlet priors, often unaware of the underlying practical implications that they bear. This research is the first to explore symmetrical and asymmetrical Dirichlet priors on topic coherence and human topic ranking when uncovering latent semantic structures from scientific research articles. More specifically, we examine the practical effects of several classes of Dirichlet priors on 2000 LDA models created from abstract and full-text research articles. Our results show that symmetrical or asymmetrical priors on the document–topic distribution or the topic–word distribution for full-text data have little effect on topic coherence scores and human topic ranking. In contrast, asymmetrical priors on the document–topic distribution for abstract data show a significant increase in topic coherence scores and improved human topic ranking compared to a symmetrical prior. Symmetrical or asymmetrical priors on the topic–word distribution show no real benefits for both abstract and full-text data.

This work was originally published as:


CHAPTER 3. EXPLORING DIRICHLET PRIORS

3.1 Introduction

Global research efforts have led to an ever-increasing amount of scientific output. Combined with the digitalization of scientific archives, this increase is threatening to overwhelm today’s scientists trying to keep track of and identify relevant literature (Larsen and von Ins, 2010). Consequently, scientists need new tools and algorithms for browsing these collections in a structured way, particularly as topics within articles, which are the ideas contained within articles that can be shared among similar articles, cannot always be detected through traditional keyword searches (Srivastava and Sahami, 2009). Probabilistic topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) and probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) are machine-learning algorithms used to automatically uncover underlying semantic structures, such as themes or topics, in large collections of documents. These underlying semantic structures can subsequently be used to categorize, summarize, and annotate large document collections in a purely unsupervised fashion.

LDA, although the simplest topic model, has received much attention from machine-learning researchers and has been adopted and extended in many ways. LDA is a generative probabilistic topic model that aims to uncover hidden or latent thematic structures from large collections of documents. LDA is a three-level hierarchical Bayesian model that models documents as discrete distributions over $K$ latent topics, and every topic is modeled as a multinomial distribution over the fixed vocabulary. Uncovering latent thematic structures proceeds through posterior inference of the latent variables given the observed words. Apart from its applicability to text, LDA has proven useful to other types of data, such as image (Fergus et al., 2005), video (Mehran et al., 2009), and audio (Kim et al., 2009).

As a conjugate prior to the multinomial distribution, LDA uses a Dirichlet prior to simplify posterior inference. Typically, these priors and related hyperparameters are set to be symmetrical, assuming that a priori all topics have equal probability to be assigned to a document and all words have an equal chance to be assigned to a topic. The reasons for choosing symmetrical priors, compared to asymmetrical priors, are not explicitly stated and are often implicitly assumed to have little or no practical effect (Wallach et al., 2009). However, hyperparameters can have a significant effect on the achieved accuracy for various inference techniques, such as Gibbs sampling, variational Bayes, or collapsed variational Bayes (Asuncion et al., 2012). In fact, inference methods have relatively similar predictive performance when the hyperparameters are optimized, thereby explaining away most differences between them.

Little research has examined the effects of Dirichlet priors on the quality of generated topics. Among the few, Wallach et al. (2009) demonstrated that using an asymmetric Dirichlet prior on the document–topic distribution shows significant performance gains concerning the likelihood of held-out documents. However, the likelihood correlates negatively with human interpretability (Chang et al., 2009), which is often considered the gold standard for topic quality. Consequently, researchers have proposed topic co-
herence measures (Aletras and Stevenson, 2013; Stevens et al., 2012; Newman et al., 2010a; Röder et al., 2015), a proxy for topic quality that shows improved correlation with human topic ranking data. The underlying idea of topic coherence is rooted in the distributional hypothesis of linguistics (Harris, 1954)—namely, words with similar meanings tend to occur in similar contexts. This paper is the first to explore the practical effects of several classes of Dirichlet priors on the coherence of generated topics. More specifically, we study topic coherence for the combinations of symmetrical and asymmetrical priors on the document–topic distribution, as well as the topic–word distribution, when uncovering latent topics with LDA. In addition, topics are ranked by a domain expert on interpretability, providing a qualitative analysis of topic quality for different classes of Dirichlet priors in addition to a quantitative measure. Such analyses can provide valuable guidance to researchers utilizing LDA tools such as Mallet and Gensim (Rehurek and Sojka, 2010) to uncover topical structures from scientific articles (Gatti et al., 2015; Sun and Yin, 2017; Westgate et al., 2015; Wang and McCallum, 2006; Alston and Pardey, 2016) and unknowingly leaving hyperparameters set to default (i.e. symmetrical).

3.2 Background

3.2.1 Latent Dirichlet Allocation

LDA is a generative probabilistic topic model that aims to uncover latent semantic structures from a set of documents, \( D \). The latent semantic structures can subsequently be used to organize, categorize, and annotate documents without the need of prior human labeling or annotation. LDA models documents as discrete distributions over \( K \) latent topics, and every topic is modeled as a discrete distribution over the fixed vocabulary. As a result, LDA captures the heterogeneity of ideas prevailing in a document collection and can be viewed as a mixed membership model (Erosheva et al., 2004). The underlying latent semantic structure is expressed by topics \( \beta \), topic proportions \( \theta \), and topic assignments \( z \) and includes hidden variables that LDA posits into the corpus. However, \( \beta \), \( \theta \), and \( z \) are unobserved, and the goal is to determine them from the observed variables (i.e. the words within the documents). LDA's structure allows the observed variables to interact with structured distributions of a hidden variable model (Blei and Lafferty, 2009). Learning the hidden variables (i.e. the underlying semantic structure) can be achieved by inferring the posterior distribution of the latent variables given the observed documents. The interaction between latent and observed variables is manifested in the generative process behind LDA, the imaginary random process in which we assume the documents come from and are based on probabilistic sampling rules. The generative process is described as follows:

1. For every topic \( k = \{1, ..., K\} \)
Figure 3.1: LDA represented as a graphical model in which the nodes denote the random variables and the edges the dependencies between them. Unshaded nodes are unobserved or hidden variables, and the shaded nodes represent the observed random variables. The boxes, called plates, indicate replication.

(a) draw a distribution over the vocabulary $V$, $\beta_k \sim \text{Dir}(\eta)$

2. For every document $d$

(a) draw a distribution over topics, $\theta_d \sim \text{Dir}(\alpha)$ (i.e. per-document topic proportion)

(b) for each word $w$ within document $d$

i. draw a topic assignment, $z_{d,n} \sim \text{Mult}(\theta_d)$, where $z_{d,n} \in \{1, ..., K\}$ (i.e. per-word topic assignment)

ii. draw a word $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$, where $w_{d,n} \in \{1, ..., V\}$

Where $K$ is the numbers of topics, $V$ is the vocabulary size, and $\alpha$ and $\eta$ are the Dirichlet hyperparameters that affect the smoothing of topic proportions within documents and words within topics, respectively. The joint distribution (see Figure 3.1 for the graphical representation) of all the hidden and observed variables becomes:

$$p(\beta_K, \theta_D, z_D, w_D | \alpha, \eta) = \prod_{k=1}^{K} p(\beta_k | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta_{z_{d,n}})$$ (3.1)

To learn the distribution of the hidden variables, we invert the generative process and fit the hidden variables onto the observed words. The hidden structure is thus described by the posterior distribution of the latent variables given the observed words:
CHAPTER 3. EXPLORING DIRICHLET PRIORS

\[ p(\beta_K, \theta_D, z_D|w_D, \alpha, \eta) = \frac{p(\beta_K, \theta_D, z_D, w_D|\alpha, \eta)}{p(w_D|\alpha, \eta)} \tag{3.2} \]

\[ p(w_D|\alpha, \eta) = \int_{\beta_K} \int_{\theta_D} p(w_d|\alpha, \eta) \tag{3.3} \]

However, the posterior is intractable to compute (Blei et al., 2003) due to the evidence as expressed in (3.3). The solution is to approximate the posterior using inference techniques. Two main posterior inference techniques can be discerned: (i) sampling-based algorithms (Newman et al., 2007; Porteous et al., 2008) and (ii) variational- or optimization-based algorithms (Blei and Jordan, 2006; Teh et al., 2006; Wang et al., 2011). Sampling-based algorithms, such as Markov Chain Monte Carlo (MCMC) sampling, sample from the posterior—usually one variable at a time—while fixing the other variables. Repeating this process for several iterations causes the process to converge, in which the sample values have the same distribution as if they came from the true posterior. Variational inference aims to find a simplified parametric distribution that is closest to the true posterior measured in the Kullback-Leibler (KL) divergence. Once inference is complete, the posterior distribution reveals the latent structure of the documents expressed by topics \( \beta \), topic proportions \( \theta \), and topic assignments \( z \).

One way to think about LDA is to imagine a document in which one highlights words with colored markers. Words that relate to one topic are colored blue, words that relate to another topic are colored red, and so on. After all of the words have been colored (excluding words such as "the", "a"), all the words with the same color are the topics, and the article will blend the colors in different proportions. Different documents will have different blends of colors, and we could use the proportion of the various colors to situate this specific document in a document collection (e.g. documents addressing mainly the blue topic). Moreover, documents with the same blend of colors discuss the topics in similar proportion and are considered closely related from a topical perspective. Technically, documents with similar topic distributions are close in Kullback-Leibler divergence, a measure to calculate the distance between two probability distributions. LDA as a statistical model captures this intuition. We refer the interested reader to (Blei, 2012) for a concise introduction to LDA.

### 3.2.2 Research Utilizing LDA

Topic modeling algorithms, and specifically LDA, have been helpful in elucidating the key ideas within a set of documents, such as articles published in the journal PNAS (Griffiths and Steyvers, 2004), political science texts (Grimmer and Stewart, 2013) or data-driven journalism (Rusch et al., 2013). Moreover, it is considered that this approach could provide insight into the development of a scientific field and changes
in research priorities (Neff and Corley, 2009), and do so with greater speed and quantitative rigor than would otherwise be possible through traditional narrative reviews (Grimmer and Stewart, 2013). As such, LDA has been applied, for example, in the domain of transportation research (Sun and Yin, 2017), computer science (Hall et al., 2008; Wang et al., 2011; Wang and McCallum, 2006), fisheries science (Syed and Weber, 2018; Syed et al., 2018a), conservation science (Westgate et al., 2015), and the fields of operations research and management science (Gatti et al., 2015).

### 3.2.3 Coherence Scores

Measures such as predictive likelihood on held-out data (Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan and Mimno, 2009) have been proposed to evaluate the quality of generated topics. However, such measures correlate negatively with human interpretability (Chang et al., 2009), making topics with high predictive likelihood less coherent from a human perspective. High-quality or coherent latent topics are of particular importance when they are used to browse document collections or understand the trends and development within a particular research field. As a result, researchers have proposed topic coherence measures, which are a qualitative approach to automatically uncover the coherence of topics (Aletras and Stevenson, 2013; Stevens et al., 2012; Newman et al., 2010a; Röder et al., 2015). Topics are considered to be coherent if all or most of the words (e.g. a topic’s top-N words) are related. Topic coherence measures aim to find measures that correlate highly with human topic evaluation, such as topic ranking data obtained by, for example, word and topic intrusion tests (Chang et al., 2009). Human topic ranking data are often considered the gold standard and, consequently, a measure that correlates well is a good indicator for topic interpretability. A recent study by Röder et al. (2015) systematically and empirically explored the multitude of topic coherence measures and their correlation with available human topic ranking data; new coherence measures obtained by combining existing elementary elements were also examined. The researchers’ systematic approach revealed a new unexplored coherence measure, which they labeled $C_V$, to achieve the highest correlation with all available human topic ranking data. This study adopts the $C_V$ coherence measure for calculating topic coherence, with a detailed description of the calculations behind this measure described below.

The calculation of $C_V$ starts with the segmentation of the topic’s top-N words into pairs of word subsets, $S_i = (W', W^*)$, where $W' \in W$, $W^* \in W$, and $W$ consists of the topic’s top-N most probable words. More formally, a pair $S$ is defined as:

$$S = \{(W', W^*) | W = \{w_i\}; w_i \in W; W^* = W\} \quad (3.4)$$

For example, if $W = \{w_1, w_2, w_3\}$, then one pair might be $S_i = (W' = w_1), (W^* = w_1, w_2, w_3)$. Such segmentation measures the extent to which the subset $W^*$ supports or
The support between word subsets of a pair $S_i = (W', W^*)$ is calculated with a confirmation measure $\phi$. $C_V$ uses an indirect confirmation measure that considers not only the words within a pair, but also all words in $W$. A direct confirmation measure, such as difference, ratio, and likelihood measure, could place a low probability on high-support but low-frequency pairs. An indirect confirmation measure overcomes this by pairing every subset with $W$, thereby increasing the semantic support of supporting pairs. Word subsets are now represented as context vectors (Aletras and Stevenson, 2013), such as $\vec{v}(W')$ by pairing them to all words in $W$, as exemplified in (3.5). The relatedness between context vectors and words in $W$ is calculated by normalized pointwise mutual information (NPMI), as shown in (3.6).

$$v(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^{\gamma} \right\}_{j=1, \ldots, |W|} \tag{3.5}$$

$$\text{NPMI}(w_i, w_j)^{\gamma} = \left( \frac{\log \frac{p(w_i, w_j) + \varepsilon}{p(w_i)p(w_j)}}{-\log(P(w_i, w_j) + \varepsilon)} \right)^{\gamma} \tag{3.6}$$

In contrast to pointwise mutual information, NPMI achieves a higher correlation with human topic ranking data (Aletras and Stevenson, 2013), which is generally a result of reducing the impact of low-frequency counts in word co-occurrences (Bouma, 2009). Given our running example of $W = \{w_1, w_2, w_3\}$, we obtain the context vector for $w_1$ as $\vec{w}_1 = \{\text{NPMI}(w_1, w_1)^{\gamma}, \text{NPMI}(w_1, w_2)^{\gamma}, \text{NPMI}(w_1, w_3)^{\gamma}\}$, with the constant $\varepsilon$ to prevent logarithms of zero, and $\gamma$ to place more weight on higher NPMI values.

Probabilities of single words $p(w_i)$ or the joint probability of two words $p(w_i, w_j)$ can be estimated using a Boolean document calculation—that is, the number of documents in which $(w_i)$ or $(w_i, w_j)$ occurs, divided by the total number of documents. The Boolean document calculation, however, ignores the frequencies and distances of words. $C_V$ incorporates a Boolean sliding window calculation in which a new virtual document is created for every window of size $s$ when sliding over the document at a rate of one word token per step. For example, document $d_1$ with words $w$ results in virtual documents $d'_1 = \{w_1, \ldots, w_s\}$, $d'_2 = \{w_2, \ldots, w_{s+1}\}$, and so on. The probabilities $p(w_i)$ and $p(w_i, w_j)$ are subsequently calculated from the total number of virtual documents. In contrast to the Boolean document calculation, the Boolean sliding window calculation tries to capture the word token proximity to some degree.

The indirect confirmation measure $\phi_S(u, w)$ is obtained by calculating the cosine vector similarity between all context vectors $v(W') \in u$ and $v(W^*) \in w$ of a pair $S_i = (W', W^*)$, as shown in (3.7).
Finally, the arithmetic mean of individual confirmation measures is used to arrive at an overall topic coherence score.

3.3 Methods

3.3.1 Dataset

We compare the influence of Dirichlet hyperparameters on two datasets containing scientific research articles related to the domain of fisheries. The first dataset, $DS_1$, contains all full-text research articles published by the journal *Canadian Journal of Fisheries and Aquatic Sciences* and the journal *ICES Journal of Marine Science* from 1996 to 2016, with $D = 8,012$ documents, vocabulary size of $V = 203,248$, a total of $N = 29,469,919$ words, and on average 3,678 words per document. The second dataset, $DS_2$, contains only abstract data from the journal *Canadian Journal of Fisheries and Aquatic Sciences*, with $D = 4,417$, $V = 14,643$, $N = 481,168$, and 109 words on average per document. Both journals are domain-specific (i.e. fisheries) journals, but employ a wide scope of research directives related to the biological, ecological, and socio-ecological aspects of fisheries.

The domain of fisheries includes a multitude of knowledge production approaches, from mono- to transdisciplinary. Biologists, oceanographers, mathematicians, computer scientists, anthropologists, sociologists, political scientists, economists, and researchers from many other disciplines contribute to the body of knowledge of fisheries, together with non-academic participants such as decision makers and stakeholders. Within the domain of fisheries, research into text analytics techniques has only been applied in a number of cases (e.g. (Syed et al., 2016; Syed and Spruit, 2017)).

These journals were chosen for several reasons. First, a fisheries domain expert was available to rank the topics manually. Second, domain-specific journals, in contrast to generic journals such as *Nature*, *Science*, or *PLOS ONE*, increase generalizability to other domain-specific journals that are often the subject of study when uncovering topical structures from scientific publications, such as research performed within the field of computational linguistics (Hall et al., 2008) or neural information processing systems (NIPS) (Wang and McCallum, 2006), thereby making our results applicable to such approaches. Finally, the two journals have the highest frequency of publication output within the analyzed period compared to all other fisheries journals.

Words that were part of a standard list of stop words ($n = 153$), single-occurrence
Table 3.1: Notation of Dirichlet classes. $\alpha = \text{document-topic distribution}$, $\eta = \text{topic-word distribution}$

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>$\alpha$</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>AS</td>
<td>Asymmetric</td>
<td>Symmetric</td>
</tr>
<tr>
<td>SA</td>
<td>Symmetric</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>SS</td>
<td>Symmetric</td>
<td>Symmetric</td>
</tr>
</tbody>
</table>

words, and words occurring in $\geq 90\%$ of the documents (e.g., fish, analysis, research) were removed. The removal of the top 90\% of words serves as an estimate to prevent frequently occurring words from dominating all topics. All documents were tokenized and represented as bag-of-word features. Apart from grouping lowercase and uppercase words, no normalization method (e.g., stemming or lemmatization) was applied to reduce inflectional and derivational forms of words to a common base form. Stemming algorithms can be overly aggressive and could result in unrecognizable words that reduce interpretability when labeling the topics. Stemming might also lead to another problem: It cannot be deduced whether a stemmed word comes from a verb or a noun (Evangelopoulos et al., 2012). As human topic ranking was part of our topic quality evaluation, interpretability was considered to be highly important.

### 3.3.2 Dirichlet Hyperparameters

Hyperparameter $\alpha$ controls the shape of the document–topic distribution, whereas $\eta$ controls the shape of the topic–word distribution. A large $\alpha$ leads to documents containing many topics, and a large $\eta$ leads to topics with many words. In contrast, small values for $\alpha$ and $\eta$ result in sparse distributions: documents containing a small number of topics and topics with a small number of words. In essence, the hyperparameters $\alpha$ and $\eta$ have a smoothing effect on the multinomial variables $\theta$ and $\beta$, respectively. Four different classes or combinations of Dirichlet priors are explored, as listed in Table 3.1, in which we follow a similar notation (i.e. AA, AS, SA, SS) as described in (Wallach et al., 2009).

Symmetrical priors are often the default setting for LDA tools such as Mallet and Gensim and assume a priori that each of the $K$ topics has an equal probability of being assigned to a document while each word has an equal chance of being assigned to a topic. For the symmetrical prior $\alpha$, the hyperparameter is a vector with the value $1/K$, where $K$ is the number of topics. The symmetrical prior $\eta$ has a scalar parameter with the value $1/V$, where $V$ is the size of the vocabulary (full-text data $DS_1 = 203,248$, and abstract data $DS_2 = 14,643$). For the asymmetrical priors, we utilize an iterative learning process to approximate the hyperparameters from the data; estimation is re-
CHAPTER 3. EXPLORING DIRICHLET PRIORS

quired as no exact closed form solution exists. Estimating hyperparameters can be used to increase model quality, and their values can reveal specific properties of the corpus: $\alpha$ for the distinctiveness in underlying semantic structures and $\eta$ for the group size of commonly co-occurring words (Heinrich, 2005). Several methods for hyperparameter estimation exist, such as gradient ascent, fixed point iteration, and Newton-Raphson method. Estimating the Dirichlet parameter $\alpha$ aims to maximize $p(D|\alpha)$ by maximizing the log-likelihood function of the data $D$, with $\log \tilde{p}_k$ being the observed sufficient statistics (the following is analogous to that of $\eta$).

$$F(\alpha) = \log p(D|\alpha) = N \log \Gamma \left( \sum_k a_k \right) - N \sum_k \log \Gamma(a_k) + N \sum_k (a_k - 1) \log \tilde{p}_k$$

with $\log \tilde{p}_k = \frac{1}{N} \log p_{i,k}$

This study adopts the Newton-Rapson (Huang, 2005) method that provides a quadratic converging method for parameter estimation. Given an initial value for $\alpha$, parameters are iteratively updated to arrive at an asymmetrical Dirichlet distribution learned from the data. The update is given in (3.9), with $\nabla F$ being the gradient descent, iteratively stepping along a positive gradient to maximize or converge the log-likelihood function $F$ (3.8).

$$a_k^{new} = a_k^{old} - \frac{(\nabla F)_k - b}{q_{kk}}$$

$$\nabla F = \frac{\partial F}{\partial \alpha_k} = N \left( \Psi \left( \sum_k \alpha_k \right) - \Psi(a_k) + \log \tilde{p}_k \right)$$

$$b = \frac{\sum_j (\nabla F)_j / q_{jj}}{1/c + \sum_j 1/q_{jj}}$$

$$c = N \Psi' \left( \sum_k \alpha_k \right)$$

$$q_{jk} = -N \Psi'(\alpha_k)$$

3.3.3 Creating LDA Models

LDA models were created for four different classes of priors on $\alpha$ and $\eta$, as listed in Table 3.1. For each class of priors, LDA models were produced by varying the number of topics parameter $K = \{1, \ldots, 50\}$ and repeating the process five times; one class resulted in 250 LDA models. The same approach was performed on both datasets: $DS_1$ for full-text data and $DS_2$ for abstract data. A total of 2000 different LDA models were created.
Given that our datasets focus on fisheries only, making them homogeneous in nature, a small number of topics is expected—typically around 10 to 20 given the scope and aims of the selected journals.

The Python library Gensim (Rehurek and Sojka, 2010) was used to create LDA models. Posterior inference approximation is performed with online variational Bayes (VB) as proposed by Hoffman et al. (2010). Online VB is based on an online stochastic optimization process and produces similar or improved (Hoffman et al., 2010) and faster (Bottou and Bousquet, 2007) LDA models compared to its batch variant. The Newton-Raphson process of iteratively learning asymmetrical Dirichlet priors can conveniently be incorporated into online LDA in linear time. The convergence iteration parameter for the expectation step (i.e. E-step) is set to 100, where per-document parameters are fit for the variational distributions [see Algorithm 2 in (Hoffman et al., 2010)].

3.3.4 Topic Coherence

The coherence of topics was calculated using the $C_V$ coherence measure as described in detail in Section 3.2.3; $C_V$ has been shown to obtain the highest correlation with all available human topic ranking data. The segmentation of the topic’s top-$N$ words and subsequent calculation of confirmation are calculated for $N = 15$, pairing every top 15 word with every other top 15 word and calculating their semantic support within the corpus. $N = 15$ was chosen, in contrast to, for example, $N = 10$ (Aletras and Stevenson, 2013), as no stemming or lemmatization was applied; with $N = 10$, several words with the same base form were among the top 10 words (e.g. sample, sampling), so analyzing the top 10 words would effectively mean analyzing fewer than 10 distinct words. The constant $\varepsilon$ for NPMI calculations (see (3.6)) avoids logarithms of zero and acts as a smoothing factor. This value is set to a very small number, $10^{-12}$, as proposed by Stevens et. al. (Stevens et al., 2012); the coherence measure is highly dependent on the smoothing constant, and a very small value significantly reduces the scores for unrelated words compared to, for example, $\varepsilon = 1$ (Mimno et al., 2011). The $\gamma$ constant for NPMI calculations is set to 1 (see (3.5)) to place equal weights on all NPMI values. In contrast to $\gamma = 2$ (Aletras and Stevenson, 2013), $\gamma = 1$ produced a higher correlation with human topic ranking data (Röder et al., 2015). The sliding window $s$ for the Boolean sliding window calculation is set to 110 (Röder et al., 2015).

3.3.5 Human Topic Ranking

A fisheries domain expert manually ranked a selection of topics by inspecting the topic’s top 15 most probable words and a selection of document titles and content. The domain expert is affiliated with the leading competence institution for fishery and aquaculture in Norway. As topic coherence scores are also obtained from the topic’s top
15 words, the manual ranking of the top 15 words allows for equal comparison between the two proposed assessments. The domain expert was asked to provide a label for each topic that best captures the semantics of the top 15 words. In addition, the domain expert was asked to rank the topics concerning semantically correct or, conversely, incorrect words. An incorrect word could be a wrong fisheries domain-related word that does not match the topic label and, thus, does not fit with the semantics of the majority of right words. For example, in cases where most of the topic words resemble the fish species cod, an incorrect domain-related word might refer to a different kind of species. Furthermore, incorrect terms may refer to noise terms (i.e. words that serve a grammatical or syntactical purpose only). Topics are subsequently ranked by the number of right terms concerning all of the top 15 words. High-quality topics have $\geq 90\%$ correct words, medium-quality topics have $\geq 80\%$ but $< 90\%$ correct words, and low-quality topics have $< 80\%$ correct words.

### 3.3.6 Relaxing LDA assumptions

At the time of writing, the original LDA method proposed by Blei et al. (2003) has over 22,000 citations. The technique has received much attention from machine learning researchers and other scholars and has been adopted and extended in a variety of ways. More concretely, relaxing the assumptions behind LDA can result in richer representations of the underlying semantic structures. The bag of words assumption has been relaxed by conditioning words on the previous words (i.e. Markovian structure) (Wallach, 2006a); the document exchangeability assumption (i.e., the order in which documents are analyzed), relaxed by the dynamic topic model (Blei and Lafferty, 2006), and the Bayesian non-parametric model can be utilized to automatically uncover the number of topics (Whye Teh et al., 2004). Furthermore, LDA has been extended in various ways. Topics might correlate as a topic about "cars" is more likely to also be about "emission" than it is about "diseases". The Dirichlet distribution is implicitly independent and a more flexible distribution, such as the logistic normal, is a more appropriate distribution to capture covariance between topics. The correlated topic model aids in this task (Blei and Lafferty, 2007). Other examples extending LDA include the author-topic model (Rosen-Zvi et al., 2004), the relational topic model (Chang and Blei, 2010), the spherical topic model (Reisinger et al., 2010), the sparse topic model (Wang and Blei, 2009), and the bursty topic model (Doyle and Elkan, 2009). Topic models that relax or extend the original LDA model bring additional computational complexity and their own sets of limitations and challenges; nevertheless, it would be interesting to explore these models in future research.
3.4 Results

3.4.1 Topic Coherence

The coherence scores for the prior classes AA, AS, SA, and SS obtained from 8,012 full-text research articles (DS₁) are shown in Figure 3.2. Additionally, the coherence scores for prior classes obtained from 4,417 abstracts (DS₂) are shown in Figure 3.3. The coherence score represents the mean coherence score from all five runs for each value of $k$.

A visual inspection of Figs. 3.2a–3.2f (full-text data) shows that similar coherence scores are obtained for AA and AS (Figure 3.2a), with both sharing an asymmetrical prior over $\alpha$ but a different prior over $\eta$. Similar results are obtained when comparing SA and SS (Figure 3.2f), sharing a symmetrical prior over $\alpha$ and a different prior over $\eta$. Thus, varying $\eta$, while maintaining a similar prior over $\alpha$, shows no real difference in obtained coherence score. A slightly increased coherence is obtained for an asymmetrical prior over $\alpha$ (e.g., Figure 3.2d) for $k > 20$. Other combinations explored (e.g. AA–SA, AA–SS, and AS–SS) show similar results: a slight increase in coherence for an asymmetrical prior over $\alpha$, with $\eta$ showing no real benefits on topic coherence.

Figs. 3.3a–3.3f show coherence scores for LDA models obtained from abstract data (DS₂). AA–AS (Figure 3.3a) show that different priors over $\eta$, while maintaining the same asymmetrical prior over $\alpha$, result in similar coherence scores. Similarly, a symmetrical prior over $\alpha$ (Figure 3.3f) with different priors over $\eta$ shows no real differences in topic coherence. However, a large difference in coherence is obtained when varying the priors over $\alpha$ (Figure 3.3d), with an asymmetrical $\alpha$ showing improved coherence over a symmetrical $\alpha$. For DS₂, priors over $\alpha$, in contrast to results from DS₁, show higher coherence scores for all values of $k$. Moreover, varying priors over $\eta$ for DS₁ and DS₂ have a negligible effect on obtained coherence scores.

Table 3.2 shows the coherence score values obtained from DS₁ for $k = \{2, ..., 50\}$, with $\bar{X}$ representing the mean coherence over 5 runs, $s$ the standard deviation, and $f$ and $p$ the one-way ANOVA F-value and p-value, respectively. The last six columns show the post hoc significance thresholds for all six comparison of Dirichlet priors.

Table 3.2 reveals that significant differences are obtained starting from $k \geq 25$, although this does not hold for every $k \geq 25$. For $k < 25$, except for $k = 6$, no significant differences are obtained for combinations of priors; asymmetrical or symmetrical priors over $\alpha$ and $\eta$ have no significant effect on topic coherence. However, the coherence score values for $k < 25$ show slightly higher values (shown in bold) for a symmetrical prior over $\alpha$ compared to an asymmetrical prior. In contrast, for $k \geq 25$, an asymmetrical prior over $\alpha$ shows higher coherence values compared to a symmetrical prior. For all $k$, where $p$ is significant, SA–SS show no significance and AA–AS show significance only for $k = 6$ and $k = 47$; indicating the marginal importance of symmetrical
Figure 3.2: A comparison of calculated $C_V$ topic coherence scores for all classes of priors (i.e. AA, AS, SA, SS). Coherence scores represent mean scores from five runs for $K = \{1, \ldots, 50\}$. $DS_1 = 8,012$ full-text articles.
Figure 3.3: A comparison of calculated $C_V$ topic coherence scores for all classes of priors (i.e. AA, AS, SA, SS). Coherence scores represent mean scores from five runs for $K = \{1, \ldots, 50\}$. $DS_2 = 4,417$ abstracts.
or asymmetrical priors over $\eta$.

Table 3.3 shows the coherence score values and ANOVA statistics for $DS_2$. For all $k > 2$, the difference is significant ($p < 0.001$). These significant differences are caused by using an asymmetrical prior over $\alpha$ compared to a symmetrical prior. Where $DS_1$ shows mixed results between different priors over $\alpha$, for $DS_2$, every combination of asymmetrical priors over $\alpha$ outperforms symmetrical priors over $\alpha$. The post hoc tests for comparisons between different priors over $\eta$ are almost in all cases not significant, following a similar trend with $DS_1$. 
<table>
<thead>
<tr>
<th>K</th>
<th>$\bar{x}_{AA}$</th>
<th>$\bar{x}_{AS}$</th>
<th>$\bar{x}_{SA}$</th>
<th>$\bar{x}_{SS}$</th>
<th>$s_M$</th>
<th>$s_S$</th>
<th>$f$</th>
<th>$p$</th>
<th>$AA-AS$</th>
<th>$AA-SA$</th>
<th>$AA-SS$</th>
<th>$AS-SA$</th>
<th>$AS-SS$</th>
<th>$SA-SS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.580</td>
<td>0.583</td>
<td>0.516</td>
<td>0.571</td>
<td>0.037</td>
<td>0.021</td>
<td>0.055</td>
<td>0.004</td>
<td>3.237</td>
<td>0.0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.559</td>
<td>0.570</td>
<td>0.564</td>
<td>0.564</td>
<td>0.026</td>
<td>0.020</td>
<td>0.021</td>
<td>0.035</td>
<td>0.108</td>
<td>0.9543</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.543</td>
<td>0.547</td>
<td>0.565</td>
<td>0.572</td>
<td>0.014</td>
<td>0.016</td>
<td>0.020</td>
<td>0.020</td>
<td>2.554</td>
<td>0.0918</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.552</td>
<td>0.561</td>
<td>0.559</td>
<td>0.578</td>
<td>0.013</td>
<td>0.010</td>
<td>0.025</td>
<td>0.013</td>
<td>1.788</td>
<td>0.1899</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.539</td>
<td>0.562</td>
<td>0.556</td>
<td>0.570</td>
<td>0.014</td>
<td>0.011</td>
<td>0.021</td>
<td>0.007</td>
<td>3.298</td>
<td>0.0475*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.547</td>
<td>0.563</td>
<td>0.565</td>
<td>0.565</td>
<td>0.010</td>
<td>0.017</td>
<td>0.013</td>
<td>0.018</td>
<td>1.308</td>
<td>0.3063</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.555</td>
<td>0.569</td>
<td>0.569</td>
<td>0.571</td>
<td>0.025</td>
<td>0.019</td>
<td>0.018</td>
<td>0.018</td>
<td>0.571</td>
<td>0.6419</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.554</td>
<td>0.567</td>
<td>0.569</td>
<td>0.571</td>
<td>0.011</td>
<td>0.017</td>
<td>0.008</td>
<td>0.018</td>
<td>0.878</td>
<td>0.4732</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.562</td>
<td>0.567</td>
<td>0.569</td>
<td>0.571</td>
<td>0.012</td>
<td>0.009</td>
<td>0.022</td>
<td>0.016</td>
<td>0.242</td>
<td>0.8660</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.572</td>
<td>0.566</td>
<td>0.567</td>
<td>0.571</td>
<td>0.006</td>
<td>0.005</td>
<td>0.005</td>
<td>0.021</td>
<td>0.270</td>
<td>0.8462</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.566</td>
<td>0.571</td>
<td>0.571</td>
<td>0.574</td>
<td>0.014</td>
<td>0.012</td>
<td>0.013</td>
<td>0.005</td>
<td>0.334</td>
<td>0.8006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.565</td>
<td>0.570</td>
<td>0.577</td>
<td>0.573</td>
<td>0.014</td>
<td>0.017</td>
<td>0.007</td>
<td>0.010</td>
<td>0.670</td>
<td>0.5828</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.569</td>
<td>0.588</td>
<td>0.581</td>
<td>0.577</td>
<td>0.014</td>
<td>0.016</td>
<td>0.012</td>
<td>0.013</td>
<td>1.309</td>
<td>0.3061</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.577</td>
<td>0.579</td>
<td>0.597</td>
<td>0.587</td>
<td>0.011</td>
<td>0.017</td>
<td>0.014</td>
<td>0.011</td>
<td>1.816</td>
<td>0.1848</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.585</td>
<td>0.583</td>
<td>0.585</td>
<td>0.593</td>
<td>0.008</td>
<td>0.010</td>
<td>0.008</td>
<td>0.012</td>
<td>0.926</td>
<td>0.4508</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.583</td>
<td>0.587</td>
<td>0.581</td>
<td>0.581</td>
<td>0.015</td>
<td>0.009</td>
<td>0.005</td>
<td>0.006</td>
<td>0.359</td>
<td>0.7837</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.586</td>
<td>0.576</td>
<td>0.578</td>
<td>0.591</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.014</td>
<td>1.639</td>
<td>0.2200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.587</td>
<td>0.588</td>
<td>0.584</td>
<td>0.591</td>
<td>0.010</td>
<td>0.017</td>
<td>0.008</td>
<td>0.012</td>
<td>0.216</td>
<td>0.8840</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.580</td>
<td>0.587</td>
<td>0.586</td>
<td>0.586</td>
<td>0.004</td>
<td>0.012</td>
<td>0.005</td>
<td>0.010</td>
<td>0.659</td>
<td>0.5890</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.594</td>
<td>0.581</td>
<td>0.588</td>
<td>0.598</td>
<td>0.010</td>
<td>0.010</td>
<td>0.007</td>
<td>0.020</td>
<td>1.385</td>
<td>0.2834</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.589</td>
<td>0.595</td>
<td>0.579</td>
<td>0.583</td>
<td>0.004</td>
<td>0.011</td>
<td>0.012</td>
<td>0.009</td>
<td>2.377</td>
<td>0.1082</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.592</td>
<td>0.599</td>
<td>0.581</td>
<td>0.582</td>
<td>0.008</td>
<td>0.019</td>
<td>0.009</td>
<td>0.012</td>
<td>1.730</td>
<td>0.2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.610</td>
<td>0.595</td>
<td>0.583</td>
<td>0.584</td>
<td>0.009</td>
<td>0.011</td>
<td>0.011</td>
<td>0.007</td>
<td>6.739</td>
<td>0.0038**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.607</td>
<td>0.595</td>
<td>0.583</td>
<td>0.584</td>
<td>0.009</td>
<td>0.011</td>
<td>0.011</td>
<td>0.007</td>
<td>6.739</td>
<td>0.0038**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Coherence score values and one-way ANOVA test statistics for $DS_1$ for $K = \{2, \ldots, 50\}$. $^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$
### Table 3.2: continued.

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \bar{X}_{AA} )</th>
<th>( \bar{X}_{AS} )</th>
<th>( \bar{X}_{SA} )</th>
<th>( \bar{X}_{SS} )</th>
<th>( s_{AA} )</th>
<th>( s_{AS} )</th>
<th>( s_{SA} )</th>
<th>( s_{SS} )</th>
<th>( f )</th>
<th>( p )</th>
<th>( AA \rightarrow AS )</th>
<th>( AA \rightarrow SA )</th>
<th>( AA \rightarrow SS )</th>
<th>( AS \rightarrow SA )</th>
<th>( AS \rightarrow SS )</th>
<th>( SA \rightarrow SS )</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>0.604</td>
<td>0.599</td>
<td>0.578</td>
<td>0.592</td>
<td>0.008</td>
<td>0.015</td>
<td>0.014</td>
<td>0.013</td>
<td>3.162</td>
<td>0.0534</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.596</td>
<td><strong>0.603</strong></td>
<td>0.594</td>
<td>0.588</td>
<td>0.009</td>
<td>0.013</td>
<td>0.008</td>
<td>0.010</td>
<td>1.471</td>
<td>0.2601</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.595</td>
<td><strong>0.597</strong></td>
<td>0.586</td>
<td>0.592</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.015</td>
<td>0.812</td>
<td>0.5056</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.601</td>
<td><strong>0.601</strong></td>
<td>0.588</td>
<td>0.596</td>
<td>0.008</td>
<td>0.005</td>
<td>0.014</td>
<td>0.017</td>
<td>1.038</td>
<td>0.4024</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td><strong>0.605</strong></td>
<td>0.605</td>
<td>0.582</td>
<td>0.595</td>
<td>0.012</td>
<td>0.007</td>
<td>0.007</td>
<td>0.006</td>
<td>6.342</td>
<td>0.0049**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.606</td>
<td><strong>0.608</strong></td>
<td>0.581</td>
<td>0.596</td>
<td>0.007</td>
<td>0.011</td>
<td>0.004</td>
<td>0.013</td>
<td>7.172</td>
<td>0.0029**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.597</td>
<td><strong>0.604</strong></td>
<td>0.583</td>
<td>0.588</td>
<td>0.010</td>
<td>0.004</td>
<td>0.005</td>
<td>0.009</td>
<td>6.359</td>
<td>0.0048**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.606</td>
<td><strong>0.607</strong></td>
<td>0.592</td>
<td>0.603</td>
<td>0.006</td>
<td>0.013</td>
<td>0.010</td>
<td>0.017</td>
<td>1.237</td>
<td>0.3290</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.606</td>
<td><strong>0.611</strong></td>
<td>0.587</td>
<td>0.584</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.016</td>
<td>5.993</td>
<td>0.0061**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.605</td>
<td>0.603</td>
<td>0.594</td>
<td>0.594</td>
<td>0.014</td>
<td>0.010</td>
<td>0.005</td>
<td>0.011</td>
<td>1.269</td>
<td>0.3185</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.602</td>
<td>0.596</td>
<td>0.596</td>
<td>0.582</td>
<td>0.007</td>
<td>0.012</td>
<td>0.004</td>
<td>0.006</td>
<td>5.398</td>
<td>0.0093**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>0.604</td>
<td>0.601</td>
<td>0.589</td>
<td>0.596</td>
<td>0.005</td>
<td>0.011</td>
<td>0.010</td>
<td>0.013</td>
<td>1.864</td>
<td>0.1764</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.600</td>
<td>0.607</td>
<td>0.586</td>
<td>0.600</td>
<td>0.004</td>
<td>0.004</td>
<td>0.016</td>
<td>0.011</td>
<td>2.937</td>
<td>0.0650</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0.605</td>
<td>0.596</td>
<td>0.596</td>
<td>0.602</td>
<td>0.004</td>
<td>0.008</td>
<td>0.013</td>
<td>0.006</td>
<td>1.529</td>
<td>0.2453</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.612</td>
<td>0.610</td>
<td>0.595</td>
<td>0.596</td>
<td>0.009</td>
<td>0.013</td>
<td>0.009</td>
<td>0.010</td>
<td>3.066</td>
<td>0.0612</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.608</td>
<td>0.605</td>
<td>0.587</td>
<td>0.589</td>
<td>0.007</td>
<td>0.008</td>
<td>0.007</td>
<td>0.013</td>
<td>5.712</td>
<td>0.0074**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.605</td>
<td>0.604</td>
<td>0.591</td>
<td>0.596</td>
<td>0.005</td>
<td>0.010</td>
<td>0.015</td>
<td>0.007</td>
<td>1.605</td>
<td>0.2274</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>0.605</td>
<td><strong>0.605</strong></td>
<td>0.585</td>
<td>0.594</td>
<td>0.015</td>
<td>0.010</td>
<td>0.006</td>
<td>0.014</td>
<td>2.664</td>
<td>0.0831</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.613</td>
<td><strong>0.614</strong></td>
<td>0.594</td>
<td>0.600</td>
<td>0.009</td>
<td>0.010</td>
<td>0.007</td>
<td>0.011</td>
<td>4.600</td>
<td>0.0167*</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.611</td>
<td>0.608</td>
<td>0.583</td>
<td>0.595</td>
<td>0.011</td>
<td>0.009</td>
<td>0.009</td>
<td>0.017</td>
<td>4.402</td>
<td>0.0194*</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>0.608</td>
<td>0.602</td>
<td>0.589</td>
<td>0.599</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.007</td>
<td>3.502</td>
<td>0.0400*</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>0.606</td>
<td><strong>0.615</strong></td>
<td>0.594</td>
<td>0.598</td>
<td>0.004</td>
<td>0.006</td>
<td>0.011</td>
<td>0.006</td>
<td>6.280</td>
<td>0.0051**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0.606</td>
<td>0.605</td>
<td>0.588</td>
<td>0.597</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.008</td>
<td>5.384</td>
<td>0.0094**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>0.619</td>
<td>0.608</td>
<td>0.590</td>
<td>0.599</td>
<td>0.009</td>
<td>0.008</td>
<td>0.006</td>
<td>0.005</td>
<td>12.376</td>
<td>0.0002***</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.609</td>
<td><strong>0.611</strong></td>
<td>0.589</td>
<td>0.594</td>
<td>0.006</td>
<td>0.007</td>
<td>0.010</td>
<td>0.010</td>
<td>6.225</td>
<td>0.0053**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
</tr>
</tbody>
</table>
**Table 3.3**: Coherence score values and one-way ANOVA test statistics for $DS_2$ for $K = \{2, \ldots, 50\}$. $^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$

<table>
<thead>
<tr>
<th>$K$</th>
<th>$\bar{X}_{AA}$</th>
<th>$\bar{X}_{AS}$</th>
<th>$\bar{X}_{SA}$</th>
<th>$\bar{X}_{SS}$</th>
<th>$s_{AA}$</th>
<th>$s_{AS}$</th>
<th>$s_{SA}$</th>
<th>$s_{SS}$</th>
<th>$f$</th>
<th>$p$</th>
<th>$AA-AS$</th>
<th>$AA-SA$</th>
<th>$AA-SS$</th>
<th>$AS-SA$</th>
<th>$AS-SS$</th>
<th>$SA-SS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.399</td>
<td>0.411</td>
<td>0.400</td>
<td>0.380</td>
<td>0.017</td>
<td>0.011</td>
<td>0.021</td>
<td>0.037</td>
<td>1.207</td>
<td>0.3390</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.493</td>
<td>0.488</td>
<td>0.425</td>
<td>0.414</td>
<td>0.013</td>
<td>0.026</td>
<td>0.025</td>
<td>0.031</td>
<td>11.187</td>
<td>0.0003***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.504</td>
<td>0.524</td>
<td>0.431</td>
<td>0.458</td>
<td>0.010</td>
<td>0.023</td>
<td>0.025</td>
<td>0.018</td>
<td>18.234</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.523</td>
<td>0.531</td>
<td>0.446</td>
<td>0.464</td>
<td>0.032</td>
<td>0.015</td>
<td>0.030</td>
<td>0.024</td>
<td>10.511</td>
<td>0.0005***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.517</td>
<td>0.517</td>
<td>0.453</td>
<td>0.474</td>
<td>0.013</td>
<td>0.015</td>
<td>0.017</td>
<td>0.014</td>
<td>17.836</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.535</td>
<td>0.545</td>
<td>0.497</td>
<td>0.468</td>
<td>0.009</td>
<td>0.010</td>
<td>0.029</td>
<td>0.013</td>
<td>17.407</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.527</td>
<td>0.532</td>
<td>0.465</td>
<td>0.483</td>
<td>0.009</td>
<td>0.029</td>
<td>0.020</td>
<td>0.029</td>
<td>7.939</td>
<td>0.0018**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.551</td>
<td>0.549</td>
<td>0.477</td>
<td>0.482</td>
<td>0.008</td>
<td>0.016</td>
<td>0.025</td>
<td>0.022</td>
<td>18.741</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.569</td>
<td>0.541</td>
<td>0.453</td>
<td>0.485</td>
<td>0.003</td>
<td>0.018</td>
<td>0.012</td>
<td>0.012</td>
<td>72.727</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.546</td>
<td>0.534</td>
<td>0.466</td>
<td>0.489</td>
<td>0.019</td>
<td>0.017</td>
<td>0.013</td>
<td>0.026</td>
<td>14.533</td>
<td>0.0001***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.556</td>
<td>0.548</td>
<td>0.462</td>
<td>0.486</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.018</td>
<td>14.272</td>
<td>0.0001***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.547</td>
<td>0.558</td>
<td>0.466</td>
<td>0.483</td>
<td>0.027</td>
<td>0.022</td>
<td>0.021</td>
<td>0.031</td>
<td>12.797</td>
<td>0.0002***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.536</td>
<td>0.570</td>
<td>0.487</td>
<td>0.484</td>
<td>0.011</td>
<td>0.016</td>
<td>0.008</td>
<td>0.019</td>
<td>33.262</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.560</td>
<td>0.562</td>
<td>0.476</td>
<td>0.492</td>
<td>0.017</td>
<td>0.014</td>
<td>0.022</td>
<td>0.015</td>
<td>26.740</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.561</td>
<td>0.568</td>
<td>0.499</td>
<td>0.481</td>
<td>0.016</td>
<td>0.014</td>
<td>0.023</td>
<td>0.023</td>
<td>19.807</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.554</td>
<td>0.577</td>
<td>0.475</td>
<td>0.494</td>
<td>0.015</td>
<td>0.011</td>
<td>0.020</td>
<td>0.013</td>
<td>41.061</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.570</td>
<td>0.573</td>
<td>0.463</td>
<td>0.489</td>
<td>0.018</td>
<td>0.016</td>
<td>0.019</td>
<td>0.020</td>
<td>38.081</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.558</td>
<td>0.559</td>
<td>0.487</td>
<td>0.483</td>
<td>0.005</td>
<td>0.011</td>
<td>0.022</td>
<td>0.016</td>
<td>33.195</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.566</td>
<td>0.549</td>
<td>0.484</td>
<td>0.486</td>
<td>0.009</td>
<td>0.017</td>
<td>0.014</td>
<td>0.022</td>
<td>27.914</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.551</td>
<td>0.554</td>
<td>0.494</td>
<td>0.477</td>
<td>0.016</td>
<td>0.018</td>
<td>0.021</td>
<td>0.014</td>
<td>19.513</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.563</td>
<td>0.553</td>
<td>0.484</td>
<td>0.495</td>
<td>0.024</td>
<td>0.015</td>
<td>0.022</td>
<td>0.012</td>
<td>17.750</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.543</td>
<td>0.553</td>
<td>0.471</td>
<td>0.499</td>
<td>0.021</td>
<td>0.021</td>
<td>0.013</td>
<td>0.019</td>
<td>16.624</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.562</td>
<td>0.560</td>
<td>0.495</td>
<td>0.494</td>
<td>0.022</td>
<td>0.014</td>
<td>0.012</td>
<td>0.025</td>
<td>15.922</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.549</td>
<td>0.559</td>
<td>0.491</td>
<td>0.481</td>
<td>0.011</td>
<td>0.009</td>
<td>0.022</td>
<td>0.007</td>
<td>33.981</td>
<td>0.0000***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3: continued.

| \( K \) | \( \bar{X}_{AA} \) | \( \bar{X}_{AS} \) | \( \bar{X}_{SA} \) | \( \bar{X}_{SS} \) | \( s_{AA} \) | \( s_{AS} \) | \( s_{SA} \) | \( s_{SS} \) | \( f \) | \( p \) | \( AA - AS \) | \( AA - SA \) | \( AA - SS \) | \( AS - SA \) | \( AS - SS \) | \( SA - SS \) |
|--------|----------------|----------------|----------------|----------------|----------|----------|----------|----------|---------|------|------------|------------|------------|------------|------------|------------|------------|
| 26     | 0.523          | 0.558          | 0.492          | 0.494          | 0.016    | 0.011    | 0.019    | 0.005    | 19.576  | 0.0000*** | **     | **         | **         | **         | ***        | ***        | ***        |
| 27     | 0.552          | 0.551          | 0.483          | 0.496          | 0.014    | 0.013    | 0.013    | 0.016    | 27.074  | 0.0000*** | ***     | ***        | ***        | ***        | ***        | ***        | ***        |
| 28     | 0.552          | 0.554          | 0.487          | 0.474          | 0.022    | 0.010    | 0.019    | 0.010    | 27.501  | 0.0000*** | **     | ***        | ***        | ***        | ***        | ***        | ***        |
| 29     | 0.538          | 0.552          | 0.480          | 0.483          | 0.015    | 0.013    | 0.021    | 0.010    | 24.880  | 0.0000*** | **     | ***        | ***        | ***        | ***        | ***        | ***        |
| 30     | 0.537          | 0.551          | 0.469          | 0.480          | 0.013    | 0.009    | 0.015    | 0.019    | 30.982  | 0.0000*** | **     | ***        | ***        | ***        | ***        | ***        | ***        |
| 31     | 0.559          | 0.547          | 0.485          | 0.486          | 0.022    | 0.018    | 0.016    | 0.021    | 16.533  | 0.0000*** | ***     | **         | ***        | ***        | **         | ***        | ***        |
| 32     | 0.551          | 0.557          | 0.474          | 0.491          | 0.017    | 0.021    | 0.009    | 0.013    | 28.414  | 0.0000*** | ***     | **         | ***        | ***        | ***        | ***        | ***        |
| 33     | 0.538          | 0.544          | 0.475          | 0.501          | 0.009    | 0.007    | 0.013    | 0.013    | 36.003  | 0.0000*** | **     | **         | ***        | ***        | **         | ***        | ***        |
| 34     | 0.539          | 0.547          | 0.484          | 0.480          | 0.014    | 0.006    | 0.019    | 0.018    | 22.547  | 0.0000*** | **     | **         | ***        | ***        | **         | ***        | ***        |
| 35     | 0.552          | 0.538          | 0.463          | 0.488          | 0.013    | 0.011    | 0.012    | 0.015    | 43.139  | 0.0000*** | ***     | **         | ***        | ***        | **         | ***        | ***        |
| 36     | 0.541          | 0.540          | 0.489          | 0.488          | 0.011    | 0.019    | 0.016    | 0.012    | 15.847  | 0.0000*** | ***     | ***        | **         | **         | **         | **         | **         |
| 37     | 0.552          | 0.553          | 0.477          | 0.489          | 0.023    | 0.007    | 0.006    | 0.007    | 39.560  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 38     | 0.545          | 0.548          | 0.485          | 0.473          | 0.014    | 0.011    | 0.008    | 0.017    | 36.334  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 39     | 0.530          | 0.548          | 0.487          | 0.494          | 0.014    | 0.005    | 0.011    | 0.008    | 32.693  | 0.0000*** | ***     | ***        | **         | **         | **         | **         | **         |
| 40     | 0.543          | 0.539          | 0.488          | 0.483          | 0.017    | 0.023    | 0.010    | 0.013    | 15.567  | 0.0001*** | ***     | **         | **         | **         | **         | **         | **         |
| 41     | 0.546          | 0.535          | 0.461          | 0.478          | 0.012    | 0.009    | 0.011    | 0.015    | 48.701  | 0.0000*** | ***     | ***        | **         | **         | **         | **         | **         |
| 42     | 0.520          | 0.543          | 0.476          | 0.488          | 0.021    | 0.019    | 0.018    | 0.017    | 10.426  | 0.0005*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 43     | 0.542          | 0.532          | 0.470          | 0.483          | 0.015    | 0.015    | 0.015    | 0.013    | 23.777  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 44     | 0.528          | 0.533          | 0.482          | 0.483          | 0.008    | 0.016    | 0.010    | 0.011    | 22.455  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 45     | 0.536          | 0.544          | 0.473          | 0.476          | 0.006    | 0.008    | 0.012    | 0.010    | 66.934  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 46     | 0.528          | 0.525          | 0.470          | 0.483          | 0.014    | 0.019    | 0.007    | 0.016    | 16.132  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 47     | 0.527          | 0.545          | 0.472          | 0.478          | 0.005    | 0.016    | 0.018    | 0.009    | 29.467  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 48     | 0.536          | 0.540          | 0.475          | 0.476          | 0.009    | 0.005    | 0.017    | 0.010    | 41.643  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 49     | 0.538          | 0.527          | 0.476          | 0.487          | 0.014    | 0.015    | 0.009    | 0.022    | 14.562  | 0.0001*** | ***     | **         | ***        | ***        | **         | **         | **         |
| 50     | 0.525          | 0.535          | 0.470          | 0.484          | 0.018    | 0.013    | 0.008    | 0.014    | 21.727  | 0.0000*** | ***     | **         | ***        | ***        | **         | **         | **         |
Table 3.4: Human topic ranking for $D_{S_2}$ (abstract) on $k = 17$ LDA model

<table>
<thead>
<tr>
<th>Class</th>
<th>High-quality</th>
<th>Medium-quality</th>
<th>Low-quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$15/17$ (88%)</td>
<td>$2/17$ (12%)</td>
<td>$0/17$ (0%)</td>
</tr>
<tr>
<td>AS</td>
<td>$15/17$ (88%)</td>
<td>$2/17$ (12%)</td>
<td>$0/17$ (0%)</td>
</tr>
<tr>
<td>SA</td>
<td>$12/17$ (70.5%)</td>
<td>$4/17$ (23.6%)</td>
<td>$1/17$ (5.9%)</td>
</tr>
<tr>
<td>SS</td>
<td>$11/17$ (64.7%)</td>
<td>$6/17$ (35.3%)</td>
<td>$0/17$ (0%)</td>
</tr>
</tbody>
</table>

3.4.2 Human Topic Ranking

The results of the fisheries domain expert’s human topic ranking are shown in Table 3.4 (see Section 3.3.5 for the classification method). Human topic ranking was performed on $D_{S_2}$ for $k = 17$ LDA models, which is the $k$-value that shows the best coherence score (via elbow method) and, simultaneously, the $k$-value with the largest difference amongst all prior classes (ANOVA $f = 41.06$). A similar pattern as found for topic coherence scores can be identified (Figs. 3.3a–3.3f): AA and AS with an asymmetrical prior over $\alpha$ result in more high-quality (88%) topics compared to SA and SS with a symmetrical prior over $\alpha$ (70.5% and 64.7% high-quality topics). Both AA and AS perform similarly, indicating that priors over $\eta$ have no effect on human topic ranking. Furthermore, SA and SS show similar lower human topic ranking, with three topics differently classified: SS has 77.5% of high-quality topics compared to 64.7% for SS, but simultaneously one low-quality topic. A two-dimensional inter-topic distance map for $D_{S_2}$ with $k = 17$ is displayed in Figure 3.4 for all classes of priors. This two-dimensional representation is obtained by computing the distance between topics (Chuang et al., 2012) and applying multidimensional scaling (Sievert and Shirley, 2014). It displays the similarity between topics concerning their probability distribution over words (i.e. $\beta$). In addition, a topic label that best captures the semantics of the top 15 words is attached.

We omitted human topic ranking results for $D_{S_1}$ as they show an equal number of high-quality and medium-quality topics for all classes of priors and for several arbitrarily chosen $k$-values ($k < 25$). These results are in line with topic coherence scores that show similar scores for all prior classes (see Figs. 3.2a–3.2f). An inspection of $k \geq 25$ LDA models (the point where significant differences between prior classes start) shows an increasing number of incorrect terms for LDA models with a symmetrical prior over $\alpha$ (SA and SS), compared to models with an asymmetrical prior over $\alpha$. 

67
Figure 3.4: A two-dimensional inter-topic distance map (via multidimensional scaling) for all classes of priors for $DS_2$ with $k = 17$. The surface of the node indicates the overall topic prevalence within the corpus. Color coding is used to indicate human topic ranking classification: white = high-quality, grey = medium-quality, and black = low-quality.
3.5 Discussion and Conclusion

Our results show that an asymmetrical prior over $\alpha$ indicates increased topic coherence and topic ranking compared to a symmetrical prior. However, this particularly holds for the $DS_2$ dataset, the collection of 4,417 abstracts, and not necessarily for the $DS_1$ dataset, the collection of 8,012 full-text documents. Thus, selecting a different prior on $\alpha$ has large practical implications for datasets containing a smaller vocabulary size and being homogenous in nature. Symmetrical or asymmetrical priors over $\eta$ show no real benefits regarding topic coherence and human topic ranking.

The results on $DS_2$ are in line with research performed by Wallach et. al. (Wallach et al., 2009), which found that an asymmetrical prior over $\alpha$ shows improved likelihood of held-out data and that different priors over $\eta$ show no real differences; the vocabulary size of the $DS_2$ data set can be compared to the vocabulary size used in their research. However, our results on $DS_1$, the full-text dataset with significantly higher vocabulary size and an average number of words per document, found no difference for combinations of priors over $\alpha$ and $\eta$, making topic coherence and manual topic ranking less influenced by full-text data.

A symmetrical prior over $\alpha$ assumes that all topics have an equal probability of being assigned to a document. Such an assumption ignores that certain topics are more prominent in a document collection and, consequently, would logically have a higher probability to be assigned to a document. Conversely, specific topics are less common and, thus, not appropriately reflected with a symmetrical prior distribution. Logically speaking, an asymmetrical prior over $\alpha$ would capture this intuition and would, therefore, be the preferred choice. We have empirically shown that this intuition indeed results in significantly higher topic coherence and a better topic ranking for $DS_2$ and $DS_1$ for $k \geq 25$. For $DS_1$ with $k < 25$, the differences are not significant, although human topic ranking shows slightly better topics for the classes with an asymmetrical prior over $\alpha$.

Concerning priors over $\eta$, we naturally want topic–word distributions to be different from each other so as to avoid conflicts between them. A symmetrical prior over $\eta$ will reflect the power-law usage of words (i.e. some words occur in all topics) while simultaneously resolving ambiguity between topics with a few distinct word co-occurrences (Wallach et al., 2009). Therefore, symmetrical priors over $\eta$ are the preferred choice. Although our empirical results indicate no real benefits when varying priors on $\eta$, the symmetrical prior shows slight, but still very marginal, overall improved coherence and ranking results for both datasets.
Table 3.5: A selection of modeling topics for $k = 17$. Terms in bold are considered incorrect words.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Class</th>
<th>Label</th>
<th>Top 15 words</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DS_1$</td>
<td>AA</td>
<td>Models (population)</td>
<td>model, cod, stock, mortality, population, recruitment, year, models, estimates, fisheries, abundance, biomass, size, spawning, years</td>
<td>High</td>
</tr>
<tr>
<td>$DS_1$</td>
<td>AS</td>
<td>Models (abundance)</td>
<td>model, length, estimates, uncertainty, abundance, models, values, distribution, gaussian, survey, simulated, simulation, acoustic, parameters, pollock</td>
<td>High</td>
</tr>
<tr>
<td>$DS_1$</td>
<td>SA</td>
<td>Models</td>
<td>model, stock, fishing, catch, fisheries, recruitment, fishery, year, estimates, models, mortality, management, assessment, biomass, effort</td>
<td>High</td>
</tr>
<tr>
<td>$DS_1$</td>
<td>SS</td>
<td>Models (fishing)</td>
<td>model, fishing, catch, stock, fishery, fisheries, mortality, effort, management, estimates, year, models, parameters, population, assessment</td>
<td>High</td>
</tr>
<tr>
<td>$DS_2$</td>
<td>AA</td>
<td>Models</td>
<td>model, data, models, <strong>used</strong>, estimates, <strong>using</strong>, stock, management, fish, recruitment, mortality, population, method, approach, <strong>based</strong></td>
<td>Medium</td>
</tr>
<tr>
<td>$DS_2$</td>
<td>AS</td>
<td>Models</td>
<td>data, model, models, <strong>used</strong>, <strong>using</strong>, fish, estimates, method, approach, <strong>based</strong>, sampling, analysis, based, methods, method, estimate, <strong>use</strong></td>
<td>Medium</td>
</tr>
<tr>
<td>$DS_2$</td>
<td>SA</td>
<td>Models</td>
<td>model, data, models, management, <strong>used</strong>, species, estimates, <strong>using</strong>, fish, fisheries, abundance, stock, habitat, analysis, recruitment</td>
<td>Medium</td>
</tr>
<tr>
<td>$DS_2$</td>
<td>SS</td>
<td>Models (growth)</td>
<td>model, growth, data, estimates, fish, <strong>using</strong>, <strong>used</strong>, models, habitat, size, mortality, parameters, length, method, population</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Human topic ranking was based on the presence of incorrect terms being part of the topic’s 15 most probable words. A closer look into the reasons why topics are ranked lower reveals that all topics contain correct domain-related terms, but are only ranked lower due to the presence of so-called noise terms (e.g. used, using, two, among, total, higher, within, great, large, high, significantly). Lower-ranked topics contain a higher number of such terms and, as such, are classified as medium- or low-quality topics. Interestingly, none of the topics have incorrect domain-related terms that could refer to, for example, the biological, ecological, socio-ecological, or social aspects of fisheries. A selection of topics and incorrect terms is shown in Table 3.5. Topics uncovered from abstract data, combined with a symmetrical prior over $\alpha$, are more prone to contain such noise terms. For full-text data, all classes of priors show an equal but low number of noise words.

A growing amount of research is utilizing LDA to uncover latent semantic structures from scientific research articles as a mean to discover topical trends and developments within a particular research area (Gatti et al., 2015; Sun and Yin, 2017; Westgate et al., 2015; Wang and McCallum, 2006; Alston and Pardey, 2016). These approaches are often characterized by (i) exploring one or several domain-specific journals (e.g. journals related to transportation research, operations research and management science), (ii) using abstract data, and (iii) using an open source tool (e.g. Mallet, Gensim) to perform LDA. Our approach touches upon all three characteristics; thus, we would recommend an asymmetrical prior over $\alpha$ and a symmetrical prior over $\eta$ for optimal topic coherence and topic ranking.

Figure 3.4 shows a visual representation of 17 latent topics for $DS_2$. We identify several overlapping topics (e.g. otoliths, population genetics) and several semantically related topics (e.g. population dynamics and population genetics; lake nutrients and lake waters). At the same time, we find several topics occurring in one of the prior classes that are absent in other prior classes. For instance, fish diseases and parasites occurs only in AA (Figure 3.4a). One reason might be that manual topic labeling is limited by the subjectivity inherent in human interpretation (Urquhart, 2001); indeed, an analysis of the topics by another domain expert could yield contradictory results. Another reason might be due to the probabilistic nature of LDA, where differences are merely a result of differences in sampling. Although such analysis is outside the scope of this research, it is an interesting directive for future research. Furthermore, the research performed by Wallach et. al. was applied on corpora related to patent, newsgroup and news data, whereas this paper analyzed scientific research articles. Future research might focus on different types of scientific articles, more broadly oriented journals, or other unexplored forms of textual data to gain more insights into the practical effects Dirichlet priors have on LDA’s latent topics.
Chapter 4

Bootstrapping a Semantic Lexicon on Verb Similarities

We present a bootstrapping algorithm to create a semantic lexicon from a list of seed words and a corpus that was mined from the web. We exploit extraction patterns to bootstrap the lexicon and use collocation statistics to dynamically score new lexicon entries. Extraction patterns are subsequently scored by calculating the conditional probability in relation to a non-related text corpus. We find that verbs that are highly domain related achieved the highest accuracy and collocation statistics affect the accuracy positively and negatively during the bootstrapping runs.

This work was originally published as:

CHAPTER 4. BOOTSTRAPPING A SEMANTIC LEXICON

4.1 Introduction

Semantic lexicons for specific domains are increasingly becoming important in many Natural Language Processing (NLP) tasks such as information extraction, word sense disambiguation, anaphora resolution, speech recognition, and discourse processing. Online lexical resources such as WordNet (Miller et al., 1990) and Cyc (Lenat, 1995) are useful for generic domains but often fall short for domains that include specific terms, jargon, acronyms and other lexical variations. Handcrafting domain-specific lexicons can be time-consuming and costly. Various techniques have been developed to automatically create semantic lexicons, such as lexicon induction, lexicon learning, lexicon bootstrapping, lexical acquisition, hyponym learning, and web-based information extraction.

We define a semantic lexicon as a dictionary of hyponym words that share the same hypernym. For example, cat, dog or cow are all hyponym words that share the hypernym ANIMAL. Likewise, the words red, green, and blue are semantically related by the hypernym COLOR. A semantic lexicon differs from an ontology or taxonomy as it does not describe the formal representation of shared conceptualizations or information about concepts and their instances, nor does it provide a strict hierarchy of classes.

Several attempts have been made to automatically create semantic lexicons from text corpora by utilizing semantic relations in conjunctions (dogs and cats and cows), lists (dogs, cats, cows), appositives (labrador retriever, a dog) and compound nouns (dairy cow) (Riloff and Shepherd, 1997; Roark and Charniak, 1998; Phillips and Riloff, 2002; Widdows and Dorow, 2002). Others have used extraction patterns (Colombia was divided, the country was divided) (Thelen and Riloff, 2002; Igo and Riloff, 2009), instance/concept (“is-a”) relationships (Pantel and Ravichandran, 2004), coordination patterns (Ziering et al., 2013b), multilingual symbiosis (Ziering et al., 2013a), or combinations thereof (Qadir and Riloff, 2012). Lexical learning from informal text, such as social media, has also been performed (Qadir et al., 2015).

Learning semantic lexicons is often based on existing corpora that may not be available for all domains. Furthermore, little research on lexicon learning has been performed on web text from informative sites, forums, blogs, and comment sections that contain content written by a variety of people, thus containing different writing idiosyncrasies. We incorporate web mining techniques to build our own domain corpus and combine the BASILISK (Thelen and Riloff, 2002) bootstrapping algorithm and the Pointwise Mutual Information (PMI) scoring metric proposed by (Igo and Riloff, 2009). We based the hyponym relationships on the extraction pattern context of verb stem similarities and used a probability scoring metric to extract the most suitable verbs. Our aim is to use existing web content to create a semantic lexicon and subsequently use extraction patterns to find semantically related words. Extraction patterns, and more specifically the verbs within these patterns, are scored against a non-related text corpus to explore if verbs that occur more frequently in domain text are more likely to be accompanied by semantically related nouns.
The remainder of this paper is structured as follows: First, we present previous work in semantic lexicon learning for hypernym-hyponym relationships. Second, we detail our bootstrapping algorithm, the extraction of the text corpus, the creation of extraction patterns, and the scoring method. Third, we analyze the semantic lexicon and the accuracy of our algorithm during each successive run of the bootstrapping process. We then conclude our work and provide new research directives.

4.2 Previous Work

Riloff and Shepherd (1997) used noun co-occurrence statistics to bootstrap a semantic lexicon from raw data. Their bootstrapping algorithm scored conjunctions, lists, appositives and nominal compounds to find category words within a context window. It is one of the first attempts to build a semantic lexicon and uses human intervention to review the words and select the best ones for the final dictionary. Roark and Charniak (1998) applied a similar technique but use a different definition for noun co-occurrence and scoring candidate words. Riloff and Shepherd (1997) ranked and selected candidate words based on the ratio of noun co-occurrences in the seed list to the total frequency of the noun in the corpus while Roark and Charniak (1998) used log-likelihood statistics (Dunning, 1993) for final ranking.

Widdows and Dorow (2002) used graph models of the British National Corpus for lexical acquisition. They focused on the relationships between nouns when they occurred as part of a list. The nodes represent nouns and are linked to each other when they conjunct with either and or or. The edges are weighted by the frequency of the co-occurrence. Their algorithm mitigates the infection of bad words entering the candidate word list by looking at type frequency rather than token frequency.

Phillips and Riloff (2002) automatically created a semantic lexicon by looking at strong syntactic heuristics. They distinguish between two types of lexicons, (1) proper noun phrase and (2) common nouns, by utilizing the syntactic relationships between them. Syntactic structures are defined by appositives, compound nouns, and "is-a" clauses—identity clauses with a main verb of to be. Statistical filtering is applied to avoid inaccurate syntactic structures and deterioration of the lexicon entries. Pantel and Ravichandran (2004) also looked at "is-a" relationships by looking at concept signatures and applying a top-down approach. Their method first uses co-occurrence statistics for semantic classes to then find the most appropriate hyponym relationship.

Thelen and Riloff (2002) proposed a weakly supervised bootstrapping algorithm, called BASILISK, to learn semantic lexicons by looking at extraction patterns. They used the AutoSlog (Riloff, 1996) extraction pattern learner and a list of manually selected seed words to build semantic lexicons for the MUC-4 proceedings, a terrorism domain related corpus. Extraction patterns capture role relationships and are used to find noun phrases with similar semantic meaning as a result of syntax and lexical semantics. This
is explained by their example verb *robbed*. The subject of the verb robbed often indicates the perpetrator while the direct object of the verb robbed could indicate the victim or target. To avoid inaccurate words entering the lexicon and to increase the overall accuracy they learned multiple categories simultaneously. (Igo and Riloff, 2009) used BASILISK and applied co-occurrence statistics for hypernym-en seed word collocation by computing a variation of Pointwise Mutual Information (PMI). They used web statistics to re-rank the words after the bootstrapping process was finished.

Qadir and Riloff (2012) used pattern-based dictionary induction, contextual semantic tagging, and coreference resolution in an ensemble approach. They combined the three techniques in a single bootstrapping process and added lexicon entries if words occur in at least two of the three methods. Since each of them exploit independent sources, their ensemble method improves precision in the early stages of the bootstrapping process, in which semantic drift (Curran et al., 2007) can decrease the overall accuracy substantially. Ziering et al. (2013a) have also employed an ensemble method and used linguistic variations between multiple languages to reduce semantic drift.

We combine the BASILISK bootstrapping algorithm (Thelen and Riloff, 2002) and co-occurrence statistics proposed by Igo and Riloff (2009) to dynamically score each word before it enters the lexicon. We compare BASILISK’s scoring metric $\text{AvgLog}$ in contrast to the PMI metric at each bootstrap run rather than using PMI scores to re-rank the lexicon after the bootstrapping process has finished.

### 4.3 Lexicon Bootstrapping

We use a list of seed words to initiate the bootstrapping process and use this set of seed words to create a highly related text corpus by mining the web. Web mining was applied because domain specific corpora are not always readily available. Before bootstrapping begins, we use the linguistic expressions of extraction patterns (Riloff, 1996) and group noun phrases when they occur with the same stemmed verb. The stemmed verbs are then scored by a probability score with respect to a non-related text corpus. Verbs that are domain specific, that is, they infrequently occur in general text are given a higher score. We then use BASILISK’s bootstrapping algorithm and a PMI scoring metric before adding new words to the lexicon. The PMI score is a measure of association between words. For this, we use web count statistics between hyponym and hypernym words to calculate collocation statistics and utilize these scores to only allow the strongest hyponyms to enter the lexicon, thus decreasing wrong lexicon entries during the bootstrapping process.

A schematic overview of the bootstrapping process is displayed in Figure 4.1 and Algorithm 1. We will discuss the bootstrapping process in detail in subsequent sections.
CHAPTER 4. BOOTSTRAPPING A SEMANTIC LEXICON

Figure 4.1: Schematic overview of bootstrapping process.

Algorithm 1 Bootstrap Lexicon

1: lexicon ← seedwords
2: corpus ← TopNSearches(seedwords)
3: patterns ← Patterns(corpus, 0.5 ≤ p ≤ 1.0)
4: for $i = 0$ to $i < m$ do
5:     patternpool ← Score(patterns, topN + i)
6:     words ← GetNouns(patternpool)
7:     lexicon ← ScoreWord(words, topN) $\notin$ lexicon
8:     $i++$
9: end for
10: return lexicon
4.3.1 Domain and Seed Words

We choose to build a single related lexicon for the fisheries domain. The fisheries domain includes a multitude of knowledge production approaches, from mono- to transdisciplinary. Biologists, oceanographers, mathematicians, computer scientists, anthropologists, sociologists, political scientists, economists and researchers from many other more disciplines contribute to the fisheries body of knowledge, together with non-academic participants, such as decision makers and stakeholders. Due to these diverse contributions of specialized language from a multitude of knowledge production approaches, the fisheries domain is characterized by a large body of words. For the same reason, this body of words is extremely rich in concepts. However, sometimes the concepts use different words to refer to the same abstraction (e.g. fishermen and fishers) and sometimes, even though they are using the same words, they are referring to different abstractions (e.g. fisher behavior may mean something for an economist and something else for an anthropologist). In addition, the fisheries domain has a high frequency of compound words (e.g. fisheries management, fishing method), in order to differentiate it from other resource management domains, such as forestry, for example. The definition of a hyponym-hypernym relation is therefore based on a more abstract level and also includes transitive relationships. For example, if \( x \) is a hyponym of \( y \), and \( y \) is a hyponym of \( z \), then \( x \) is a hyponym of \( z \).

The seed words were chosen by experts from the Arctic University of Norway where they were asked for a list of 10 nouns or compound nouns that best cover the domain. The list contains the phrases fishery ecosystem, fisheries management, fisheries policy, fishing methods, fishing gear, fishing area, fish, fish species, fishermen, fish supply chain.

4.3.2 Building the Corpus

The fisheries corpus was created by extracting web text from a Google search for each of the seed words and extracting the content of the first 30 URLs. We enclosed the 10 search terms with quotations marks to force Google into an exact match. We found that the first 30 searches provided sufficient text for the corpus and still contained search related content. After mining the pages we started an extensive data cleaning process. Besides cleaning HTML, JavaScript, and CSS tags, we needed to clean text from e.g. headers, footers, labels, sidebars that entered the corpus. We removed content such as "click here", "all rights reserved", "copyright". Finally, we scored the mined text fragments with an en-US and en-GB lookup dictionary \( D \) as:

\[
S_{text_i} = \frac{\sum_{i=1}^{V} w_i}{V_i}
\]  

(4.1)
Figure 4.2: The lexical dispersion plot shows how often a word, displayed on the y-axis, occurs in the cleaned corpus. The x-axis displays the position within the corpus from beginning to end.

Where $V_i$ is the vocabulary size of $text_i$, and $w_i$ a word match such that $w_i \in D$. We found $S_{text_i} \geq 0.3$ sufficient to clean the corpus even further as it removes non-English and improper formatted text. For example, phrases like "Not logged inTalkContributionCreate accountLog in" bear no meaning and serves navigational purposes only when rendered by the web browser, yet it is extracted from the HTML content. Figure 4.2 shows the dispersion plot for the seed words unigrams after the cleaning phase.

4.3.3 Chunking

We used the Punkt sentence tokenizer and a part-of-speech (POS) tagger to create individual sentences with their grammatical properties. We then used a shallow parser with regular expressions (RE) on POS tags to extract noun phrases (NP) when they occur as a direct object or subject. The verb that precedes or follows the NP is used to group related NPs when they share the same stem.
CHAPTER 4. BOOTSTRAPPING A SEMANTIC LEXICON

Figure 4.3: Parse tree for the noun phrase *European fish* with its accompanying transitive verb *manage* and noun phrase *the marine environment* with transitive verb *protect*. Both NPs occur as a direct object.

The RE grammar for the parser is defined as (1) \(<\text{VB}.*<\text{DT}?<\text{JJ}>*<\text{NN}>+\) for direct object noun phrases and (2) \(<\text{DT}?<\text{JJ}>*<\text{NN}>+<\text{VB}.*>+\) as subject noun phrases. (1) extracts a verb in any tense \(<\text{VB}.*>\), followed by an optional determiner \(<\text{DT}?>\), followed by 0 or more adjectives \(<\text{JJ}>*\) and ending with at least one noun \(<\text{NN}>+\). Figure 4.3 shows the parse tree of the phrase *manage European fish and protect the marine environment*. We extract two patterns *manage European fish* and *protect the marine environment*.

(2) extracts patterns starting with an optional determiner \(<\text{DT}?>\), followed by 0 or more adjectives \(<\text{JJ}>*\), followed by at least one noun \(<\text{NN}>+\) and ending in any verb tense \(<\text{VB}.*>\). Figure 4.4 shows the parse tree for the phrase *the northern fisherman catches* in which the NP occurs as a subject.

Figure 4.4: Parse tree for the noun phrase *the northern fisherman* with its accompanying transitive verb *catches*. The NP occurs as a subject.

4.3.4 Scoring Verbs

The verb tenses that precede an NP as a direct object, or follow the NP as a subject are stemmed with the Porter algorithm (Porter, 1980) to subsequently group NPs when creating the extraction patterns. Stemming groups verb tenses into the same stem—not necessarily the root of the verb. To score stemmed verbs we used a non-related text corpus containing rural information. The idea behind scoring the domain related
verbs is to limit the extraction patterns to verbs that are more frequently found in the fisheries domain. For example, the verb create is often found in all sorts of domains but the verb catch not. The noun phrases that are linked to the verb catch are more likely to contain semantically related words.

Let $D_1$ be the set of stemmed verbs from the seed word related corpus (Section 4.3.2) and $D_2$ be the set of stemmed verbs from the non-related corpus (rural corpus) and that the vocabulary size $|V|_{D_1} \approx |V|_{D_2}$. Now let $L$ be the combined set of stemmed verbs such that for every $l \in L$ we have that $l \in D_1 \cup D_2$. Let $N_{l,D_1}$ denote the number of instances $l$ contain in $D_1$. We estimate the probability that stemmed verb $l$ contained within the seed word related set $D_1$:

$$P(l|D_1) = \frac{N_{l,D_1}}{N_{l,D_1} + N_{l,D_2}} \quad (4.2)$$

For example, the verb supervise, which includes supervised and supervising, gets stemmed into supervis. A frequency of $D_1 = 12$ and $D_2 = 3$ results in a score of 0.8. The distribution of scores is shown in Figure 4.5 and shows the number of verbs with their estimated probability. For example, there are 50 verbs with a score $P \geq 0.9$ which are highly domain related, such as prohibit, catch, rescue, exploit, breath, deplete, fish.

Figure 4.5: Frequency of stemmed verbs and estimated probability.
Table 4.1: Example phrases where the transitive verb of a phrase gets stemmed into the same stem.

<table>
<thead>
<tr>
<th>stem</th>
<th>verb</th>
<th>noun</th>
<th>phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>regul</td>
<td>regulated</td>
<td>industry</td>
<td>... regulated industry ...</td>
</tr>
<tr>
<td>regul</td>
<td>regulated</td>
<td>seafood trade</td>
<td>... regulated seafood trade ...</td>
</tr>
<tr>
<td>regul</td>
<td>regulated</td>
<td>quota</td>
<td>... quota regulated ...</td>
</tr>
<tr>
<td>regul</td>
<td>regulates</td>
<td>legislation</td>
<td>... legislation regulates ...</td>
</tr>
<tr>
<td>catch</td>
<td>catching</td>
<td>fish</td>
<td>... catching wild fish ...</td>
</tr>
<tr>
<td>catch</td>
<td>catching</td>
<td>marine life</td>
<td>... catching marine life ...</td>
</tr>
<tr>
<td>catch</td>
<td>catch</td>
<td>deep-dwelling fish</td>
<td>... catch deep-dwelling fish ...</td>
</tr>
<tr>
<td>catch</td>
<td>catches</td>
<td>fisherman</td>
<td>... the fisherman catches ...</td>
</tr>
</tbody>
</table>

4.3.5 Verb Extraction Pattern

We selected verbs in which \( P \geq 0.5 \) and group them in steps of 0.1 such that \( 0.9 \leq P \leq 1.0, 0.8 \leq P \leq 1.0, \ldots, 0.5 \leq P \leq 1.0 \) so that we analyze the top 10%, 20%, 50% of verbs. To create the extraction pattern, any noun \(<\text{NN}>\) or compound noun \(<\text{NN}>+\) extracted by the chunker is grouped together if they share the same stemmed verb in either subject and direct object cases. For example, Table 4.1 shows some phrases for the root verb regulate which gets stemmed into regul and root verb catch which has an identical stem. The nouns industry, seafood trade, quota, and legislation are grouped together into an extraction pattern because they all share the same stem regulate.

For every NP, we have extracted the noun or compound noun. We have not restricted ourselves to the head noun as compound nouns are generally more informative for the fisheries domain. For example, marine science was accepted as a fisheries related word but science not. An overview of some extraction patterns are listed below:

- \{industry, seafood trade, quota, legislation\}
- \{fish, marine life, deep-dwelling, fish, fisherman\}
- \{catch, conservation, sustainability, growth, fishing, participation, tuna, management approach\}
- \{oxygen, air, fish, equipment, right\}
- \{sediment resuspension, world, bycatch, potential, fishing technique, hunger\}

We found 468 different extraction patterns in our corpus. An overview of the distribution is shown in Figure 4.6 (y-axis logarithmically scaled).
Figure 4.6: The frequency of nouns in an extraction pattern. The x-axis shows the number of extraction patterns found in our corpus (468). The y-axis shows the number of nouns grouped within that pattern.
4.3.6 Bootstrapping

The bootstrapping process starts with a list of 10 seed words (Section 4.3.1). Next, all the extraction patterns are scored by calculating the \( RlogF \) score (Riloff, 1996):

\[
RlogF(pattern_i) = \begin{cases} 
\frac{F_i}{N_i} \times \log_2(F_i) & \text{if } F_i \geq 1 \\
-1 & \text{if } F_i = 0 
\end{cases} \tag{4.3}
\]

Where \( F_i \) is the number of lexicon words found in \( pattern_i \) and \( N_i \) the total number of nouns in \( pattern_i \). Extraction patterns that contain nouns that are already part of the lexicon will get a higher score. The first iteration selects the top \( N \) patterns which are then placed into a pattern pool. We used \( N=20 \) for the first iteration and incremented it by 1 every next run to allow new patterns to enter the process.

The next step is to score all the nouns that are part of the pattern pool. We evaluated two scoring metrics: (1) BASILISK’s \( \text{AvgLog} \), (2) PMI that uses search counts from the Bing search engine. The \( \text{AvgLog} \) score is defined as:

\[
\text{AvgLog}(word_i) = \frac{\sum_{j=1}^{P_i} \log_2(F_j + 1)}{P_i} \tag{4.4}
\]

(1) \( \text{AvgLog} \) uses all the patterns to score the nouns found in the top \( N \) patterns. \( P_i \) is the number of patterns in which \( word_i \) occurs and \( F_j \) the number of lexicon words found in pattern \( j \). The nouns that are part of the pattern pool are given a higher score, thus being more semantically related, when they also occur in other extraction patterns with a high number of lexicon word matches.

(2) is based on hypernym collocation statistics proposed by (Igo and Riloff, 2009). We implement the PMI scoring metric within the bootstrapping process and dynamically calculate collocation statistics before adding new words in the lexicon. We hypothesize that lexicon words that occur more often in collocation with its domain are more likely to be semantically related. We use the number of hits between a lexicon word (hyponym) and its hypernym word by utilizing the \textsc{Near} operator from the Microsoft’s BING search engine. We choose collocation range of 10 and define the PMI score as:

\[
PMI(x, y) = \log \left( \frac{P(x, y)}{P(x) \cdot P(y)} \right) \tag{4.5}
\]

\[
PMI(x, y) = \log(N) + \log \left( \frac{\text{Count}(x, y)}{\text{Count}(x) \cdot \text{Count}(y)} \right) \tag{4.6}
\]
PMI(x, y) is the Pointwise Mutual Information that lexicon word x occurs with hypernym y, where P(x, y) is the probability that x and y occur together on the web (using the NEAR operator), P(x) the probability that x occurs on the web, and P(y) the probability that y occurs on the web. We would have to calculate probabilities such as P(x) by dividing the counts of x by the total number of web pages N. However, N is not known and can be omitted because it will be the same for each lexicon word. We can rewrite PMI(x, y) again by taking the log of the number of hits from the collocation statistics and dividing it by statistics of their individual parts.

Each noun that was part of the extraction pattern was given a score and the top-N nouns were selected to enter the lexicon. We added the noun with the highest score (N=1) to the lexicon and repeated the bootstrapping process.

4.4 Evaluation

We evaluated the lexicon entries by a gold standard dictionary. Domain experts have labeled every noun or compound noun that was found in the extraction patterns. A value of 1 was assigned if the word was related to the fisheries domain and 0 if it had no relationship. We did not distinguish between highly relevant domain words and less relevant words. For example, the word marine science is highly relevant and has often a direct link to the domain, but the word conservation in itself can be ambiguous. It could be related to e.g. conserving artwork or to prevent depletion of natural resources. However, a word was considered to be correct if any sense of the word is semantically related. Furthermore, unknown words were manually looked up for their meaning. For example, plecoglossus altivelis, oncorhynchus keta, leucosternon, peach anthias and khaki grunter are types of fish unknown to the annotators, yet are semantically related.

We have run the bootstrapping algorithm until the lexicon contained 100 words. We repeated the process for the top 10% to top 50% scored verbs as discussed in Section 4.3.5 and compared the AvgLog (S_b) and PMI (S_{pmi}) scoring metric. Examples of lexicon entries are: invertebrate seafood, ground fish, shellfish, demersal fish, mackerel, carp, crabs, deepwater shrimp, school, life, squid, hake, fisherman, cod, tuna, conservation reference size, trout, quota, sea, shrimp, mortality, freshwater, trawl, salmon, tenkara, snapper, method, license, attractor, pocket water, fee, vessel license, style fly, artisan, plastic worm, freshwater fly, saltwater, jack mackerel, trawler.

Figure 4.7 shows the accuracy (percentage of correct lexicon words) for S_b when learning 100 semantically related words. The lines represent the verb probability groupings (e.g. top 10%, 20%). The accuracy degrades when incorrect words enter the lexicon.
and starts to contribute in learning more incorrect words. The accuracy converges to around 70% for all probability groupings. Figure 4.8 shows the accuracy for $S_{pmi}$. It shows that using web statistics for hyponym-hypernym words affects the accuracy positively and negatively, yet still converges to around 70% after 100 words are learned. The use of $S_{pmi}$ affects the top 30% verbs substantially, outperforming $S_b$ up to learning 90 words. However, $S_{pmi}$ negatively affects the accuracy when looking at the top 50% verbs compared to $S_b$.

Figure 4.7: Graph that shows the accuracy when learning 100 words for the top 10% to top 50% verbs when scoring candidate words with BASILISK's AvgLog ($S_b$).

Scoring verbs before creating extraction patterns causes differences in accuracy up to a lexicon size of 90. Verbs that occur more often in domain related text, such as discussed in Section 4.3.4, essentially benefit the accuracy of the lexicon up to a certain size yet have limited effect on large lexicons. We would, however, have expected that the top 10% verbs outperform the top 20% and so on. This does not seem to hold for both $S_b$ and $S_{pmi}$. For example, when using $S_b$ and learning 50 words, the top 50% verbs achieved a higher accuracy (0.74), compared to the top 20% (0.66). Similarly, when learning 40 words and using $S_{pmi}$, the top 20% verbs achieved lower accuracy (0.68) than the top 30% verbs (0.78). The top 10% verbs for $S_b$ and $S_{pmi}$ achieve the highest accuracy in nearly all stages of the bootstrapping process. Small lexicons would benefit from selecting only the top 10% verbs to create extraction patterns, achieving an accuracy of around 0.8 when learning 50 words. An overview of all accuracy scores is given in Table 4.2.
Figure 4.8: Graph that shows the accuracy when learning 100 words for the top 10% to top 50% verbs when scoring candidate words with Pointwise Mutual Information ($S_{pmi}$).
Table 4.2: Lexicon accuracy when learning up to a 100 words. Accuracy scores are given for the top 10%, ..., 50% verbs for both $S_b$ (scoring new entries with BASILISK’s \textit{AvgLog}), and $S_{pmi}$ (scoring new entries with PMI).

<table>
<thead>
<tr>
<th>Lexicon entries</th>
<th>$0.5 \leq P \leq 1.0$</th>
<th>$0.6 \leq P \leq 1.0$</th>
<th>$0.7 \leq P \leq 1.0$</th>
<th>$0.8 \leq P \leq 1.0$</th>
<th>$0.9 \leq P \leq 1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_b$</td>
<td>$S_{pmi}$</td>
<td>$S_b$</td>
<td>$S_{pmi}$</td>
<td>$S_b$</td>
</tr>
<tr>
<td>20</td>
<td>0.8</td>
<td>0.75</td>
<td>0.8</td>
<td>0.75</td>
<td>0.8</td>
</tr>
<tr>
<td>30</td>
<td>0.73</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>40</td>
<td>0.68</td>
<td>0.65</td>
<td>0.68</td>
<td>0.6</td>
<td>0.68</td>
</tr>
<tr>
<td>50</td>
<td>0.74</td>
<td>0.68</td>
<td>0.64</td>
<td>0.68</td>
<td>0.64</td>
</tr>
<tr>
<td>60</td>
<td>0.72</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>70</td>
<td>0.69</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>80</td>
<td>0.68</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>90</td>
<td>0.69</td>
<td>0.7</td>
<td>0.69</td>
<td>0.7</td>
<td>0.69</td>
</tr>
<tr>
<td>100</td>
<td>0.7</td>
<td>0.69</td>
<td>0.69</td>
<td>0.68</td>
<td>0.7</td>
</tr>
</tbody>
</table>
4.5 Conclusion

In this paper, we presented a bootstrapping algorithm based on BASILISK and a highly related corpus that was created by mining web pages. We have created the corpus by utilizing the same set of seed words that initially was used to start the bootstrapping process. We used extraction patterns to group noun phrases with similar semantic meaning by grouping them when they share the same stemmed verb. We scored the extraction patterns with a non-related text corpus and calculated accuracy scores for the top 10%, 20%, 30%, 40% and 50%. Next to using BASILISK original scoring metric, we used a PMI score by looking at hyponym-hypernym collocation statistics.

We found varied results between the scored extraction patterns. Patterns that were created by looking at strong verbs that most often occur in domain related text, and less frequent in general (non-related) text, created a higher accuracy lexicon when looking at the top 10% scored verbs while other top scores showed mixed results. The use of collocation statistics by utilizing the NEAR operator of Microsoft’s Bing search engine provided better accuracy for a number of scores but simultaneously caused a degrade in the accuracy for other verb scores.

The achieved accuracy covers web text for the fisheries domain and more research is needed concerning the generalizability into other domains and other forms of text, such as scientific literature and other technical language. Furtermore, research is needed to explain why accuracy varies between verb scores and why collocation statistics work better in some cases. Finally, research is also necessary when scoring the verbs against a non-related text corpus to see which types or genres of non-related domain corpora affects the domain under study.
Despite increased fisheries science output and publication outlets, the global crisis in fisheries management is as present as ever. Since a narrow research focus may be a contributing factor to this failure, this study uncovers topics in fisheries research and their trends over time. This interdisciplinary research evaluates whether science is diversifying fisheries research topics in an attempt to capture the complexity of the fisheries system, or if it is multiplying research on similar topics, attempting to achieve an in-depth, but possibly marginal, understanding of a few selected components of this system. By utilizing latent Dirichlet allocation as a generative probabilistic topic model, we analyze a unique dataset consisting of 46,582 full-text articles published in the last 26 years in 21 specialized scientific fisheries journals. Among the 25 topics uncovered by the model, only one (Fisheries management) refers to the human dimension of fisheries understood as socio-ecological complex adaptive systems. The most prevalent topics in our dataset directly relating to fisheries refer to Fisheries management, Stock assessment, and Fishing gear, with Fisheries management attracting the most interest. We propose directions for future research focus that most likely could contribute to providing useful advice for successful management of fisheries.

This work was originally published as:

5.1 Introduction

Following a similar trend to scientific research at large, fisheries science research output has significantly increased in the last three decades (Aksnes and Browman, 2016), in parallel with an increase in the number of fisheries scientific journals (Mather et al., 2008). This rapid expansion of the field is attributed to the growing concern about the state of global fish stocks and to the major role that science has been playing in fisheries management (Jarić et al., 2012). However, despite the increased volume of fisheries science output and publication outlets, the global crisis in marine capture fisheries management is as present as ever, with unforeseen consequences ranging from fisheries-induced evolutionary changes among wild fish populations (Belgrano and Fowler, 2013) to conflicts between states over the implementation of best available science (Brooks et al., 2016). There are various hypotheses regarding causes and contributing factors for failures of fisheries management, including data uncertainty, model inadequacy, ecosystem structure, institutional efficacy, economic discord or research focus (Smith and Link, 2005). Among these, research focus is the least explored (Smith and Link, 2005). Using hybrid content analysis of a unique dataset consisting of 46,582 fisheries science full-text articles published in the last 26 years, we uncover focus topics in fisheries research and their trends.

Fisheries are socio-ecological complex adaptive systems (SECASs) in which macroscopic properties emerge from local actions that spread to higher scales due to agents’ (fish and humans) collective behavior; these properties then feed back, in a nonlinear way, influencing individuals’ options and behaviors, but they typically only do so diffusely and over long timescales (Levin et al., 2013; Ostrom, 2009). A fishery can be defined as "the complex of people, their institutions, their harvest and their observations associated with and including a targeted stock or group of stocks (i.e. usually fish), and increasingly, the associated ecosystems that produce said stocks" (Link, 2010). Deconstructing the concept, the two main dimensions of a fishery are the human dimension (i.e. human agents, communities of these, and their institutions) and the natural dimension (i.e. biotic, such as predator species and prey species, and abiotic, such as water temperature and nutrients) (Charles, 2000; Lennox et al., 2017; Österblom et al., 2013). The purpose of this study is to assess whether fisheries science output is reflecting this conceptual diversity of fisheries as SECASs, and if so, to what extent. Is science diversifying fisheries research topics in an attempt to capture the complexity of the fisheries system, or is it multiplying research on similar topics, trying to achieve an in-depth, but possibly marginal, understanding of a few selected components of this system? Based on the critical reflection that "the majority of fisheries scientists have a biologically oriented background, they can be a bit naïve regarding other factors when it comes to the prominence of economic or social considerations" (Link, 2010), the working hypothesis of this study is that the human dimension of fisheries might be under-represented in the fisheries specialty literature.

The assessments of the development and trends in fisheries science have so far been
mostly based on reviews (e.g. (Johnson et al., 2013)) or bibliometric evaluations (e.g. (Aksnes and Browman, 2016)) of scientific publications in the field. Limitations of these studies include: taking into account only a limited number of publications (e.g. (Jarić et al., 2012)), or a limited time period (e.g. 2000–2009; (Jarić et al., 2012)); having a limited scope (e.g. artisanal coral reef fisheries research (Johnson et al., 2013); fish stock assessment research (Kumaresan et al., 2014); shark by-catch research (Molina and Cooke, 2012)); using proxies for full-text articles (e.g. titles (Jarić et al., 2012); abstracts (Aksnes and Browman, 2016)), or proxies for topics of research (e.g. one word per topic (Jarić et al., 2012; Aksnes and Browman, 2016)). Most importantly, all these previous attempts to map the fisheries science field are top-down approaches, with topics of interest manually predefined by the analysts (e.g. species, region, habitat, study object (Jarić et al., 2012)), and the analyzed data manually assigned to these topics. However, such approaches are limited due to the subjectivity inherent in human decisions, and the analysis of the same research field could yield opposite results (e.g. (Rose et al., 2011) vs. (Hill and Lackups, 2010) evaluation of the field of cetacean research).

In contrast to previous approaches, we follow a completely novel strategy for the field of fisheries science, a bottom-up approach (Debortoli et al., 2016) by utilizing topic modeling to uncover hidden research topics within fisheries science publications. Topic modeling algorithms are machine learning methods to automatically uncover hidden or latent thematic structures from large collections of documents. Topic models can produce a set of interpretable topics that can be viewed as groups of co-occurring words that are associated with a single topic or theme (DiMaggio et al., 2013). Such groups of co-occurring words (i.e. topics) are words that tend to frequently come up together within the same linguistic context; more frequently than one would expect by chance alone. These co-occurring words tend to purport similar meaning and refer to a similar subject. For example, in the context of fisheries science, an author might write a text to which she/he gave the key words 'community structure', 'subtropical areas', 'reference points', and 'weight'. This text might use more frequently the words 'parameters', 'estimation', 'stock', 'modeling', 'male', 'female', 'sex', and 'spawning'. If we wanted to use topic modeling to uncover the latent topics of this hypothetical text, based on how often these most used words would appear together (i.e. co-occur), the automated topic model would group the first four words in one topic and the last four words in a different topic. These two topics would then be manually labeled by a domain expert most likely as 'stock assessment modeling' and 'reproduction', respectively. Note that the subject of these two topics is not similar to the one that might be inferred from the key-words given to this hypothetical text. Thus, these topics are latent, they are hidden in the pattern of co-occurring words. In essence, topic models are able to exploit the co-occurrence structure of texts and produce the topics as lists of words that frequently come up together, within and between documents; technically, such lists of words are probability distributions over words.

The topics emerge from the statistical properties of the documents and, thus, overcome the need for manual annotation of the collection of texts, though manual interpreta-
CHAPTER 5. TOPIC ANALYSIS OF FISHERIES SCIENCE

Identification of the subject of a topic might still be needed, as it is yet considered the gold standard in the domain of topic modeling (Lau et al., 2011). As such, we allow the documents to speak for themselves and view the documents through the computational lens of the topic model, rather than relying on the manifest or reported content by their authors. Document collections that are too large to explore manually can now be analyzed to study phenomena of the sort that can only be viewed through the macroscopic computational lens of the topic model (Mohr and Bogdanov, 2013). Topic modeling approaches have been helpful in elucidating the key ideas within a set of documents, such as articles published in the journal PNAS (Griffiths and Steyvers, 2004), political science texts (Grimmer and Stewart, 2013) or data-driven journalism (Rusch et al., 2013). Moreover, it is considered that this approach could provide insight into the development of a scientific field and changes in research priorities (Neff and Corley, 2009), and do so with greater speed and quantitative rigor than would otherwise be possible through traditional narrative reviews (Grimmer and Stewart, 2013). As such, this topic modeling method has been applied, for example, in the domain of transportation research (Sun and Yin, 2017), computer science (Hall et al., 2008; Wang et al., 2011; Wang and McCallum, 2006), fisheries modeling (Syed and Weber, 2018), conservation science (Westgate et al., 2015), and the fields of operations research and management science (Gatti et al., 2015).

After identifying the hidden topics of fisheries science, we analyze the extent to which these topics cover the complexity of the fisheries domain. Afterwards, we examine topic similarity, topic co-occurrence, topic prevalence, and topical trends over the last 26 years. We furthermore identify patterns in increasing and decreasing topic trends over specific periods of time (i.e. hot and cold topics in 1990–1995, 1995–2000, 2000–2005, 2005–2010, and 2010–2016), and describe the distribution of uncovered topics over journals.

5.2 Methods

5.2.1 Latent Dirichlet Allocation

This paper utilizes the topic model latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is a Bayesian probabilistic topic model and follows the assumption that documents exhibit multiple topics in mixing proportions, thus capturing the heterogeneity of, for example, research topics within scientific publications. In statistics, this is often referred to as a mixed-membership model (Erosheva et al., 2004). Technically, a topic is a multinomial distribution of words in the vocabulary, where each word has a different probability within each topic; within a topic, more prominent words have a higher probability and groups of high probability words can be considered as co-occurring clusters or constellation of words that describe a certain underlying topic or theme. A document might be 60% about the topic fisheries management and 40% about the
topic "about" a subject (e.g. fisheries management) relates to the probability distribution of words that places high probability to words that would be used to describe the subject (DiMaggio et al., 2013). Note that the underlying topics and to what extent the document exhibits these topics are not known in advance. These details are the output of the LDA analysis and emerge automatically from the statistical properties of the documents and the assumptions behind LDA.

One way to think about LDA is to imagine a document in which one highlights words with colored markers. Words that relate to one topic are colored blue, words that relate to another topic are colored red, and so on. After all of the words have been colored (excluding words such as 'the', 'a'), all the words with the same color are the topics, and the article will blend the colors in different proportions. Different documents will have different blends of colors, and we could use the proportion of the various colors to situate this specific document in a document collection (e.g. documents addressing mainly the blue topic). Moreover, documents with the same blend of colors discuss the topics in similar proportion and are considered closely related from a topical perspective. Technically, documents with similar topic distributions are close in Kullback-Leibler divergence, a measure to calculate the distance between two probability distributions. LDA as a statistical model captures this intuition. We refer the interested reader to Blei (2012) for a concise introduction to LDA.

LDA is best described by its generative process, i.e. the imaginary probabilistic recipe that generates the documents as well as the hidden structure. The hidden structure is the topics, modeled as distributions of words, and the topic proportions per document, where each document has some probability for each latent topic (i.e. mixing topic proportions). More formally, the generative process also assigns each word to a topic as to allow for documents to exhibit multiple topics, analogous to the colored words example. Given the observed documents, the aim now is to infer the hidden structure to answer the question "what is the likely hidden topical structure that have generated these documents?", a process that can be seen as reverse-engineering the generative process. Technically, we want to infer the posterior distribution of the latent variables given the observed documents. An analogy to this process is described by the local farmers market example; one might estimate what vegetables and what quantities are being sold at the local farmers market by post-hoc inspection of people’s shopping basket. Seeing more baskets refines the estimation of the products and their quantities and provides an estimate of the market’s produce (Rhody, 2013). Mainly two types of inference techniques can be discerned: sampling-based algorithms (e.g. (Newman et al., 2007; Porteous et al., 2008) and variational-based algorithms (e.g. (Blei and Jordan, 2006; Teh et al., 2006; Wang et al., 2011). To simplify posterior inference, LDA uses a Dirichlet distribution as a conjugate prior for the multinomial distribution, hence the name latent Dirichlet Allocation. The posterior distribution will reveal the probability distributions of words for each topic, and the topic proportions per document. Note that the obtained structure is latent, and therefore, the probability distributions of words are not semantically labeled. However, when sorted, the words with the highest probability within a topic will relate to what one would call a topic or theme (DiMaggio
et al., 2013; Mohr and Bogdanov, 2013). In this context, it is important to mention that the LDA model does not give a name to the identified latent topics (i.e. the model does not label the topics). The output of the model groups the co-occurring words under numbered topics (i.e. topic 1, topic 2, topic 3 etc.). Albeit a subjective endeavor, possibly affecting the statistical objectiveness of the LDA method, in order to increase the readability and interpretability of these topics, a human analyst can be used to interpret what is the common subject of the words within each topic and consequently give a name (i.e. a post-hoc label) to this topic (DiMaggio et al., 2013). Research into automatic assignment of topic labels exists, however manual annotation by a human expert is still considered the gold standard in labeling topics (Lau et al., 2011). For our study, instead of using the topic numbering provided by the LDA model (i.e. topic 1, topic 2, topic 3 etc.), and in order to increase readability of the text and interpretability of the results, we chose to give a specific label to each topic using the gold standard in this domain, i.e. manual annotation (see the section Labeling Topics).

5.2.2 Assumptions behind LDA

LDA is a bag-of-words model in which documents are represented as unordered sequences of words. Such an assumption neglects word order and possibly important cues to the content of a document (Steyvers and Griffiths, 2007). Although an unrealistic assumption, it is reasonable when uncovering semantic structures of text (Blei, 2012; Blei and Lafferty, 2006). Consider a thought experiment where the words of a document are shuffled. After finding a high number of words like spawning, eggs, and growth, one can imagine that the document deals with some aspects of reproduction. LDA further assumes document exchangeability, that is, the order in which documents are analyzed is unimportant, yet all documents are analyzed at the end of the LDA analysis. Consequently, LDA is unable to explicitly capture evolving topics from documents that cover large time spans (e.g. centuries). To do that, we would need to resort to a more complicated and computationally expensive dynamic topic model (Blei and Lafferty, 2006). Such an approach is currently not feasible given the large dataset used here, but would be interesting to explore in future work. Nevertheless, the assumption of document exchangeability captures the fact that current literature builds on top of previous literature, but is a limitation for topics that have radically changed in the way they are described (e.g. terminology) in literature. For example, the field of atomic physics was described by words relating to “matter” in the late 19th Century, “electron” in middle of the 20th Century, and “quantum” in the late 20th Century. Likewise, the field of neuroscience evolved from being described by words relating to “nerve”, to “neuron”, to “ca2” over the last 100 years (Blei and Lafferty, 2006). The dynamic topic model uses a sequence of time slices in which topics are conditioned on the previous topics, which is a limitation of the standard LDA model used in this study.
5.2.3 Creating the Data Set

Taking into consideration the issue of having access to the electronic version of the text, we decided to include in our analysis only journal articles, as the majority of these are now available for download from online databases. Thus, we have excluded books, books chapters, and reports, something that may have introduced bias in our results. Furthermore, due to computational and time constraints, we have limited the number of journals included in our analysis. Thus, the total volume of fisheries publications is underestimated in our analysis, something that limits the results of this study. The dataset was constructed following a set of inclusion criteria to obtain a diverse set of journals that reflect fisheries science while maintaining computational feasibility. First, we included all journals with the term “fisheries” or “fishery” in their title that are listed by the Fisheries Science Citation Index Extended (SCIE) 2016 provided by Thomson Reuters, and having an impact factor of $\geq 1.0$. Second, we included all journals from the Fisheries SCIE 2016 that do not include these words in their titles, but explicitly address fisheries in their aims and scopes, and having an impact factor of $\geq 1.0$. Third, we included the top four journals with the highest 2016 impact factor with the term “marine” in their title, indexed by any list from SCIE or Social Science Citation Index (SSCI), and explicitly addressing fisheries in their aims and scopes. All journals were subject to the University of Tromsø—The Arctic University of Norway subscription rights. A total of 21 journals satisfied these criteria (Table 5.1). Although journals that do not match these criteria also publish fisheries research, such journals were not considered to be specialized fisheries research outlets (e.g. the journal *Ecology and Society*).
Table 5.1: An overview of the dataset used when creating the LDA model to uncover latent topics from fisheries publications. The dataset consists of 46,582 full-text publications from 21 top-tier fisheries journals. Fisheries rank and impact factor are extracted from the 2016 Fisheries ISI Journal Citation Reports (JCR) provided by Thomson Reuters. Journals without a rank are “marine” journals not covered by the JCR fisheries index, but cover fisheries aspects within their aims and scopes. IF is the impact factor, \( Y_{\text{min}} \) is the lowest publication year, \( Y_{\text{max}} \) the highest publication year, \( N \) the number of documents (publications) deemed fit for further analysis, \( \bar{W} \) the mean number of words, Std. \( W \) the estimated standard deviation of number of words, and \( \bar{V} \) the mean vocabulary size. Note that word and vocabulary statistics are obtained after the data cleaning process.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Rank</th>
<th>IF</th>
<th>( Y_{\text{min}} )</th>
<th>( Y_{\text{max}} )</th>
<th>( N )</th>
<th>( \bar{W} )</th>
<th>Std. ( W )</th>
<th>( \bar{V} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish and Fisheries</td>
<td>1</td>
<td>9.013</td>
<td>2000</td>
<td>2016</td>
<td>419</td>
<td>4161</td>
<td>2022</td>
<td>1085</td>
</tr>
<tr>
<td>Reviews in Fish Biology and Fisheries</td>
<td>3</td>
<td>3.575</td>
<td>1991</td>
<td>2016</td>
<td>659</td>
<td>4142</td>
<td>3014</td>
<td>1070</td>
</tr>
<tr>
<td>Fisheries</td>
<td>5</td>
<td>3</td>
<td>1997</td>
<td>2016</td>
<td>477</td>
<td>1977</td>
<td>1313</td>
<td>667</td>
</tr>
<tr>
<td>Aquaculture Environment Interactions</td>
<td>6</td>
<td>2.905</td>
<td>2010</td>
<td>2016</td>
<td>203</td>
<td>3172</td>
<td>1045</td>
<td>844</td>
</tr>
<tr>
<td>ICES Journal of Marine Science</td>
<td>7</td>
<td>2.76</td>
<td>1990</td>
<td>2016</td>
<td>3903</td>
<td>2432</td>
<td>1039</td>
<td>689</td>
</tr>
<tr>
<td>Reviews in Fisheries Science &amp; Aquaculture</td>
<td>9</td>
<td>2.545</td>
<td>1997</td>
<td>2016</td>
<td>375</td>
<td>4443</td>
<td>4429</td>
<td>1061</td>
</tr>
<tr>
<td>Canadian Journal of Fisheries and Aquatic Sciences</td>
<td>10</td>
<td>2.466</td>
<td>1996</td>
<td>2016</td>
<td>4423</td>
<td>3205</td>
<td>1121</td>
<td>828</td>
</tr>
<tr>
<td>Fisheries Research</td>
<td>11</td>
<td>2.185</td>
<td>1995</td>
<td>2016</td>
<td>3610</td>
<td>2491</td>
<td>1084</td>
<td>679</td>
</tr>
<tr>
<td>Ecology of Freshwater Fish</td>
<td>13</td>
<td>2.054</td>
<td>1996</td>
<td>2016</td>
<td>932</td>
<td>2471</td>
<td>972</td>
<td>709</td>
</tr>
<tr>
<td>Marine Resource Economics</td>
<td>14</td>
<td>1.911</td>
<td>2010</td>
<td>2016</td>
<td>159</td>
<td>3609</td>
<td>1279</td>
<td>836</td>
</tr>
<tr>
<td>Fisheries Oceanography</td>
<td>18</td>
<td>1.578</td>
<td>1997</td>
<td>2016</td>
<td>752</td>
<td>3036</td>
<td>1123</td>
<td>765</td>
</tr>
<tr>
<td>Journal of Fish Biology</td>
<td>21</td>
<td>1.519</td>
<td>1990</td>
<td>2016</td>
<td>7075</td>
<td>2113</td>
<td>1551</td>
<td>652</td>
</tr>
<tr>
<td>Transactions of the American Fisheries Society</td>
<td>22</td>
<td>1.502</td>
<td>1997</td>
<td>2016</td>
<td>2381</td>
<td>3168</td>
<td>1186</td>
<td>791</td>
</tr>
<tr>
<td>CCAMLR Science</td>
<td>24</td>
<td>1.429</td>
<td>1990</td>
<td>2016</td>
<td>314</td>
<td>1722</td>
<td>1123</td>
<td>506</td>
</tr>
<tr>
<td>Knowledge and Management of Aquatic Ecosystems</td>
<td>26</td>
<td>1.217</td>
<td>1997</td>
<td>2016</td>
<td>590</td>
<td>1860</td>
<td>958</td>
<td>622</td>
</tr>
<tr>
<td>Marine and Coastal Fisheries</td>
<td>28</td>
<td>1.177</td>
<td>2009</td>
<td>2016</td>
<td>274</td>
<td>3539</td>
<td>1144</td>
<td>876</td>
</tr>
<tr>
<td>Aquatic Conservation: Marine and Freshwater Ecosystems</td>
<td>n/a</td>
<td>3.13</td>
<td>1991</td>
<td>2016</td>
<td>1328</td>
<td>2873</td>
<td>1251</td>
<td>842</td>
</tr>
<tr>
<td>Marine Ecology Progress Series</td>
<td>n/a</td>
<td>2.292</td>
<td>1990</td>
<td>2016</td>
<td>12674</td>
<td>3046</td>
<td>2242</td>
<td>812</td>
</tr>
<tr>
<td>Marine Policy</td>
<td>n/a</td>
<td>2.235</td>
<td>1990</td>
<td>2016</td>
<td>2516</td>
<td>3145</td>
<td>1264</td>
<td>889</td>
</tr>
</tbody>
</table>

TOTAL 46582
Moreover, even though some of the most influential and highly-cited fisheries papers are published in high-impact journals such as *Nature* and *Science*, they only marginally contribute to the total number of papers published in fisheries-related journals. Including in our analysis all publications from Nature and Science would result in a high number of fisheries-irrelevant topics (e.g. astrophysics), as these journals typically publish a broad range of topics. Using keyword searches to obtain only fisheries related publications would be a top-down approach, and hence, would be biased by: (i) the search terms used, and (ii) the way publications are indexed and, subsequently, retrieved.

We downloaded full-text research articles published in the 21 journals covering fisheries aspects for a time span of 26 years (1990–2016) to allow for enough variation in publication trends. Analyzing full-text articles, compared to just abstract data, results in more detailed and higher quality topics (Syed and Spruit, 2017). Only research articles were considered and other types of publications, such as errata, conference reports, forewords, announcements, dedications, letters, comments, and book reviews, were excluded. A total of 46,582 articles were deemed fit for further analysis. The year of publication was chosen to be the issue year in which the article appeared in print, regardless of the accepted year or (first) online publication. Information about the journal name, the time range for which articles were collected, the journal’s impact factor, the total number of articles deemed fit for further analysis, and word statistics are given in 5.1. Additionally, an overview of the number of publications per journal per year is shown in Fig. 5.1. Not all journals provided articles for the complete time span of 26 years. For example, the journal *Fish and Fisheries* started in 2000 and, therefore, only articles from 2000 to 2016 were included in the study. Another example relates to the journal subscription rights, which did not allow data collection for all years. For example, the *Canadian Journal of Fisheries and Aquatic Sciences* started in 1901, but our subscription only allowed access from 1996.
**Figure 5.1:** The number of publications (y-axis) per journal (color-coded) for the years 1990–2016 (x-axis) that were used to create the LDA model. The total number of documents was 46,582.
All articles appeared in portable document format (PDF) and were first converted to their plain text representation. This resulted in a complete transformation from PDF to plain text for all elements of each article, including the header, title, author info, affiliation info, abstract, keywords, content, tables, bibliography, and captions. Several articles, mainly from the early 1990s, were image-based PDFs that were unsuitable for direct conversion from PDF to plain text. In these cases, the Tesseract optical character recognition (OCR) software library was used to subsequently convert these articles into text-based PDFs and then to their plain text representation. To make sure we only analyzed the content text of each article, we used regular expressions, an advanced text pattern search method, to remove boilerplate content such as journal information, article metadata, acknowledgments, and bibliographies. Additionally, multi-language abstracts or non-English articles (e.g. articles that appeared in French), were also removed.

LDA is a bag-of-words model in which documents are represented as sequences of individual word features. As such, every document was tokenized. Tokenization is the process of obtaining individual words (also known as unigrams) from sentences. Unigrams lose important semantic cues that are encoded by compound words. To overcome this, bi-grams were included by combining two consecutive unigrams that occurred ≥ 20 times within each document. As a result, compound words, such as “rainbow trout”, are preserved. Additionally, we used named entity recognition (NER), a technique from natural language processing (NLP), to retrieve entities related to names, nationalities, companies, locations, objects etc. from the documents. Entities such as “the European Union”, “the Norwegian Research Council”, and “marine protected areas” are thus preserved and included in the analysis. The inclusion of bi-grams and entities allows for a richer bag-of-words representation than a standard unigram representation.

Although all tokens within a document serve an important function, for topic modeling they are not all equally important. We proceed by filtering out numbers, punctuation marks and single-character words as they bear no topical meaning. Furthermore, we removed stop words (e.g. the, is, a, which), words that occurred only once (e.g. mainly typos and incorrectly hyphenated words), and words that occurred in ≥ 90% of the documents (e.g. result, study, show) as they serve no discriminative topical significance. Omitting frequently occurring words prevents such words from dominating all topics.

For grammatical reasons, different word forms or derivationally related words can have a similar meaning and, ideally, we would want such terms to be grouped. Stemming and lemmatization are two NLP techniques to reduce inflectional and derivational forms of words to a common base form. Stemming heuristically cuts off derivational affixes to achieve some normalization, albeit crude in most cases. Stemming loses the ability to relate stemmed words back to their original part-of-speech, such as verbs or nouns, and decreases the interpretability of topics in later stages (Evangelopoulos et al., 2012). For example, the term 'fishing' will be stemmed to 'fish', likewise, “modeling” will be stemmed to “model”, and cannot be returned to its original part-of-speech (i.e.
verb). Our analysis uses lemmatization, which is a more sophisticated normalization method that uses a vocabulary and morphological analysis to reduce words to their base form, called lemma. It is best described by its most basic example, normalizing the verbs 'am', 'are', 'is' to 'be', although such terms will be filtered out from our analysis. Likewise, lemmatization correctly normalizes 'fisheries' and 'fishery', and 'policies' and 'policy'. Additionally, uppercase and lowercase words were grouped. The final corpus consisted of 46,582 full-text publications with around 130 million words and 170,000 unique words.

5.2.4 Creating the LDA Model

LDA assumes that the number of topics to uncover is known in advance and is set by the K-parameter. As the optimal number of topics is not known in advance, we created 50 different LDA models by varying the K-parameter from 1 to 50. Measures to determine the optimal LDA model are described in the next section. The LDA models are created using the Python library Gensim (Rehurek and Sojka, 2010). Since LDA is a Bayesian probabilistic model, we can incorporate some prior knowledge into the model. Prior knowledge can be encoded by symmetrical or asymmetrical Dirichlet priors. A symmetrical prior distribution of topics within documents assumes that all topics have an equal probability of being assigned to a document. Such an assumption ignores that certain topics are more prominent in a document collection and, consequently, would logically have a higher probability to be assigned to a document. Conversely, specific topics are less common and, thus, not appropriately reflected with a symmetrical prior distribution. Logically speaking, an asymmetrical prior would capture this intuition and would, therefore, be the preferred choice (Syed and Spruit, 2018a; Wallach et al., 2009). Additionally, we iteratively optimize the prior using the Newton-Rapson method (Huang, 2005) by learning it from the data. To infer the hidden variables (i.e. inferring the posterior distribution of the hidden variables given the observed documents) we use variational inference called ‘online LDA’ (Hoffman et al., 2010).

5.2.5 Calculating Model Quality

Analogous to choosing the right number of clusters for techniques such as k-nearest neighbors, choosing the right number topics is an important task in topic modeling, including LDA, to appropriately capture the underlying topics in a dataset. A low number of topics will result in a few too broad topics, with high values capturing meaningless topics; such topics are merely the result of the statistical nature of LDA. Several approaches to determine the optimal number of topics have been proposed. One such approach is to fit various topic models to a training set of documents and calculate a model fit on a test set (held-out data) (Scott and Baldridge, 2013). The model that fits best on the test set would be considered a better model. However, topic models are
used by humans to interpret and explore the documents, and there is no technical rea-
son that the best-fitted model would aid best in this task (Blei, 2012). In fact, research
has shown that such measures negatively correlates with human interpretation (Chang
et al., 2009).

Another approach is to assess the quality of topics with human topic ranking, which
is considered the gold standard when assessing the interpretability of generated top-
ics. Such ranking is often based on word or topic intrusion tests, in which an intruder
word or topic needs to be recognized within a set of related or cohesive words or top-
ics (Chang et al., 2009). However, this approach is time-consuming and expensive as
for every created topic model (e.g. 1 to 50), and for every topic within that model,
the interpretability of individual words and sets of words need to be assessed. To
circumvent this, a more quantitative approach is preferred while maintaining human
interpretability. One way is to assess the quality of topics with coherence measures that
are based on the distributional hypothesis (Harris, 1954), which states that words with
similar meanings tend to co-occur within similar contexts. Such an approach, drawing
on the philosophical premise that a set of statements or facts is said to be coherent if
its statements or facts support each other, informs us about the understandability and
interpretability of topics from a human perspective. This study utilized the \(C_V\) coher-
ence measure (Röder et al., 2015), which has shown the highest correlation with all
available human topic ranking data, and is thus an appropriate quantitative approach.
The \(C_V\) coherence score for all 50 LDA models was calculated, and an elbow method,
estimating the (inflection) point where adding more topics will not increase coherence,
was used to obtain the optimal number of topics.

5.2.6 Labeling Topics

As previously described, the topical structure that permeates the document collection
is latent, and the probability distributions of words (i.e. topics) are not semantically
labeled (i.e. they are not given a name by the LDA model). When sorted, the top 10 or
top 15 high probability words within each topic are used to describe the topic. How-
ever, algorithmic analyses of content remain limited in their capacity to understand
latent meanings or the subtleties of human language (Lewis et al., 2013) and manual
labeling is still consider the gold standard in topic modeling (Lau et al., 2011). Thus,
the labeling of each topic (i.e. giving a name to each topic; e.g. a topic with the five
most probable words being 'pig', 'cow', 'sheep', 'goat', 'horse' would be labeled as 'do-
mestic animals') was performed by a human analyst, i.e. a fisheries domain expert.
When identifying the common subject of each topic (i.e. the name or the label of the
topic), the analyst used the following procedure. First, the analyst closely inspected
the 15 most probable words from each topic. Second, the analyst inspected the titles
of the documents in the dataset that were included by the topic model in that respec-
tive topic. The interested reader can find a sample of publication titles that have high
probability within a single topic in Table 5.4 in the Appendix. Third, based on the
previous two steps, the analyst labeled a topic (i.e. gave it a name; e.g. if the LDA model included in topic 1 the words 'pig', 'cow', 'farm' and the titles of the documents included by the model in this topic have in common the subject of domestic animals, then the analyst gave topic 1 the label of 'domestic animals'). Furthermore, to validate the labeling of the topics, we visualized the topics in a two-dimensional area by computing the distance between topics (Chuang et al., 2012) and applying multidimensional scaling (Sievert and Shirley, 2014). This two-dimensional topic representation displays the similarity between topics with respect to their word distribution over topics, i.e. the words and their corresponding probability within the topic. Clustering and overlapping nodes indicate similar word distributions and the surface of the node indicates the relative topic prevalence in the complete set of documents. The topic prevalence indicates how widespread a topic is within all the documents, as all topic proportions add up to 100%. In a fourth step, the analyst used this visualization in order to validate the choice of the final label for each topic (e.g. topics using similar vocabulary usually refer to similar subjects; thus, for example, the topics labeled by the analyst 'domestic animals' and 'astrophysics' appearing close together in the two-dimensional topic representation would raise suspicions and the analyst would have to go through the labeling procedure again in order to find labels that make sense for the two vocabulary-close topics). The labels were further validated in a fifth step, as described in the section Validation of Results.

5.2.7 Calculating Topical Trends over Time

To gain insight into the topical temporal dynamics of the fisheries field, we aggregated the document topic proportions for each year and for every individual topic into a composite topic-year proportion. Doing so provides a sense of how the prevalence of each topic within fisheries science publications has changed over the last 26 years. Additionally, to obtain insight into increasing and decreasing topical trends, we fit a one-dimensional least square polynomial for different time intervals. The polynomial coefficient is used as a proxy for the trend and defines the slope of the composite topic-year proportions for a range of years. Coefficients are multiplied by the number of years within each time interval to obtain the change measured in percentage points. Positive values indicate increasing or hot topics, and negative values decreasing or cold topics. The time intervals allow for historical comparisons between 1990–1995, 1995–2000, 2000–2005, 2005–2010, and 2010–2016. Color coding is used to resemble the hot (i.e. red) and cold (i.e. blue) topical trends.

5.2.8 Calculating Topic over Journals

Following a similar approach as topical trends over time, we aggregate topic proportion per journal to gain insight into how topics are covered by the journals included in this study. Doing so enables us to identify broadly oriented or more focused oriented
journals. Note that aggregation of topic proportions is handled per journal and covers only the years for which articles are downloaded (see Table 5.1). For example, for the journal *Fish and Fisheries*, journal topic distributions cover the time range 2000–2016, whereas the journal *ICES Journal of Marine Science* covers the complete time range of 1990–2016.

5.2.9 Relaxing LDA Assumptions and Future Research Directions

At the time of writing, the original LDA method proposed by Blei et al. (2003) has over 20,000 citations. The technique has received much attention from machine learning researchers and other scholars and has been adopted and extended in a variety of ways. More concretely, relaxing the assumptions behind LDA can result in richer representations of the underlying semantic structures. The bag of words assumption has been relaxed by conditioning words on the previous words (i.e. Markovian structure) (Wallach, 2006b); the document exchangeability assumption, relaxed by the previously mentioned dynamic topic model (Blei and Lafferty, 2006), and the Bayesian non-parametric model can be utilized to automatically uncover the number of topics (Whyte Teh et al., 2004). Furthermore, LDA has been extended in various ways. Topics might correlate as a topic about ‘cars’ is more likely to also be about ‘emission’ than it is about ‘diseases’. The Dirichlet distribution is implicitly independent and a more flexible distribution, such as the logistic normal, is a more appropriate distribution to capture covariance between topics. The correlated topic model aids in this task (Blei and Lafferty, 2007). Other examples extending LDA include the author-topic model (Rosen-Zvi et al., 2004), the relational topic model (Chang and Blei, 2010), the spherical topic model (Reisinger et al., 2010), the sparse topic model (Wang and Blei, 2009), and the bursty topic model (Doyle and Elkan, 2009). Apart from its applicability to text, LDA can be applied to audio (Kim et al., 2009), video (Mehran et al., 2009), and image (Fergus et al., 2005) data. Those topic models that relax or extend the original LDA model bring additional computational complexity and their own sets of limitations and challenges; nevertheless, it would be interesting to explore these models in future research.

5.3 Results and Discussion

5.3.1 Uncovering Fisheries Topics

The LDA model with the optimal coherence score contains 25 topics \( k = 25 \). The ten most probable words (i.e. the words with the highest probabilities), together with the semantically attached label for each uncovered latent topic, are shown in Table 5.2. The manually assigned labels for the 25 topics are: (1) Conservation, (2) Morphology,
(3) Salmon, (4) Reproduction, (5) Non-fish species, (6) Coral reefs, (7) Biochemistry, (8) Freshwater, (9) Diet, (10) North Atlantic, (11) Southern hemisphere, (12) Development, (13) Genetics, (14) Assemblages, (15) Growth experiments, (16) Stock assessment, (17) Growth, (18) Tracking and movement, (19) Fishing gear, (20) Primary production, (21) Models, (22) Salmonids, (23) Acoustics and swimming, (24) Estuaries, and (25) Fisheries management. These 25 topics can be grouped into overarching themes: aquatic organisms biology (n = 11), specific aquatic organisms (n = 4); aquatic habitats (n = 3), geographical areas (n = 2), modeling (n = 2), management (n = 2), and fishing technology (n = 1). A visual representation of the topics, their proportions within the complete corpus, and their grouping in overarching themes can be found in Fig. 5.2.
Table 5.2: Table showing the 25 uncovered topics from 46,582 fisheries science articles published in 21 fisheries-specialized journals in the period 1990–2016. Each topic displays the ten most probable words (i.e. words with the highest probability). The topics are manually labeled with a logical topic description (top) that best captures the semantics of the top words, with the topic ID in parentheses.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Label</th>
<th>Top-10 words</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Morphology</td>
<td>Specimen, Mm, Body, Dorsal, Morphological, Length, Shape, Head, Form, Morphology</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>3</td>
<td>Salmon</td>
<td>Salmon, Chinook, Chinook Salmon, Pacific, River, Columbia, Fish, Year, Stock, Juvenile</td>
<td>Specific Aquatic Organisms</td>
</tr>
<tr>
<td>4</td>
<td>Reproduction</td>
<td>Female, Male, Sex, Size, Reproductive, Shark, Spawn, Mature, Maturity, Oocyte</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>5</td>
<td>Non-Fish Species</td>
<td>Sediment, Crab, Site, Mussel, Seagrass, Density, Treatment, Effect, Plant, Shell</td>
<td>Specific Aquatic Organisms</td>
</tr>
<tr>
<td>6</td>
<td>Corals</td>
<td>Reef, Coral, Site, Habitat, Area, Community, Abundance, Colony, Island, Depth</td>
<td>Aquatic Habitats</td>
</tr>
<tr>
<td>7</td>
<td>Biochemistry</td>
<td>Concentration, Cell, Tissue, Acid, Protein, Lipid, Level, Sample, Activity, Exposure</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>8</td>
<td>Freshwater</td>
<td>Lake, Fish, Bass, Reservoir, Angler, Largemouth, Walleye, Population, Year, Perch</td>
<td>Aquatic Habitats</td>
</tr>
<tr>
<td>9</td>
<td>Diet</td>
<td>Prey, Diet, Food, Predator, Size, Feed, Fish, Trophic, Value, Consumption</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>10</td>
<td>North Atlantic</td>
<td>Sea, Cod, Area, North, Lamprey, Fish, Arctic, Herring, Parasite, Baltic</td>
<td>Geographical Areas</td>
</tr>
<tr>
<td>11</td>
<td>Southern Hemisphere</td>
<td>Sea, Area, Water, Region, Island, Shelf, Whale, Temperature, South, Depth</td>
<td>Geographical Areas</td>
</tr>
<tr>
<td>12</td>
<td>Development</td>
<td>Egg, Larval, Larvae, Spawn, Stage, Temperature, Larva, Early, Day, Hatch</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>13</td>
<td>Genetics</td>
<td>Population, Genetic, Sample, Analysis, Individual, Gene, Allele, Dna, Loci, River</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>14</td>
<td>Assemblages</td>
<td>Habitat, River, Site, Stream, Water, Flow, Area, Channel, Fish, Reach</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>15</td>
<td>Growth Experiments</td>
<td>Fish, Temperature, Treatment, Experiment, Effect, Water, Rate, Control, Tank, Test</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>16</td>
<td>Stock Assessment</td>
<td>Year, Population, Stock, Mortality, Rate, Recruitment, Estimate, Model, Biomass, Change</td>
<td>Modeling</td>
</tr>
<tr>
<td>17</td>
<td>Growth</td>
<td>Growth, Otolith, Length, Fish, Sturgeon, Sample, Mm, Size, Growth Rate, Rate</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>18</td>
<td>Tracking and Movement</td>
<td>Fish, Tag, River, Release, Dam, Movement, Hatchery, Migration, Survival, Rate</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>19</td>
<td>Fishing Gear</td>
<td>Catch, Fishing, Fishery, Fish, Gear, Net, Trawl, Size, Hook, Vessel</td>
<td>Fish Technology</td>
</tr>
<tr>
<td>20</td>
<td>Primary Production</td>
<td>Concentration, Water, Rate, Nutrient, Phytoplankton, Production, Sediment, Carbon, Chl, Sample</td>
<td>Specific Aquatic Organisms</td>
</tr>
<tr>
<td>21</td>
<td>Models</td>
<td>Model, Estimate, Value, Variable, Parameter, Analysis, Effect, Distribution, Base, Sample</td>
<td>Modeling</td>
</tr>
<tr>
<td>22</td>
<td>Salmonids</td>
<td>Trout, Fish, Stream, Rainbow, Population, Brook, Cutthroat, Creek, Brown, Salmonid</td>
<td>Specific Aquatic Organisms</td>
</tr>
<tr>
<td>23</td>
<td>Acoustics and Swimming</td>
<td>Depth, Fish, Velocity, Water, Speed, Distance, Vertical, Surface, Sound, Night</td>
<td>Aquatic Organism Biology</td>
</tr>
<tr>
<td>24</td>
<td>Estuaries</td>
<td>Fish, Estuary, Water, Bay, Salinity, Estuarine, Area, Freshwater, Habitat, River</td>
<td>Aquatic Habitats</td>
</tr>
<tr>
<td>25</td>
<td>Fisheries Management</td>
<td>Fishery, Management, Fishing, State, Resource, Economic, Vessel, Policy, Area, Fish</td>
<td>Management</td>
</tr>
</tbody>
</table>
Conditioning the topics on the word 'fishery' or 'fisheries' (i.e. taking into consideration the probability assigned to this word), these 25 topics can be divided into four groups, the first two of which we considered to be directly relating to fisheries: using the word often \( (n = 3) \), moderately \( (n = 5) \), infrequently \( (n = 8) \) or almost not at all \( (n = 9) \) (Fig. 5.3). The topics using this word often are, in descending order: (25) Fisheries management, (19) Fishing gear, and (16) Stock assessment. Almost one-fifth of all the topics are using the word 'fishery' or 'fisheries' moderately: (1) Conservation, (3) Salmon, (4) Reproduction, (8) Freshwater, (24) Estuaries. One third of all the topics are using the word 'fishery' or 'fisheries' infrequently: (2) Morphology, (7) Biochemistry, (10) North Atlantic, (13) Genetics, (15) Growth experiments, (17) Growth, (18) Tracking and moving, (22) Salmonids. Another one-third of all the topics does not use this word almost at all: (5) Non-fish species, (6) Corals, (9) Diet, (11) Southern Hemisphere, (12) Development, (14) Assemblages, (20) Primary production, (21) Models, (23) Acoustics and swimming.

Considering all the 25 topics, only one (i.e. (25) Fisheries management) refers explicitly to the human dimension component of the fishery system, something that confirms our working hypothesis that the human dimension of fisheries is under-represented in the fisheries specialty literature. To evaluate whether the human dimension of fisheries as SECASs is further refined within the Fisheries management topic, following the same methodology as described above, we created a new LDA model that zooms in on this topic, thereby creating subtopics from the broader Fisheries management topic. The new model uncovered 12 subtopics from the topic Fisheries management (Table 5.3), out of which eight assign higher probability to the term 'fishery' or 'fisheries' (i.e. use this word often or moderately) and, thus, were considered directly relating to fisheries: three on various management approaches (i.e. Co-management, Precautionary approach, Quota systems); three on economics (i.e. Markets, Bioeconomics, Blue economy); and two on type of fishery (i.e. Small scale fisheries, Recreational fisheries).

Out of the total of 25 topics uncovered by our analysis, three were considered generic (i.e. (10) North Atlantic, (11) Southern hemisphere, (21) Models)). Out of the remaining 22 topics, 20 cover the natural dimension of fisheries, reasonably mirroring the curriculum of fisheries biology and fisheries ecology higher education courses (e.g. Jennings et al., 2009; King, 2007)), but not addressing topics such as climate change. However, considering the focus of the two remaining topics, i.e. (1) Conservation and (25) Fisheries management, it is apparent that the research focus in fisheries during the last 26 years has not entirely captured the complexity of the fisheries domain, especially of the human dimension component, something also observed for specialized research areas such as, for example, by-catch reduction technology (Campbell and Cornwell, 2008; Molina and Cooke, 2012). These results seem to be confirmed by the bibliometric analyses published in (Aksnes and Browman, 2016) and (Jarić et al., 2012), where no human dimension related words were identified among the most frequent words used in fisheries publication titles and abstracts. This situation might not be surprising given the institutional context in which fisheries research is performed. For example, within the International Council for the Exploration of the Sea (ICES), which is one of
Figure 5.2: Inter-topic distance map that shows a two-dimensional representation (via multidimensional scaling) of the 25 uncovered fisheries topics. The distance between the nodes represents the topic similarity with respect to the distributions of words (i.e., LDA's output). The surface of the nodes indicates the topic prevalence within the corpus, with bigger nodes representing topics being more prominent within the document collection (all nodes add up to 100%).
Figure 5.3: Inter-topic distance map showing topics conditioned on the word 'fishery' (including 'fisheries'). The figure is similar to Fig. 5.2 but expresses the differences in probability assigned to the word 'fishery'. Bigger nodes place higher probability to the word 'fishery' and can be considered more directly related to fisheries science.
Table 5.3: Table showing the 12 uncovered sub-topics from the documents (n = 3,390) dealing with the topic Fisheries management. The sub-topics provide a “zoomed-in” view of the topical decomposition from the subset of documents on fisheries management. Each topic displays the ten most probable words and the semantic label that best describes the underlying latent topic.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Label</th>
<th>Top-10 words</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spatial Planning</td>
<td>Marine, Policy, Stakeholder, Process, Coastal, Development, Sea, Plan, Environmental, Regional</td>
<td>Non-Fisheries</td>
</tr>
<tr>
<td>2</td>
<td>Markets</td>
<td>Price, Market, Fish, Product, Production, Model, Value, Seafood, Estimate, Sector</td>
<td>Economics</td>
</tr>
<tr>
<td>3</td>
<td>Bioeconomics</td>
<td>Cost, Model, Stock, Fishery, Effort, Value, Scenario, Harvest, Fish, Rate</td>
<td>Economics</td>
</tr>
<tr>
<td>4</td>
<td>Conservation and MPA</td>
<td>Marine, Mpa, Conservation, Protect, Ocean, Ecosystem, Mpas, Protection, Sea, Habitat</td>
<td>Non-Fisheries</td>
</tr>
<tr>
<td>5</td>
<td>Small-Scale Fisheries</td>
<td>Fishing, Fisher, Fish, Study, Catch, Fishery, Boat, Fisherman, Local, Community</td>
<td>Type of Fishery</td>
</tr>
<tr>
<td>6</td>
<td>Blue Economy</td>
<td>Marine, Fish, Aquaculture, Fishery, Shark, Water, Coastal, Development, Production, Specie</td>
<td>Economics</td>
</tr>
<tr>
<td>7</td>
<td>Pollution</td>
<td>Ship, Vessel, Oil, Port, Shipping, Pollution, Risk, International, Trade, Country</td>
<td>Non-Fisheries</td>
</tr>
<tr>
<td>8</td>
<td>Legislation</td>
<td>Sea, Law, International, Convention, Agreement, Country, China, Water, Coastal, Maritime</td>
<td>Non-Fisheries</td>
</tr>
<tr>
<td>9</td>
<td>Co-Management</td>
<td>Fishery, Community, Social, Fishing, System, Right, Group, Fisher, Government, Local</td>
<td>Management Approaches</td>
</tr>
<tr>
<td>10</td>
<td>Quota Systems</td>
<td>Fishery, Vessel, Fishing, Catch, Quota, Fish, Fleet, System, Total, Stock</td>
<td>Management Approaches</td>
</tr>
<tr>
<td>11</td>
<td>Precautionary Approach</td>
<td>Fishery, Stock, Specie, Catch, Fishing, Datum, Assessment, Fish, Whale, Ecosystem</td>
<td>Management Approaches</td>
</tr>
<tr>
<td>12</td>
<td>Recreational Fisheries</td>
<td>Angler, Recreational, Fish, Fishing, Survey, Respondent, Catch, Fishery, Estimate, Value</td>
<td>Type of Fishery</td>
</tr>
</tbody>
</table>
the most important fisheries-related intergovernmental organizations, despite having along the years various groups working more or less directly with different aspects of this human dimension, now only one out of more than 45 expert groups is working explicitly with this dimension of fisheries. This group, the Strategic Initiative Human Dimension (SIHD), became operational in 2015.

Persisting in having this heavily unbalanced focus between the two dimensions of fisheries systems (i.e. the human dimension (i.e. human agents, communities of these, and their institutions) and the natural dimension (i.e. biotic, such as predator species and prey species, and abiotic, such as water temperature and nutrients)) will not help in understanding the behavior of fisheries stakeholders (from fishers to consumers), leading to unintended, and too often undesirable, management outcomes (Fulton et al., 2011), and thus unsustainable fisheries. Responding to the challenges posed by sustainable fisheries necessitates the development of stronger networks within the family of human dimension sciences and across disciplinary boundaries with the natural dimension sciences (Symes and Hoefnagel, 2010a). Without providing an exhaustive list and in random order, the human dimension in fisheries could be included in fisheries science by addressing topics such as: institutional aspects (enforcement and compliance, policy interactions etc.), social aspects (gender, religion/beliefs, welfare, social cohesion, social networks, education and learning, human agency, health, safety and security at sea, food security, perception, attitudes, social norms, compliance, mental models of various actors involved in fisheries etc.), economic aspects (poverty, innovation, distribution of benefits, spiritual, inspirational, and aesthetic services of fisheries etc.), political aspects (power structures, transparency etc.), and cultural aspects (traditional/local ecological knowledge, history, cultural dimensions, culinary choices, heritage, blue humanities, fisheries literacy etc.) (Charles, 2000; De Young et al., 2008; ICES, 2016; Österblom et al., 2013; Sowman, 2011; Spalding et al., 2017; Stone-Jovicich, 2015).

Continuing our analysis, three topic clusters can be identified in Fig. 5.2, indicating a similar probability distribution over words (i.e. topics that are, to some extent, related to the words they use to describe the theme): a growth cluster (the topics Growth experiments (15), Diet (9), Non-fish species (5), Primary production; Development (12), Reproduction (4), and Growth (17)); an institutions cluster (the topics Fisheries management (25) and Conservation (1)); and a salmonids cluster (the topics Freshwater (8), Tracking and movement, and Salmonids (22); one would expect to find here also the topic Salmon (3), but, interestingly from a linguistics point of view, this topic seems rather isolated). The most isolated topics are Morphology (2) and Biochemistry (7), indicating most probably the use of a very specific topic distribution over words.

The most frequent aquatic organisms mentioned in our corpus are salmonids (e.g. salmon, trout) and other freshwater organisms (e.g. perch); shark (within the topic Reproduction); crab, mussel, and oyster (within the topic Non-fish species); cod, lamprey, and herring (within the topic North Atlantic); whale (within topic Southern hemisphere); sturgeon (within the topic Growth); tuna (within the topic Fishing gear); and
shrimp (within topic Estuaries). Commercially important species, such as anchoveta, pollock, and tilapia, were not included among the most frequent words of any of the 25 topics. These findings are relatively consistent with Aksnes and Browman (2016) and Jarić et al. (2012), who reported that the most frequently studied group of species was the Salmonidae, followed by the Atlantic cod, and that there is no correlation between the production of various species and the number of publications about these species.

Aquaculture has not been identified as a topic of its own in our dataset, and the word aquaculture was not included in the top 10 most frequent words of any of the 25 topics. However, the word aquaculture was included in the top 10 most frequent words for the sub-topic Blue economy under the topic Fisheries management, possibly indicating the interest in this relatively new industry in a context that focuses on increasing economic activities in the marine and maritime sector.

With regard to the typology of fishers by Charles (2000), only the recreational type is specifically mentioned among the most frequent words in our corpus, with the word angler included in the topic (8) Freshwater. This might be because the research focusing on the other types (e.g. subsistence, artisanal) does not employ a very specific vocabulary, or that there might be a lag in research on, for example, small-scale and artisanal fishery (Purcell and Pomeroy, 2015). However, if we look only at topic (25) Fisheries management, the recreational type and small-scale type have each its subtopic, indicating that, from a management perspective, these two types of fisheries have been relatively extensively explored by fisheries scientists.

Out of the 25 topics uncovered by our LDA model, two refer to large geographical areas: the North Atlantic (10) and Southern hemisphere (11). The words Norwegian (within the topic North Atlantic) and Florida (within the topic Estuaries (25)), are the only specific geographic references among the top 10 most probable words. These very few specific geographic references might indicate that most of the fisheries research is focused on a few areas around the globe, leaving large zones underexplored, as also indicated in Molina and Cooke (2012), or that research about other regions is published in other languages than English.

5.3.2 Topic Proportions within Documents

For every document, the LDA model infers the topical decomposition, indicating which topics are found in that document and in what proportions. The assumptions behind LDA causes documents to exhibit mainly a small number of main topics, with other topics very close to zero (note that all topics per document sum up to 1). This structure assumes that documents are often about some topics, rather than being about all topics equally. Following this line of reasoning, we analyzed the remaining topic proportions for documents exhibiting one of the topics as the dominant topic, this being defined as the document’s topic proportion that exceeds all other topic proportions. Such an anal-
Fig. 5.4 shows the average remaining topic proportions with respect to the dominant topic displayed as a heat map. The dominant topic, shown on the left (i.e. rows), has the average dominant topic proportion within parentheses. A higher number indicates that more of the document content deals with the dominant topic, while the other topics make up the smaller remaining portion of the document. Conversely, lower numbers reflect a dominant topic making up a smaller portion of the document, leaving more room for other topics to be part of that document. For example, documents dealing with the topic Fishing gear are on average allocating 45% of their content to their own topic, leaving 55% of the remaining content to other topics (for example, 9% to the topic Models and between 4 and 6% to each of the topics Fisheries management, Acoustics and swimming, Stock assessment, and Southern hemisphere). Furthermore, the remaining average topic proportions, shown at the top (i.e. columns), are sorted from high to low and reflect which topics more frequently, or to a higher extent, co-occur with other topics. For example, given any document with a dominant topic, the remaining topic proportion deals more often with the topics Models, Stock assessment, Fish reproduction, or Growth experiments (i.e. these three topics have the highest co-occurrence), and less often with the topics Salmonids, Genetics, or Salmon (i.e. these three topics have the lowest co-occurrence).

Documents dealing with the most prevalent topic in the corpus directly relating to fisheries, i.e. Fisheries management, allocate to this topic almost 60% of their content, about 13% to Conservation, and between 3 and 5% to topics such as (in descending order): Fishing gear, Models, and Stock assessment. Documents dealing with the most prevalent topic in the rest of the corpus, i.e. Primary production, allocate to this topic 55% of their content, and between 3 and 6% to topics such as (in descending order): Non-fish species, Biochemistry, Models, Diet, Southern hemisphere, and Growth experiments. Co-occurrence of such topics might be something natural. However, it would be interesting to consider whether other mixtures of topics would bring novel, and possibly also innovative, insight into fisheries science. Such an insight is highly needed in order to achieve sustainable fisheries exploitation and implement the fisheries-related actions of the international ocean governance objectives (European Commission, 2016a).
FIGURE 5.4: Heat map displaying the dominant topic (left) and the remaining average topic proportions (top) for all 46,582 documents from the corpus. This indicates the extent to which documents about one main topic relate to the other uncovered topics (i.e. the degree of topic co-occurrence), decreasing from left to right. For example, documents that primarily focus on Fisheries management also focus on Conservation (12.7%) or Fishing gear (5.3%).
25 broad fisheries topics, as well as the topical trends and prevalence for the Fisheries management sub-topics, can be found in Figs. 5.7, 5.8, 5.9 in the Appendix.

With regard to the groups of topics described in Table 5.2, during the entire period of 26 years, taken together, the 11 topics referring to aquatic organisms biology themes made up 40% of the corpus; the groups of topics on the themes of specific species, modeling, management, and habitats, between 11% and 14% each; the two topics on the theme of geographical regions, 7% together; and the fishing gear topic, almost 5%. The topics Models, Primary production, and Fisheries management are the most prevalent in the corpus during the entire period of 26 years, with approximately 7, 6.5, and 6% respectively from the total number of documents. The topics Salmon, Salmonids, and Morphology are the least prevalent, with only around 2% of the corpus accounting for each. The topic Models has been among the top three most prevalent topics since 1995. Primary production was the most prevalent topic in the first three time periods we analyzed (from 1990 to 2005). Topics directly relating to fisheries, such as Fisheries management and Stock assessment, were among the top 3 of most prevalent topics in the period 2005–2010 (both topics), and 2010–2016 (only Fisheries management). The topic Conservation joined this top during the last 6 years.

Taken together, topics that could be grouped as relating to freshwater fisheries (i.e. Freshwater, Tracking and movement, Salmonids, Salmon, Estuaries, Assemblages) account for one-sixth of the corpus. These results, corroborated with the declining interest in these topics in the last 16 years (see Fig. 5.5), confirm the findings of Jarić et al. (2012) that indicate a general decline in the frequency of studies focused on freshwater habitats.

The top four hottest topics of the last 26 years (overall column) are (in descending order): Fisheries management, Conservation, Models, and Fishing gear (with Stock assessment, the third topic directly relating to fisheries, in 7th place). The interest in models has also been confirmed by Jarić et al. (2012), and this, in addition to the prevalence of the modeling topics, which was described above, provides empirical evidence for the fact that modeling is one of the most important research methods in fisheries science (Angelini and Moloney, 2007). Topic Fisheries management, the third most prevalent topic in the corpus, remained among the top three hottest topics during the last 11 years, while it was the coldest topic in the first half on the 1990s.

The configuration of top three hottest and top three coldest topics has fluctuated during the five time periods we have analyzed. However, the topic Primary production has constantly been ranked among the coldest topics (with the topic Biochemistry also being in this top in the period 1990–2005). Overall the 26 year period, it is interesting to note that Primary production was the most prevalent topic, but at the same time had the steepest decline in interest. The topics Models and Conservation have been a constant occurrence in the top three hottest topics since the turn of the Century, whereas the topic Fisheries management only joined this top category in 2005. Another topic directly relating to fisheries, Fishing gear, was part of this top in the period 2000–2005.
CHAPTER 5. TOPIC ANALYSIS OF FISHERIES SCIENCE

Figure 5.5: Topical trends and prevalence for all 25 uncovered fisheries topics displayed as a heat map. The topical trends (left) show the increasing/hot (red) and decreasing/cold (blue) topics for different time intervals. The value represents the fitted (via linear regression) change in percentage points within each time interval. The topic prevalence (right) shows the average cumulative topic proportion in percentage for different time intervals for each of the 25 fisheries topics. It shows how present a topic is within a certain time interval given all the scientific output within that time interval. Individual trend lines per year can be found in the Fig. 5.7 in the Appendix.
The increased prevalence in the corpus of the topic Fisheries gear and the constant interest shown in this topic, together with the constant prevalence in the corpus of the topic Stock assessment and the constant interest shown in this topic, indicate that fishing gear and stock assessment have been the central elements of fisheries science in the last 26 years. The increasing prevalence and interest in the topic Fisheries management might indicate the strengthening of the connection between fisheries science and management processes, in the light of the growing concern about the status of fish stocks worldwide. Stock assessments provide a scientific and quantitative basis to the process of developing and implementing a management plan (Hoggarth et al., 2005) and, as such, are indispensable to management processes (Hoggarth et al., 2006). The interest in the topics Growth and Reproduction seems to be most stable among all the 25 topics when looking at the entire period of 26 years, even though the prevalence of these topics is rather small (around 3%). The constant interest in the latter can be explained by the importance of fisheries reproduction for fisheries assessment and management (Jakobsen et al., 2016).

5.3.4 Topical Trends over Journals

Although many journals included in our analysis overlap to some extent in their content, it is possible to identify journals that seem specialized in specific topics (Fig. 5.6). For example, almost one-fifth of the publications that appeared in the journal Fish and Fisheries relate to the topic Fisheries management, whereas another approximately one-fifth is related to the topic Conservation. Among the topics directly relating to fisheries, the journals Marine Policy and Marine Resource Economics are highly specialized in the topic of Fisheries management, and the journals Fisheries research and CCAMLR Science, in the topic Fishing gear. It appears that no journal is highly specialized in Stock assessment, with this topic being addressed by almost all the journals. The top three journals publishing this topic are ICES Journal of Marine Science, Fisheries Oceanography, and Fish and Fisheries, with 10-11% of the publication space of each of these journals covering this topic.
## Figure 5.6: Topic distribution over journals displayed as a heat map. For each included journal (left), the coverage of topics (top) in percentages is displayed (percentage values in cells, row total = 100%). The heat map displays which journals publish which topics and in what proportions. See Table 5.1 for journal year coverage, as this differs between journals (e.g. 1990–2016, 1997–2016, 2000–2016).
5.3.5 Validation of Results

We validated the output of the LDA model (including its labeling) by comparing the hot/cold LDA topics in the period 2000–2009 with the hot/cold words used in publication titles in the same period identified by the bibliometric study of Jarić et al. (2012). Some of the words that they identified as having the greatest increase in frequency can be directly linked to some of the hot topics identified by our analysis, e.g. by-catch, longline, seabird in the bibliometric study relate to the topic Fishing gear (19) identified by our study; genetic relates to the topic Genetics (13). Likewise, some of the words with the largest decrease in frequency can be directly linked to some of the cold topics identified by the LDA analysis over the same period, e.g. Atlantic in the bibliometric study relates to the topic North Atlantic identified by our study; growth, to the topics Growth (17) and Growth experiments (15); recruitment, to the topic Stock assessment (14); feeding, to the topic Diet (9).

5.4 Conclusion and Recommendations

From the analysis of more than 46,500 full-text articles published in 21 top fisheries journals, it is apparent that, during the last 26 years, the research focus of fisheries science has been predominately on the natural dimension of the fisheries system, with 22 out of 25 topics referring to this dimension. While the natural dimension of fisheries was split into various aspects covering topics from specific species to fish catch technology, the human dimension was explicitly expressed only through one, albeit the hottest topic in the data set and the second most prevalent: Fisheries management. Although there is undoubtedly some scientific production addressing various human dimension of fisheries, it could be that the narrative used to describe the human dimension is not explicit enough to be captured by word co-occurrence, or that the human dimension is not prevalent enough to be recognized as a general topic or specific subtopic by the LDA model. Additionally, it might be that the scientific production on the human dimension is published in journals other than those specialized in fisheries or other types of outlets, such as books and book chapters. We could advance various hypotheses as to why this might be the case (e.g. most fishery scientists are biological/natural sciences trained and oriented (Link, 2010); those who are non-natural sciences trained and oriented tend to publish in outlets that foster recognition from within their own scientific communities, such as books, rather than journals), but this exercise would be outside the scope of this study. Instead, we want to emphasize two important recommendations: 1. Diversification of the scientific focus so that it covers more of the complexity of fisheries, especially the human dimension (funding bodies play a crucial role here by setting the research agenda); and 2. Publication of fisheries research in outlets more likely to reach the intended audience (i.e. top interdisciplinary journals or specialized fisheries journals, if the objective of publishing is to contribute to fisheries sustainability by reaching fisheries policy- and decision-makers). A lack of interdisci-
Plenary synthesis in research is one of the major factors in fisheries collapses (Smith and Link, 2005). Thus, more integrative research and research focused on the underrepresented topics might provide insight into the fine mechanisms of fisheries as socio-ecological complex adaptive systems, and, thus, a critical input for developing successful fisheries management approaches.
Appendix

Calculating Topical Trends over Time

For every topic \( k = \{1, \ldots K\} \) and for every year \( y = \{1990, \ldots 2016\} \), we calculate the composite topic-year proportion \( \hat{\phi}_{y,k} \) as expressed in Equation 5.1.

\[
\hat{\phi}_{y,k} = \frac{\sum_{d \in y} \phi_{d,k}}{n_y}
\]  

(5.1)

\( \hat{\phi}_{y,k} \) aggregates for every document \( d \) of year \( y \) the \( k^{th} \) topic proportion \( \phi_{d,k} \) and divides it by the number of documents for that year \( n_y \) to obtain averages.
### Table 5.4: A selection of publication titles that have high probability for a single topic.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Publication title</th>
<th>Year</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Biodiversity, ecosystems and coastal zone management: linking science and policy</td>
<td>Biodiversity, ecosystems and coastal zone management: linking science and policy</td>
<td>2011</td>
<td>82%</td>
</tr>
<tr>
<td>2. Description of Careproctus patagonicus sp. nov. and C. magellanicus sp. nov. (Pisces: Scorpaeniformes) from the lower slope of Drake Passage</td>
<td>Description of Careproctus patagonicus sp. nov. and C. magellanicus sp. nov. (Pisces: Scorpaeniformes) from the lower slope of Drake Passage</td>
<td>2000</td>
<td>74%</td>
</tr>
<tr>
<td>3. Magnitude and Trends in Abundance of Hatchery and Wild Pink Salmon Chum Salmon and Sockeye Salmon in the North Pacific Ocean</td>
<td>Magnitude and Trends in Abundance of Hatchery and Wild Pink Salmon Chum Salmon and Sockeye Salmon in the North Pacific Ocean</td>
<td>2010</td>
<td>80%</td>
</tr>
<tr>
<td>4. Fishery self-governance in fishing communities of South Korea</td>
<td>Fishery self-governance in fishing communities of South Korea</td>
<td>2015</td>
<td>96%</td>
</tr>
<tr>
<td>5. The role of salinity in structuring the fish assemblages in a tropical estuary</td>
<td>The role of salinity in structuring the fish assemblages in a tropical estuary</td>
<td>2005</td>
<td>75%</td>
</tr>
<tr>
<td>7. Drift algal-epiphyte-seagrass interactions in a subtropical Thalassia testudinum meadow</td>
<td>Drift algal-epiphyte-seagrass interactions in a subtropical Thalassia testudinum meadow</td>
<td>2009</td>
<td>91%</td>
</tr>
<tr>
<td>8. Effect of silicate on particulate carbon production in phytoplankton</td>
<td>Effect of silicate on particulate carbon production in phytoplankton</td>
<td>2003</td>
<td>75%</td>
</tr>
<tr>
<td>10. Impact of density-dependent growth and maturation on assessment advice to rebuild depleted U.S. silver hake (Merluccius bilinearis) stocks</td>
<td>Impact of density-dependent growth and maturation on assessment advice to rebuild depleted U.S. silver hake (Merluccius bilinearis) stocks</td>
<td>2011</td>
<td>91%</td>
</tr>
<tr>
<td>11. Utility of Alligator Gar Age Estimates from Otoliths Pectoral Fin Rays and Scales</td>
<td>Utility of Alligator Gar Age Estimates from Otoliths Pectoral Fin Rays and Scales</td>
<td>2012</td>
<td>86%</td>
</tr>
<tr>
<td>12. Assessing the performance of linear geostatistical tools applied to artificial fisheries data</td>
<td>Assessing the performance of linear geostatistical tools applied to artificial fisheries data</td>
<td>2006</td>
<td>93%</td>
</tr>
<tr>
<td>13. In situ swimming behaviour of individual mesopelagic fish studied by split-beam echo target tracking</td>
<td>In situ swimming behaviour of individual mesopelagic fish studied by split-beam echo target tracking</td>
<td>2001</td>
<td>87%</td>
</tr>
<tr>
<td>14. Geographical and typological changes in fish guilds of South African estuaries</td>
<td>Geographical and typological changes in fish guilds of South African estuaries</td>
<td>2005</td>
<td>87%</td>
</tr>
<tr>
<td>15. Interpreting Stream Physical Characteristics in Index of Biotic Integrity Classifications: General Similarities and Specific Differences</td>
<td>Interpreting Stream Physical Characteristics in Index of Biotic Integrity Classifications: General Similarities and Specific Differences</td>
<td>2009</td>
<td>81%</td>
</tr>
<tr>
<td>16. Influence of silicate on particulate carbon production in phytoplankton</td>
<td>Influence of silicate on particulate carbon production in phytoplankton</td>
<td>2003</td>
<td>85%</td>
</tr>
<tr>
<td>17. Oxygen consumption, ammonia excretion and protein use in response to thermal changes in juvenile Atlantic salmon Salmo salar</td>
<td>Oxygen consumption, ammonia excretion and protein use in response to thermal changes in juvenile Atlantic salmon Salmo salar</td>
<td>2005</td>
<td>84%</td>
</tr>
<tr>
<td>18. Competition and Predation as Mechanisms for Displacement of Greenback Cutthroat Trout by Brook Trout</td>
<td>Competition and Predation as Mechanisms for Displacement of Greenback Cutthroat Trout by Brook Trout</td>
<td>2007</td>
<td>86%</td>
</tr>
</tbody>
</table>
Figure 5.7: Topic prevalence for each of the 25 uncovered fisheries science topics. The x-axis represents the year with the y-axis displaying the prevalence of the topic within all the documents published in that year. The prevalence indicates the cumulative topic proportion relative to all the documents published in that year. For example, in 2016, almost 9% of the papers published in this year cover aspects of fisheries management.
Figure 5.8: Topical trends and prevalence for all 12 uncovered fisheries management sub-topics displayed as a heat map. The topical trends (left) show the increasing/hot (red) and decreasing/cold (blue) topics for different time intervals. The value represents the fitted (via linear regression) change in percentage points within each time interval. The topic prevalence (right) shows the average cumulative topic proportion in percentage for different time intervals for each of the 12 fisheries management sub-topics. It shows how present a sub-topic is within a certain time interval given all the scientific output within that time interval for documents relating primarily to the broad topic of fisheries management.
Figure 5.9: Topic prevalence for each of the 12 uncovered fisheries management sub-topics. The x-axis represents the year with the y-axis displaying the prevalence of the topic within all the documents within that year relating primarily to fisheries management. The prevalence indicates the cumulative topic proportion relative to all the fisheries management documents published in that year.
Chapter 6

Using Machine Learning to Uncover Latent Research Topics in Fishery Models

Modelling has become the most commonly used method in fisheries science, with numerous types of models and approaches available today. The large variety of models and the overwhelming amount of scientific literature published yearly can make it difficult to effectively access and use the output of fisheries' modelling publications. In particular, the underlying topic of an article cannot always be detected using keyword searches. As a consequence, identifying the developments and trends within fisheries modelling research can be challenging and time-consuming. This paper utilises a machine learning algorithm to uncover hidden topics and subtopics from peer-reviewed fisheries modelling publications and identifies temporal trends using 22,236 full-text articles extracted from 13 top-tier fisheries journals from 1990–2016. Two modelling topics were discovered: estimation models (a topic that contains the idea of catch, effort, and abundance estimation) and stock assessment models (a topic on the assessment of the current state of a fishery and future projections of fish stock responses and management effects). The underlying modelling subtopics show a change in the research focus of modelling publications over the last 26 years.

This work was originally published as:

CHAPTER 6. SUB-TOPIC ANALYSIS OF FISHERY MODELS

6.1 Introduction

Global research efforts have increased significantly in recent years (Oecd, 2008), as has publication output within fisheries science (Aksnes and Browman, 2016). This growth has been partly driven by growing concerns about the state of fish stocks and the need to provide information for policy and decision makers globally. Since each fish stock is typically unique, and experimental approaches cannot be used to predict their response to fishing, it follows that the modelling and simulation of fisheries play a major role in providing management advice; these are among the most frequently used methods in fisheries science (Jarić et al., 2012). Models offer a feasible approach to the approximation of trends and processes, and they advance the understanding of fisheries and ecosystem dynamics (Angelini and Moloney 2007) while guiding data collection and illuminating core uncertainties (Epstein, 2008). For this reason, and in contrast to common perceptions, a multitude of fisheries models is available besides standard stock assessment models, and these models take on many different shapes and forms depending on their method and purpose. They may include individual-based models to investigate fleet behaviour (Bastardie et al., 2014); Bayesian belief networks to better understand stakeholder viewpoints and perceptions (Haapasaaari et al., 2012); or conceptual models to analyse fisheries from a socio-ecological complex adaptive system perspective (Ostrom, 2009; Partelow, 2015).

The frequent use of models and their wide range of applications, in combination with the growing global collections of scholarly literature, have led to an ever-increasing number of publications on the various types of models and approaches. As a result, scientists are suddenly faced with millions of publications, overwhelming their capacity to effectively use these collections and to keep track of new research (Larsen and von Ins, 2010). Online collections can be browsed and explored using keyword searches, through which publications can be collected manually; however, in addition to being time-consuming, the size and growth of the body of research often has the effect of limiting the possibility of identifying all the relevant literature. Another problem is that the underlying topic of an article is not readily available in most collections. Thus, the topic of an article—that is, the idea underlying the article, which may be shared with similar articles—cannot always be detected using keyword searches (Srivastava and Sahami, 2009). Given such challenges, an assessment of the field of fisheries models could reveal overlooked research topics, identify important changes in research directions (i.e., trends), assess the diversity of topics in publication outlets, and ultimately help in identifying new and emerging modelling topics. Furthermore, an improved understanding of fisheries modelling approaches could help researchers to more easily synthesise historical and current research developments.

The developments and trends in fisheries science and fishery models are usually assessed through reviews (e.g., (Bjørndal et al., 2004; Prellezo et al., 2012) and bibliometric studies (Jarić et al., 2012; Aksnes and Browman, 2016)). These types of studies have several limitations, such as taking into account only a limited number of publica-
tions (e.g., only 61 publications, (Gerl et al., 2016)); a limited time period (e.g., from 2000-2009, (Jarić et al., 2012); a limited scope or very specialised focus (e.g., stock assessment methods, (Cadrin and Dickey-Collas, 2015); bio-economic models, (Prellezo et al., 2012); models of an ecosystem approach to fisheries, (Plagányi, 2007); and models of the Celtic Sea, (Minto and Lordan, 2014)). Other limitations include proxies for full text, such as titles (Jarić et al., 2012), and abstracts (Aksnes and Browman, 2016), and proxies for research topics, such as one word per topic (Jarić et al., 2012; Aksnes and Browman, 2016). Most importantly, previous attempts to identify trends in fisheries and fisheries modelling are based on top-down approaches, in which research topics are predefined by the researcher (Debortoli et al., 2016), such as region, species, habitat, or study area. Such approaches are prone to human subjectivity; researchers may end up with different results (Urquhart, 2001), or the mapping of text features to categories may not be explicitly known (Quinn et al., 2010).

This study aims to overcome the limitations of previous approaches by applying a bottom-up approach in which research topics automatically emerge from the statistical properties of the documents. In doing so, the topics are automatically uncovered without prior human labelling, categorisation, or predefined classification of publications, and they are thus not biased by researchers’ top-down subjective choices. For this purpose, a probabilistic topic model algorithm called latent Dirichlet allocation (LDA) (Blei et al., 2003), which belongs to the field of unsupervised machine learning algorithms, was used to reveal research topics within the field of fisheries models that are published in peer-reviewed journals and have a strong focus on fisheries. Topic model algorithms can automatically uncover hidden or latent thematic structures (i.e., topics) from large collections of documents. The unsupervised nature of LDA allows documents to “speak” for themselves, and topics emerge without human intervention. They have proven to be very useful in automatically identifying and interpreting scientific themes in relation to the journal’s existing themes or categories (Griffiths and Steyvers, 2004).

By utilising unsupervised machine learning, this study aims to provide comprehensive information on topical trends within fisheries modelling research for fisheries scientists and stakeholders. In particular, this study analyses 22,236 full-text scientific publications published within the period from 1990 to 2016 in 13 top-tier fisheries journals. Thus, a unique dataset for the field of fisheries models was created, and topics in fisheries modelling and their underlying sub-topics were identified to determine historical and current research interests. In addition, the species, areas, and methods occurring within the identified topics were assessed.
6.2 Methods

6.2.1 Latent Dirichlet Allocation

The LDA model is a generative probabilistic topic model that represents documents (i.e., fisheries publications) as discrete distributions over K latent topics; each topic is subsequently represented as a discrete distribution over all the words (i.e., vocabulary) used. The words with high probability within the same topic are frequently co-occurring words, which can be seen as clusters or constellations of words that are often used to describe an underlying topic or theme (DiMaggio et al., 2013). In this way, LDA captures the heterogeneity of research ideas or topics within publications. The topics and their relative proportions within documents are hidden (i.e., latent) variables that LDA infers from the observable variables—that is, the words within the documents. The generative process behind LDA involves an imaginary random process, through which documents are created based on probabilistic sampling rules. The topics and their proportions are subsequently inferred from these generated documents by applying statistical inference techniques, such as variational and sampling-based algorithms (Blei and Jordan, 2006; Teh et al., 2006; Hoffman et al., 2010; Wang et al., 2011). LDA extends other popular topic model algorithms such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) while also overcoming their limitations.

The LDA model makes two assumptions when analysing and uncovering latent topics from documents. First, documents are represented as “bags of words” (i.e., unordered lists of words) in which the word order is neglected. Although this is an unrealistic assumption, it is reasonable if the aim is to uncover semantic structures from text (Blei and Lafferty, 2006; Blei, 2012). Consider a thought experiment where one imagines shuffling all the words in a document. Even when shuffled, one might find words such as “population”, “size”, “virtual”, “minimum”, and “recruitment” and expect that the document deals with aspects of population dynamics. One of the core underlying principles of LDA is based on word co-occurrences, and a small number of co-occurring words is sufficient to resolve problems of ambiguity. Second, LDA assumes that the order in which documents are analysed is unimportant (i.e., document exchangeability is assumed); however, at the end of the analysis, all documents are analysed. As a result, LDA is unable to explicitly capture the evolution of topics over decades or centuries of work. This would require a more complicated and computationally expensive dynamic topic model (Blei and Lafferty, 2006), which is currently not feasible given the large dataset; however, this is a potential approach for future work. Document exchangeability is a limitation in the case of topics whose presentation in the literature has dramatically changed (e.g., in terms of the terminology used to describe the topic), but it still captures the phenomenon by which current literature builds upon previous literature. Nonetheless, the assumption of document exchangeability is especially problematic when analysing topics that span 50 to 100 years of research.
6.2.2 Topic Interpretation

The topics emerge from the statistical properties of the documents and the statistical assumptions behind LDA. The topics are represented as discrete distributions over all the words, in which the top words (e.g., top 15) for each topic—that is, the words with the highest probability and those that more frequently co-occur together—provide insights into the semantic meaning of the topic. Topics are thus a reference to these probability distributions over words to exploit text-oriented intuitions. No epistemological claims are made beyond this representation. Furthermore, by no means is the topic distribution over words limited to these top 15 words; in fact, every word occurs in every topic, but with different probabilities. The topics are used to uncover the themes prevailing the documents, as well as the extent to which such themes are present in each document. In doing so, the main ideas of a publication can be extracted and used to track how they have developed over time. Note that the underlying topics and to what extent the document exhibits these topics are not known in advance. These details are the output of the LDA analysis and emerge automatically from the statistical properties of the documents and the assumptions behind LDA.

6.2.3 Creating the Dataset

This paper aims to identify latent fisheries modelling topics from scientific research articles published in peer-reviewed journals specialising in fisheries. In this manner, the selection of publications was restricted exclusively to fisheries journals; therefore, it follows that some subjective choices were made to achieve this. All journals included in this analysis contain the term “fishery” or “fisheries” in their title and have an impact factor of 1.0 or higher. Additionally, the journal *The ICES Journal of Marine Science* was included, because it is part of the International Council for the Exploration of the Seas (ICES), which channels science-based advice to decision makers for sustainable fisheries, and fisheries models are an important focus of this journal. A total of 13 fisheries journals were included in the study (see Table 6.1). A time frame of 26 years, from 1990 to 2016, was chosen to allow for enough variation within publication trends. Due to difficulties with journal subscription rights and the fact that some journals started after 1990 (e.g., *Fish and Fisheries* was first published in 2000), coverage was incomplete for the complete time range of 26 years for a few journals. Documents that did not constitute a type of research article (e.g., book reviews, forewords, errata, conference reports, comments, policy notes, corrigenda and letters) were discarded. In total, 22,236 full-text research articles from 13 top-tier fisheries journals were downloaded using automated download scripts, as well as by utilising the available application programming interfaces (APIs) offered by the publishers. The use of full-text articles, in contrast to only using abstracts, has shown to increase topic quality and provide a more detailed overview of the latent topics permeating a document collection (Syed and Spruit, 2017). Table 6.1 provides an overview of the complete dataset utilised in this study.
Table 6.1: Overview of the dataset (i.e., corpus): years represents the years for which documents (i.e., articles) are downloaded, IF the journal's impact factor according to ISI Journal Citation Reports 2016, N the number of documents, N/T the percentage of journal articles in relation to the total number of articles, $\bar{W}$ the mean number of words within each document, std. W the estimated standard deviation of words within each document, and $\bar{V}$ the mean vocabulary size (number of unique words) within each document. The total number of documents is 22,236.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Years</th>
<th>IF</th>
<th>N</th>
<th>N/T</th>
<th>$\bar{W}$</th>
<th>Std. W</th>
<th>$\bar{V}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Journal of Fisheries and Aquatic Sciences</td>
<td>1996-2016</td>
<td>2.44</td>
<td>4,427</td>
<td>19.90%</td>
<td>4,076</td>
<td>1,306</td>
<td>1,267</td>
</tr>
<tr>
<td>Fish and Fisheries</td>
<td>2000-2016</td>
<td>8.26</td>
<td>419</td>
<td>1.90%</td>
<td>5,893</td>
<td>2,801</td>
<td>1,757</td>
</tr>
<tr>
<td>Fisheries</td>
<td>1997-2016</td>
<td>2.43</td>
<td>477</td>
<td>2.10%</td>
<td>3,410</td>
<td>1,633</td>
<td>1,312</td>
</tr>
<tr>
<td>Fisheries Management and Ecology</td>
<td>1994-2016</td>
<td>1.51</td>
<td>1,001</td>
<td>4.50%</td>
<td>2,692</td>
<td>1,136</td>
<td>956</td>
</tr>
<tr>
<td>Fisheries Oceanography</td>
<td>1997-2016</td>
<td>2.73</td>
<td>752</td>
<td>3.40%</td>
<td>3,867</td>
<td>1,354</td>
<td>1,188</td>
</tr>
<tr>
<td>Fisheries Research</td>
<td>1995-2016</td>
<td>2.23</td>
<td>3,610</td>
<td>16.20%</td>
<td>3,204</td>
<td>1,326</td>
<td>1,064</td>
</tr>
<tr>
<td>Fishery Bulletin</td>
<td>1990-2016</td>
<td>1.51</td>
<td>1,441</td>
<td>6.50%</td>
<td>3,356</td>
<td>2,037</td>
<td>1,074</td>
</tr>
<tr>
<td>ICES Journal of Marine Science</td>
<td>1990-2016</td>
<td>2.63</td>
<td>3,903</td>
<td>17.60%</td>
<td>3,380</td>
<td>1,379</td>
<td>1,119</td>
</tr>
<tr>
<td>Marine and Coastal Fisheries</td>
<td>2009-2016</td>
<td>1.44</td>
<td>274</td>
<td>1.20%</td>
<td>4,474</td>
<td>1,364</td>
<td>1,368</td>
</tr>
<tr>
<td>North American Journal of Fisheries Management</td>
<td>1997-2016</td>
<td>1.01</td>
<td>2,517</td>
<td>11.30%</td>
<td>3,289</td>
<td>1,421</td>
<td>1,037</td>
</tr>
<tr>
<td>Reviews in Fish Biology and Fisheries</td>
<td>1991-2016</td>
<td>3.22</td>
<td>659</td>
<td>3.00%</td>
<td>5,800</td>
<td>3,994</td>
<td>1,750</td>
</tr>
<tr>
<td>Reviews in Fisheries Science &amp; Aquaculture</td>
<td>1997-2016</td>
<td>2.03</td>
<td>375</td>
<td>1.70%</td>
<td>6,186</td>
<td>6,020</td>
<td>1,737</td>
</tr>
<tr>
<td>Transactions of the American Fisheries Society</td>
<td>1997-2016</td>
<td>1.47</td>
<td>2,381</td>
<td>10.70%</td>
<td>3,888</td>
<td>1,382</td>
<td>1,203</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22,236</td>
</tr>
</tbody>
</table>
The selection of fisheries journals and underlying fisheries publications comes with some limitations. First, some of the highly influential and most cited papers on fisheries models are published in high-impact journals such as *Nature*, *Science*, and *PNAS*. Although highly influential, such publications would constitute only a small number of our sample and would only marginally or even negligibly contribute to the overall number of 22,236 publications downloaded from fisheries journals for this study. Two other reasons exist to exclude such generic journals. The first reason is that including all publications published in such outlets would drastically increase the number of uncovered topics, as fisheries make up a small portion of the publications in *Nature*, *Science* and *PNAS*. While one might be able to use keyword searches and include only those publications that match fisheries-related terms, this brings up the second reason to exclude such journals: publication filtering is based on the subjective choice of relevant keywords and is limited in terms of how publications are indexed and subsequently can be retrieved (e.g., title, abstract or full text) from these journals. Through the inclusion of publications from only fisheries journals, such subjective choices and associated limitations are avoided.

The second limitation concerns the exclusion of non-fisheries-specialised journals in which fisheries-modelling-related publications might appear. Such journals focus on, but not limited to, the field of marine science (e.g., *Marine Policy* and *Advances in Marine Biology*), the field of coastal areas or zones (e.g., *Coastal Management* and *Ocean and Coastal Management*), the field of toxicology (e.g., *Environmental Toxicology* and *Pharmacology and Aquatic Toxicology*), and the field of modelling (e.g. *Environmental Modelling & Software* and *Ecological Modelling*), in addition to a number of other journals, such as *Developmental Dynamics*, *Bulletin of the American Meteorological Society*, *Environmental Science and Technology*, *Philosophical Transactions of the Royal Society*, *Environmental Health Perspectives*, *BioScience*, *Journal of Fish Biology*, and *Progress in Oceanography*. Some publications related to fisheries modelling approaches are published in these outlets, which is a potential limitation of this study. Again, filtering for fisheries modelling publications in these journals would be biased by the subjective choice of keywords and limitations due to indexing and retrieval functionalities. Consequently, publications with a focus on the novelty in modelling approaches, which are commonly published in specialised modelling journals such as *Ecological Modelling*, were not assessed in this study. On the other hand, the modelling publications captured within the fisheries journals included in this study can potentially address other topics besides fisheries, such as climate change or habitat loss, which are likely to be included in the analysis of modelling publications.

The third limitation relates to the focus on peer-reviewed journals only. As a result, fisheries modelling research that appears in grey literature was excluded. As grey literature is not indexed in the same way as peer-reviewed studies, selecting only relevant grey literature would, again, introduce bias due to human subjectivity in the search and retrieval.
6.2.4 Pre-processing the Data Set

Several important pre-processing steps were required to transform the documents into appropriate bag-of-word representations. First, each document was converted from PDF format into a plain-text representation. Image-based PDFs, mainly old documents from the 1990s, were converted using the Tesseract optical character recognition (OCR) library. Second, documents were tokenised, which involved creating individual words (e.g., from paragraphs and sentences); meanwhile, numbers, single characters, punctuation marks, and words with only a single occurrence were removed, since they bear no topical meaning. Additionally, words that occurred in ≥90% of the documents were discarded due to their lack of distinctive topical significance. Boilerplate content, such as title pages, article metadata, footnotes, margin notes and so on, was also removed. The reference list of each article was maintained so as to allow for referenced titles and names of authors to be part of the word distributions of topics. An advantage of this approach is that author names can be part of specific topics, but they can simultaneously introduce bias when the referenced articles have no direct link to the underlying topics. A standard English stop word list (n = 153) was used to remove words that serve only syntactical and grammatical purposes, such as 'the', 'and', 'were' and 'is'. Finally, other than grouping lowercase and uppercase words, no normalisation method was applied, such as stemming or lemmatisation, to reduce the inflectional and derivational forms of words to a common base form (e.g., fishing and fishery to fish). Normalisation reduces the interpretability of topics at later stages, as stemming algorithms can be overly aggressive and may result in unrecognisable words when interpreting topics. Stemming might also lead to another problem, as it cannot be deduced whether a stemmed word comes from a verb or a noun (Evangelopoulos et al., 2012). For these reasons, and considering that the interpretability of the topics at a later stage was considered to be highly significant, an extensive normalisation phase was omitted.

6.2.5 Creating LDA Models

The LDA models were created with the Python library Gensim (Rehurek and Sojka, 2010). The number of topics to be uncovered (i.e., K parameter) varied from one to 50, thus creating 50 different LDA models. The hyper-parameters for the LDA models, which affect the sparsity of the topics created and their relative proportions, were set to be symmetrical. Technically, since LDA is a Bayesian probabilistic model, the symmetrical hyper-parameters encode prior knowledge that a priori assign equal probabilities to topics within documents, and words within topics. The quality of each topic was calculated using a topic coherence measure to find the optimal value for K (analogous to finding the right number of clusters for, e.g., k-nearest neighbours). A coherence measure calculates the degree of similarity between a topic's top N words. This provides a quantitative approach for assessing the interpretability of topics from a human perspective. As such, coherence measures aim to find coherent topics—a topic with
CHAPTER 6. SUB-TOPIC ANALYSIS OF FISHERY MODELS

top words 'apple', 'pear', and 'banana' is more coherent than 'apple', 'pear', and 'car'—rather than topics that are merely artefacts of the statistical assumptions behind LDA. The $C_v$ coherence measure was adopted, since it has shown the highest accuracy of all available coherence measures (Röder et al., 2015). An elbow method was employed to find the K value with the best performing topic coherence score.

6.2.6 Identifying Subtopics

For each modelling topic identified, a zoom-in was employed with the aim of uncovering underlying subtopics within each of the general modelling topics by applying an approach similar to that described above. These subtopics provide a more detailed deconstruction of the respective general modelling topics. A zoom-in is performed on a subset of the data consisting of documents that have the general modelling topic as the dominant topic. The dominant topic is defined as the topic with the highest relative proportion—that is, the topic that exceeds all other topic proportions within a document. Since documents are modelled as mixtures of topics, the dominant topic represents the primary topic of a document.

6.2.7 Labelling the Topics

The LDA model outputs the uncovered topics as probability distributions over all the words used; when sorted, the top 15 words are used to label the topic semantically. Representing the words as probabilistic topics has the distinct advantage that each topic is now individually interpretable (Griffiths et al., 2007), compared to a purely spatial representation like the topic model of latent semantic analysis (Deerwester et al., 1990). As stated before, the distributions of words, and specifically the words with the highest probability within each topic, are used to describe an underlying theme; however, such themes are latent, and a semantic label that best captures those words needs to be attached. For example, a topic with the top 5 words apple, banana, cherry, pear, and mango describes the underlying theme of fruits and can be labelled as such.

To provide a semantically meaningful and logical interpretation of these probability distributions, a fisheries domain expert manually labelled the topics by close inspection of the top 15 high-probability words, together with an inspection of the document titles and content. Furthermore, to improve the labelling of the topics, the topics were visualised in a two-dimensional area by computing the distance between topics (Chuang et al., 2012) and applying multidimensional scaling (Sievert and Shirley, 2014). This two-dimensional topic representation aided in identifying similarities between topics and thus similarities between topic labels.
6.2.8 Calculating Sub-Topical Modelling Trends

To gain insight into the sub-topical temporal dynamics of the modelling subtopics, document topic proportions were aggregated into a composite topic-year proportion. Such composite values provide insights into the prevalence of a modelling subtopic within a certain year, given all the publications within that year. It furthermore enables the analysis of changing topic proportions over the course of 26 years, as proportions increase or decrease for each subtopic and for each year. Additionally, to obtain insight into increasing and decreasing topical trends, a one-dimensional least square polynomial was fitted for different time intervals. The time intervals chosen were 1990–1995, 1995–2000, 2000–2005, 2005–2010, and 2010–2016, so as to allow for historical comparison. The polynomial coefficient is used as a proxy for the trend and defines the slope of the composite topic-year proportions for a range of years. Coefficients are multiplied by the number of years within each time interval to obtain the change measured in percentage points. Positive values indicate increasing or “hot” topics, and negative values indicate decreasing or “cold” topics. Color coding is used to represent the hot (i.e., red) and cold (i.e., blue) topical trends.

6.3 Results and Discussion

6.3.1 General Modelling Topics

The optimal LDA model for the complete corpus (N = 22,236 documents) uncovered 31 general fisheries topics. The calculated coherence scores to obtain the optimal number of topics, referred to as the K parameter, can be found in Fig. 6.7 in the Appendix. Among these general fisheries topics, two topics deal with the aspects of fisheries modelling. The publications dealing with these two modelling topics account for 12% (N = 2,761 documents) of the total number of publications. The remaining 29 topics, which relate to other aspects of fisheries research, are listed in Table 6.3 in the Appendix. A bibliometric analysis of trends in fisheries science found a higher proportion of publications employing models—around 30%, as estimated from publication titles and abstracts from a dataset containing 695 fisheries-related publications (Jarić et al., 2012). Several reasons can be offered to explain why these two percentages differ, such as the used time range and the selected journals; most importantly, the present paper identifies publications which predominantly deal with fisheries modelling aspects, in contrast to publications in which a modelling method is employed.

Figure 6.1 shows the top 15 words and their probabilities for the two modelling topics. The first modelling topic concerns catch-effort and abundance estimation methods and is, therefore, given the short name estimation models. It contains the words ‘catch’, ‘survey’, ‘sampling’, ‘effort’, and ‘sample’ among its top 15 words. These words reflect the collection of both fisheries-independent data, which are usually gathered through
FIGURE 6.1: The two uncovered fisheries modelling topics (i.e., estimation models and stock assessment models) from the dataset containing 22,236 fisheries publications (1990–2016; 13 journals). The figure displays the topic label (top) and the top 15 high-probability words.

survey and sampling methods, and fisheries-dependent data (e.g., collected through logbooks), which commonly provide information on catch and effort. These and other obtained data feed into models in order to estimate intermediate parameters such as natural mortality rate or catchability (Hoggarth et al., 2006); this is a phase of research reflected in estimation models through the words ‘model’, ‘estimates’, ‘estimated’, and ‘estimate’. These types of models might also be called retrospective models, since they interpret the past based on collected data.

The second modelling topic concerns modelling approaches for the assessment of the current state of a fishery and future projections and is assigned the short name stock assessment models. It contains the words ‘stock’, ‘mortality’, ‘biomass’, ‘rate’, and ‘estimate’, which reflect the most commonly used indicators (i.e., fish catch, stock biomass, stock size and fishing mortality; (Hoggarth et al., 2006)) to measure the status of the fishery and the state of the stock (Le Gallic 2002). These indicators link to reference points, which give quantitative meaning to the goals and objectives set for a fishery (Jennings et al., 2009). Reference points are usually estimated through models that use stock and recruitment data, which is reflected in the words ‘stock’, ‘population’, ‘recruitment’, ‘management’, ‘parameters’, and ‘estimates’ in stock assessment models. Together, indicators and reference points play a crucial role in fisheries management and can be used to give quantitative meanings to the objectives of a fishery (Hoggarth et al., 2006).

The distinction between these two topics shows how they are treated separately in fisheries research publications, whereas in practice (i.e., in fisheries stock assessments
for management), these two topics are connected and combined into one model but reflect the different phases of the model development (Hoggarth et al., 2006). The distribution of publication frequencies for both general modelling topics is shown in Figure 6.2, which highlights the increased research interest in stock assessment models compared to estimation models. Additionally, the top five publications with the highest topic prevalence for each of the two modelling topics, indicating to what extent the content of a publication relates to the modelling topic, are shown in Table 6.2.
<table>
<thead>
<tr>
<th>Modelling Topic</th>
<th>Title</th>
<th>Year</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimation Models</strong></td>
<td>Trawl survey based abundance estimation using data sets with unusually large catches.</td>
<td>1999</td>
<td>95.69%</td>
</tr>
<tr>
<td></td>
<td>Covariances in Multiplicative Estimates.</td>
<td>1999</td>
<td>94.35%</td>
</tr>
<tr>
<td></td>
<td>Use of simulation-extrapolation estimation in catch-effort analyses.</td>
<td>1999</td>
<td>93.90%</td>
</tr>
<tr>
<td></td>
<td>Reducing Bias and Filling in Spatial Gaps in Fishery Dependent Catch per Unit Effort Data by Geostatistical Prediction I Methodology and Simulation.</td>
<td>2014</td>
<td>92.23%</td>
</tr>
<tr>
<td></td>
<td>Confidence intervals for trawlable abundance from stratified-random bottom trawl surveys.</td>
<td>2011</td>
<td>90.48%</td>
</tr>
<tr>
<td><strong>Stock Assessment Models</strong></td>
<td>The Structure of Complex Biological Reference Points and the Theory of Replacement.</td>
<td>2009</td>
<td>99.37%</td>
</tr>
<tr>
<td></td>
<td>Analytical models for fishery reference points.</td>
<td>1998</td>
<td>98.50%</td>
</tr>
<tr>
<td></td>
<td>Implications of life-history invariants for biological reference points used in fishery management.</td>
<td>2003</td>
<td>98.14%</td>
</tr>
<tr>
<td></td>
<td>The estimation and robustness of FMSY and alternative fishing mortality reference points associated with high long-term yield.</td>
<td>2012</td>
<td>97.33%</td>
</tr>
<tr>
<td></td>
<td>Age-specific natural mortality rates in stock assessments: size-based vs. density-dependent.</td>
<td>2014</td>
<td>94.87%</td>
</tr>
</tbody>
</table>
Interestingly, only the topics of estimation models and stock assessment models were uncovered (both of which focus on the ecological dimension of fisheries), whereas topics on economic and social fisheries aspects were not found within the modelling publications. This finding might be a result of the selection of journals used in this study. Most of the included fisheries journals declare a multidisciplinary or interdisciplinary scope, while some specifically include socio-economic considerations and the human dimension as subjects of interest. Therefore, at least one social or economic modelling topic could be expected to be identified by the LDA model. Another reason for the absence of other modelling topics may be that fisheries are still perceived as a natural science. The International Council for the Exploration of the Sea (ICES) only recently established the Strategic Initiative on the Human Dimension (SIHD) “to support the integration of social and economic science into ICES work” (ICES, 2017), and the majority of the ICES workgroups still lack social science input (ICES, 2016). As a result, social scientists and economists may pursue publication of their models not in a journal related to fisheries, but rather in a journal related to their respective disciplines or having a broader scope, such as *Ecology and Society*, *Marine Resource Economics* or *Marine Policy*. Merit issues could also contribute to the topic bias. Different scientific disciplines receive publication merits for different journals, which is more often dependent on the index of a journal (e.g., Science Citation Index (SCI), Social Science Citation Index (SSCI), or International Scientific Index (ISI)) than on its impact factor. As a result, non-biological and non-ecological disciplines are less likely to use top-tier fisheries journals as publication outlets. This might, in turn, lead to low visibility of non-ecological models among fisheries stakeholders, because many fisheries journals such as *Fish and Fisheries* and *Fisheries Research* intend to reach fisheries managers, administrators, policy makers and legislators.

### 6.3.2 Subtopics within Estimation Models

The zoom-in (i.e., the process of uncovering subtopics from general topics) on the general topic of estimation models (N = 1,124 documents) identified 14 sub-topics (see Figure 6.7 in the Appendix). Figure 6.3 provides an overview of the 14 estimation model sub-topics, the top 15 words of the topics with their probabilities, and the manually attached label that best captures the semantics of the top words. Furthermore, a two-dimensional topic representation can be found in the topic similarity map in Figure 6.4 (A), showing the topic similarity with respect to the distribution of the words. The trends (i.e., the change in overall topic proportion, in percentage points) and prevalence (i.e., the size of the overall topic proportion as a percentage) are presented in Figure 6.5 (A).

Most of the uncovered subtopics can be grouped. The principal group consists of the five subtopics focusing on the biological aspects of fisheries (i.e., catch and abundance, mortality rate, fish distribution, spawning, and length and growth). This highlights the importance and scientific focus of the biological dimension in fisheries re-
### Figure 6.3: The 14 uncovered subtopics from the documents (N = 1,124) exhibiting the topic estimation models as the dominant topic. The figure displays the subtopic label (top) and the top 15 high-probability words.
Figures 6.4: Topic similarity map that shows a 2-dimensional representation (via multi-dimensional scaling). A: 14 estimation model subtopics. B: 15 stock assessment model subtopics. The distance between the nodes represents the topic similarity with respect to the distributions of the words (i.e., nodes closer together have more related word probabilities). The surface of the nodes represents the prevalence of the topic within the corpus.

Catch and abundance shows the biggest overall increase over time (+15.46%) and had the largest proportion (14.84%) within the last six years (Figure 6.5 (A)). Most of the other biological subtopics show very little variation over time, and some only make a small contribution in terms of proportion (e.g., spawning), with only 3.82% overall topic proportion (Figure 6.5 (A)). Length and growth showed the highest overall decrease over time (-14.04%), indicating a diminishing scientific interest. The subtopic of length and growth remained relatively high in terms of topic proportion, with an average of 9.13% between 2010 and 2016, possibly because growth is an important parameter for stock assessments (Lorenzen, 2016; Maunder et al., 2016) and is also most frequently discussed in fisheries, as shown by a previous trend analysis (Jarić et al., 2012). The subtopic of parameters and estimators relates more to the technical aspects of estimation modelling, but appears to be similar to the biological subtopic of mortality rate, as apparent from the similarity map (Figure 6.4 (A)). Vessel and fleet showed a large topic proportion (between 8% and 10%) over the last 16 years (Figure 6.5 (A)). Both the topic of vessel and fleet and that of net selectivity likely relate to biological considerations, but they could also hint at a slightly more economic perspective on industry (fleet) and gear-related matters; however, additional words such as ‘firm’, ‘prices’, or ‘market’ would have to be present to confirm this hypothesis further. The four subtopics of abundance (survey), sampling, abundance (sampling), and trawl survey focus on survey and sampling, which are essential methods for gathering data and information on fisheries. In particular, information on catch and stock abundance is required by almost all stock assessment models (Hoggarth et al., 2006). These four subtopics account for a combined overall topic prevalence of 30.73%, indicating their importance to fisheries research. The subtopic of recreational fisheries
Figure 6.5: Trends in changing topic proportions for different time intervals for all subtopics. The left-hand side (A) displays the 14 uncovered estimation model subtopics. The surface of the node represents the topic prevalence within a certain time range and indicates how present a topic was within all the published material of that time frame. The colours indicate the trend in topic proportion (i.e., change in percentage points) and indicate whether a topic increased in popularity (hot topic) or decreased in popularity (cold topic) within that time frame. The right-hand side (B) displays the information for the 15 uncovered stock assessment model subtopics.
CHAPTER 6. SUB-TOPIC ANALYSIS OF FISHERY MODELS

refers to a type of fishery that differs in the estimation process compared to commercial fisheries, as it often employs surveys on anglers. This type of estimation process may refer not only to marine but also to freshwater fisheries. Recreational fisheries underwent an increase in topic proportion from 2.11% in the 1990–1995 period to 7.90% in the 2010–2016 period, indicating the growing importance of recreational fisheries assessments in fisheries science. The increased importance of recreational fishing on the commercial fish stocks (Griffiths and Fay 2015) is in line with the observed trend in this study. Apart from recreational fisheries, no other types of fisheries (e.g., small-scale, artisanal, or commercial fisheries) were identified by the topic model. The distance of recreational fisheries from the other subtopics in the similarity map may explain this, as authors writing about recreational fisheries use distinctive words that are different from the discourse on other types of fisheries. Another possible explanation may be that there are more studies on recreational fisheries than on other types of fisheries. Salmon is the only topic that focuses on one particular species. The similarity map shows how the topic of salmon differs within the words used, indicating the particularity and specialised research niche of the topic (Figure 6.4 (A)). Salmon showed a positive trend (+5.61%) over the study period; however, this result is in conflict with previous research that showed a diminishing research interest in the species (Jarić et al. 2012). This could be due to the increasing effort within aquaculture and the growing economic importance of the species over the period (FAO, 2016) that separates this study from that of Jarić et al. (2012).

Within the top 15 words of the subtopics, important subjects such as species and names/methods can be identified. Three subtopics contain species names (i.e. ‘shrimp’ in sampling, ‘cod’ and ‘crab’ in fish distribution, and ‘salmon’ and ‘chinook’ in salmon). Methods mentioned within the subtopics of estimation models are ‘regression’ in parameters and estimators and ‘Bayesian’ in abundance (sampling). Parameters for fish stock assessments can be estimated through the least square method, represented in the form of regression analysis; however, maximum likelihood methods are now preferred, as they allow for a better specification in the form of errors in the models. Bayesian methods are commonly used to incorporate uncertainty into management advice, but this could also involve other methods such as maximum likelihood, bootstrapping, or Monte-Carlo modelling (Hoggarth et al., 2006). The two methods ‘regression’ and ‘Bayesian’ do not reflect the current diversity of modelling methods, nor necessarily the most conventional models used in fisheries assessments today, but they seem to have a strong association with the two topics of parameters and estimators and abundance (sampling). Note that references to names of species and methods highlight the importance and relation of such words within a specific topic—technically, they co-occur more frequently to describe the latent topic—but are by no means mutually exclusive (i.e., methods and species can occur in different subtopics simultaneously). They provide information from a topical perspective (i.e., a high-level decomposition of the document into clusters of co-occurring words), but fail to address on what basis such species and methods are linked within a specific topic.
6.3.3 Subtopics within Stock Assessment Models

The zoom-in on the topic of stock assessment models (N = 1,637 documents) revealed 15 sub-topics (see Figure 6.7 in the Appendix for the calculated topic coherence scores). Figure 6.6 provides an overview of the 15 sub-topics, the top 15 words with their probabilities, and the label attached to each topic. The topic similarity for these subtopics can be found in Figure 6.4 (B). The subtopic trends and prevalence are displayed in Figure 6.5 (B).

Most of the subtopics of stock assessment models evolve around biological aspects and processes (i.e., growth and length, movement, predation, cod recruitment, fecundity and reproduction, population dynamics, life history, and stock recruitment). The majority of these subtopics show a slight increase over the study period (Figure 6.5 (B)); together, these subtopics have an overall topic proportion of 42.91%, which shows their consistent importance within fisheries science and fisheries management (Hilborn and Walters 1992). Within the biological subtopics, predation stands out as the only subtopic that refers to ‘interaction’, ‘multispecies’ and the ‘ecosystem’. The subtopic of predation increased by 4.67% during the period from 1990 to 1995 (Figure 6.5 (B)), which reflects the increased scientific awareness of predator-prey interaction and model implications in the early 1990s (e.g., (Yodzis, 1994)). The topic proportion of predation shows a positive trend, as it rose from 3.75% in the period of 1990–1995 to 5.07% in the period of 2010–2016; this might indicate the increased attention of the scientific community to an ecosystem approach to fisheries and the implementation of multispecies and ecosystem considerations within stock assessments, modelling frameworks, and management advice (Maynou, 2014; Möllmann et al., 2014; Gaichas et al., 2017). The four subtopics of harvest strategy, management effects, management tools and reference points all concern management measures and effects, but they mainly address biological components such as ‘recruitment’, ‘abundance’, and ‘biomass’. Reference points shows the strongest overall negative trend of all subtopics (-26.55%), indicating that the popularity of this topic among fisheries scientists has decreased over the years. Nevertheless, the topic of reference points still makes up a relatively large proportion, 9.82% (Figure 6.5 (B)); this is the second largest proportion in the period of 2010–2016 after estimator performance, which has a 15.19% topic proportion within the same period. This highlights the continuity of research on reference points from the 1990s to the present day (Caddy and Mahon, 1995; Caddy, 2004; Froese et al., 2017).

The subtopic of estimator performance shows the highest increase (+11.11%) within the overall study period (i.e., 1990–2016) and makes up a large proportion within the last six years of the time frame, from 2010–2016 (15.19%); this finding could be related to the increased overall importance of models in fisheries science (Jarić et al., 2012). The subtopic of freshwater fisheries shows an overall positive trend (+6.28%), even though freshwater fisheries habitats have been found to be less studied than marine fisheries (Jarić et al., 2012). The topic proportion of freshwater fisheries rose over the study period, from 1.82% in 1990–2000 to 8.08% in 2010–2016 (Figure 6.5 (B)). The importance of freshwater fisheries in areas such as Africa and India may explain the increase in research efforts within this field (FAO, 2016).
Figure 6.6: The 15 uncovered subtopics from the documents (N = 1,637) exhibiting the topic stock assessment models as the dominant topic. The figure displays the subtopic label (top) and the top 15 high-probability words.
CHAPTER 6. SUB-TOPIC ANALYSIS OF FISHERY MODELS

From the top 15 words (Figure 6.6), related subjects were identified, such as regions, species, and names/methods. The two marine regions mentioned are ‘Atlantic’ and ‘Pacific’, possibly because these are some of the world’s major fishing areas (FAO, 2016). The various species names found within the top 15 words, such as ‘cod’, ‘herring’, and ‘anchovy’, cover many of the commercially important species in marine capture production (FAO, 2016). These results stand in stark contrast to a bibliometric study on trends in fisheries science, which found virtually no research on many commercially important species (Aksnes and Browman, 2016); however, these results were based on word frequencies in publication titles and abstracts, which may not mention the species of concern. This finding highlights the strength of the full-text LDA analysis. Other mentioned species, such as ‘abalone’, ‘lobster’, and ‘shark’, may have high probabilities for occurrence in the subtopics because they represent species of great economic value and also are often a focus of conservation efforts (Turpie et al., 2003; Simpfendorfer and Dulvy, 2017).

Several names within the words of the subtopics refer to a method named after a scientist, such as ‘Bayesian’, ‘Bertalanffy’, ‘Ricker’, and ‘Punt’, which could be a direct consequence of the inclusion of the reference list in the analysis. The subtopic of Bayesian approach indicates the importance of this methodology in fishery science and for fisheries models. A Bayesian approach can be used for stock assessments and decision analysis and resembles an improved way of fitting models to data and decision making (Hoggarth et al., 2006). The scientists von Bertalanffy and Ricker both made substantial contributions to fisheries science—von Bertalanffy in metabolism and growth (von Bertalanffy, 1957) and Ricker in the computation and interpretation of computational statistics of fish populations (Ricker, 1975). Their methods are still applied today in the form of growth models (Allen, 1966; Piner et al., 2016) and in stock-recruitment models (Baker et al., 2014). The author Punt has not developed any particular method that takes his name; however, his name may occur within the top 15 words due to his significant contribution to research and his publications on estimator performance and data standardisation, as well as his many citations by other scientists within the field. Although Punt is, relatively speaking, a newcomer compared to some of the early influential researchers in the field (e.g., Hjort, Beverton, and Holt), the occurrence of his name is perhaps a result of the timeframe examined, or it may indicate that the names of senior scientists and methods have become somewhat common knowledge and are therefore not always explicitly stated or cited.

6.4 Conclusions

The aim of this paper was to uncover fisheries modelling topics from 22,236 scientific publications from 13 peer-reviewed fisheries journals. Additionally, subtopics from general modelling topics were uncovered to provide insights into their developments and trends over the last 26 years. Overall, two main fisheries modelling topics were identified: estimation models and stock assessment models. This study demonstrates
that research in the field of fisheries modelling shows a shift of scientific focus in topics and subtopics over the last 26 years. Stock assessment models are outperforming estimation models, and their underlying subtopics have moved from length and growth to catch and abundance, and from reference points to estimator performance over the last 26 years. Economically important species and areas show a high presence within the modelling subtopics.

Both general modelling topics focus primarily on the biological aspects of fisheries; however, since this study was limited to publications in 13 fisheries journals, other topics in fisheries modelling (e.g., with a focus on social, management or economic aspects of fisheries) may well exist in publications of other journals. Possible disciplinary merit issues and the remaining understanding of fisheries as a natural science discipline might further limit fisheries journals to models with an ecological focus, despite their multidisciplinary scope.

In conclusion, this novel machine learning approach revealed interesting insights into the topical trends of a large dataset of models published in fisheries journals. This approach enables researchers to identify research topics and shifts in research focus, and it provides a bigger picture that captures the main ideas prevailing scientific publications.
Appendix
Figure 6.7: Calculated coherence scores (y-axis) for the number of topics (x-axis) (i.e., K parameter) for three different runs. The average coherence score is calculated by averaging the scores over all three runs for the same K parameter. The figures represent the following: all documents (N = 22,236); documents that exhibit the topic estimation models as the dominant topic (N = 1,124); documents that exhibit the topic stock assessment models as the dominant topic (N = 1,637).
Table 6.3: The top 15 words (i.e., the words with highest probability) for each of the 31 uncovered general fisheries topics. Topics in bold (i.e., 4 and 9) are the identified modelling topics used in the analysis of this paper, with 4 being the topic estimation models and 9 being the stock assessment models.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top 15 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>crab, crabs, lobster, eel, eels, size, traps, mm, lobsters, trap, american, anguilla, blue, females, fishery</td>
</tr>
<tr>
<td>2</td>
<td>salmon, hatchery, chinook, river, wild, atlantic, survival, coho, sockeye, juvenile, oncorhynchus, fisheries, smolts, pacific, steelhead</td>
</tr>
<tr>
<td>3</td>
<td>river, species, sampling, electrofishing, colorado, fishes, chub, capture, population, suckers, sucker, abundance, reach, sites, site</td>
</tr>
<tr>
<td>4</td>
<td>model, estimates, catch, survey, sampling, estimated, models, estimate, distribution, abundance, mean, effort, sample, method, size</td>
</tr>
<tr>
<td>5</td>
<td>genetic, populations, population, river, loci, samples, among, structure, individuals, microsatellite, within, stock, alleles, diversity, sample</td>
</tr>
<tr>
<td>6</td>
<td>prey, larvae, growth, larval, food, predation, size, feeding, diet, juvenile, zooplankton, abundance, mm, predator, rates</td>
</tr>
<tr>
<td>7</td>
<td>red, reef, gulf, species, snapper, florida, marine, mexico, reefs, fishes, shrimp, coral, habitat, artificial, drum</td>
</tr>
<tr>
<td>8</td>
<td>atlantic, bay, striped, tuna, bass, flounder, estuary, north, marine, new, river, carolina, chesapeake, estuaries, estuarine</td>
</tr>
<tr>
<td>9</td>
<td>model, stock, mortality, population, recruitment, models, biomass, year, rate, management, parameters, assessment, fisheries, estimates, fishing</td>
</tr>
<tr>
<td>10</td>
<td>species, variables, environmental, sites, lakes, assemblages, community, water, assemblage, richness, communities, diversity, index, models, spatial</td>
</tr>
<tr>
<td>11</td>
<td>cod, sea, atlantic, north, species, herring, size, cm, trawl, length, stock, area, mesh, baltic, fishing</td>
</tr>
<tr>
<td>12</td>
<td>fisheries, management, fishing, fishery, catch, economic, marine, effort, fishers, species, recreational, information, anglers, use, new</td>
</tr>
<tr>
<td>13</td>
<td>habitat, water, flow, use, depth, river, velocity, substrate, channel, areas, sites, site, area, movement, spawning</td>
</tr>
<tr>
<td>14</td>
<td>spawning, females, eggs, egg, males, female, reproductive, male, fecundity, sex, maturity, mature, stage, size, development</td>
</tr>
<tr>
<td>15</td>
<td>species, sharks, bycatch, shark, catch, longline, fishery, fisheries, fishing, gear, hooks, caught, hook, cm, atlantic</td>
</tr>
</tbody>
</table>
### Table 6.3: continued.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top 15 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>mortality, tag, tagged, tags, tagging, release, survival, released, movement, rates, fisheries, studies, capture, effects, transmitters</td>
</tr>
<tr>
<td>17</td>
<td>lake, lakes, perch, michigan, yellow, walleye, great, fisheries, northern, walleyes, mean, ontario, journal, consumption, population</td>
</tr>
<tr>
<td>18</td>
<td>growth, length, mm, size, otoliths, body, ages, otolith, cm, mean, years, first, weight, differences, lengths</td>
</tr>
<tr>
<td>19</td>
<td>temperature, water, growth, effects, swimming, treatment, energy, levels, temperatures, experiment, activity, body, effect, experiments, experimental</td>
</tr>
<tr>
<td>20</td>
<td>sea, squid, mediterranean, distribution, area, anchovy, species, waters, larvae, sardine, marine, spawning, shelf, temperature, mackerel</td>
</tr>
<tr>
<td>21</td>
<td>bass, largemouth, reservoir, river, species, lake, catfish, smallmouth, fisheries, shad, water, management, reservoirs, white, black</td>
</tr>
<tr>
<td>22</td>
<td>species, fishes, freshwater, carp, new, native, river, water, aquaculture, crayfish, populations, introduced, conservation, tilapia, many</td>
</tr>
<tr>
<td>23</td>
<td>species, dna, genetic, gene, mtDNA, samples, molecular, mitochondrial, sequence, haplotypes, infection, atlantic, identification, disease, sequences</td>
</tr>
<tr>
<td>24</td>
<td>acoustic, depth, vertical, water, bottom, surface, ts, distribution, speed, range, target, density, measurements, night, behaviour</td>
</tr>
<tr>
<td>25</td>
<td>otolith, otoliths, sr, marine, river, ratios, samples, water, differences, juvenile, chemistry, isotope, freshwater, values, campana</td>
</tr>
<tr>
<td>26</td>
<td>fishing, marine, species, fisheries, areas, sea, area, fishery, catch, australia, effort, total, south, effects, coastal</td>
</tr>
<tr>
<td>27</td>
<td>river, sturgeon, dam, chinook, columbia, lower, passage, migration, salmon, downstream, steelhead, upstream, juvenile, spawning, dams</td>
</tr>
<tr>
<td>28</td>
<td>trout, brook, rainbow, brown, lake, sea, fry, lamprey, river, stocking, lampreys, salvelinus, arctic, streams, stocked</td>
</tr>
<tr>
<td>29</td>
<td>water, concentrations, phytoplankton, production, concentration, samples, nutrient, sediment, carbon, total, food, values, biomass, organic, levels</td>
</tr>
<tr>
<td>30</td>
<td>stream, trout, streams, creek, habitat, cutthroat, sites, reaches, river, effects, brook, temperature, watershed, abundance, aquatic</td>
</tr>
<tr>
<td>31</td>
<td>sea, pacific, marine, species, climate, ocean, north, alaska, rockfish, change, ecosystem, changes, abundance, temperature, california</td>
</tr>
</tbody>
</table>
Mapping the Global Network of Fisheries Science Collaboration

As socio-environmental problems have proliferated over the past decades, one narrative which has captured the attention of policymakers and scientists has been the need for collaborative research that spans traditional boundaries. Collaboration, it is argued, is imperative for solving these problems. Understanding how collaboration is occurring in practice is important however, and may help explain the idea space across a field. In an effort to make sense of the shape of collaboration in fisheries science, here we construct a co-authorship network of the field, from a dataset comprising 73,240 scientific articles, drawn from 50 journals, and published between 2000 and 2017. Using a combination of social network analysis and unsupervised machine learning, the work first maps the global structure of scientific collaboration among fisheries scientists at the author, country and institutional levels. Second, it uncovers the hidden subgroups—here country clusters and communities of authors—within the network, detailing also the topical focus of the largest fisheries science communities. We find that while the fisheries science network is becoming more geographically and institutionally extensive, it is simultaneously becoming more intensive. The uncovered network exhibits characteristics suggestive of a thin style of collaboration, and groupings that are more regional than they are global. Although likely shaped by an array of overlapping micro- and macro-level factors, the analysis reveals a number of political-economic patterns that merit reflection by both fisheries scientists and policymakers.

Submitted for publication:

7.1 Introduction

Over the past number of decades, as socio-environmental crises have multiplied, the question of research collaboration has captured the attention of policymakers and scientists alike (Katz and Martin, 1997), with calls for “intensive cooperation” and a widening of perspectives becoming commonplace (Palsson et al., 2013). The underlying assumption driving these calls is that solutions to the complex—social, ecological, socio-ecological—problems facing humanity today are in many instances not going to be found within the confines of traditional disciplinary, thematic, sectoral, or territorial boundaries (European Commission, 2008). In this respect, fisheries have not been an exception. Amidst ongoing dissatisfaction with the outcomes of traditional fisheries science and management—for example, in terms of the increasingly precarious status of fish stocks and the communities that depend upon them (Symes et al., 2015)—policymakers and scientists have shifted their gaze to the production of knowledge within this space, and actively sought to broaden collaborative efforts in this area (European Commission, 2008, 2016b; Geoghegan-Quinn et al., 2013; IOC-UNESCO, 2017; Rozwadowski, 2002; Smith and Link, 2005; Symes and Hoefnagel, 2010b).

It is unsurprising then, if not entirely consequential, that research collaboration has increased exponentially over the past decades (Wuchty et al., 2007). Further, given the significant amount of empirical research suggesting that social relationships and the networks these relationships constitute are important in explaining processes of knowledge production (Bourdieu, 1975, 1991; Forsyth, 2003; Granovetter, 1983; Law, 1987; Moody, 2004; Phelps et al., 2012; Schott, 1991, 1993), it is of no surprise that this shifting character of science (Adams, 2013) has drawn the attention of scholars. Patterns of co-authorship amongst scientists—long recognized as providing a window into collaboration within the academic community (Newman, 2004)—have proven a particularly fruitful line of inquiry in this respect (Adams, 2012, 2013; Azoulay et al., 2010; Ding, 2011; Katz, 1994; Katz and Martin, 1997; Leydesdorff and Wagner, 2008; Liu and Xia, 2015; Martin et al., 2013; Newman, 2001; Wagner et al., 2015a). Regarding the field of fisheries, however, whilst scholars have directed their attention towards characterizing the direction and content of fisheries science publications (Aksnes and Browman, 2016; Jarić et al., 2012; Natale et al., 2012; Nikolic et al., 2011; Syed et al., 2018a), and studies have highlighted that collaboration is increasing within this space (Jarić et al., 2012), we know comparatively little about the structure these collaborations are taking. The small body of work that has analyzed co-authorship networks in fisheries science has been narrowly confined in terms of timespan and journal inclusion (Elango and Rajendran, 2012), or a particular type of fishing (Oliveira Júnior et al., 2016).

Given the applied nature of fisheries science, with science playing a critical role in informing fisheries management decisions (Campling, 2012), and hence having practical consequences for fish and people, understanding how knowledge is produced in this area is especially pertinent. Thus, with an eye to making sense of the shape of fisheries science, here we take scientific collaboration, measured as co-authorship amongst
scientists in this field, as our analytical vantage point. Using a combination of social network analysis and topic modeling (a variant of unsupervised machine learning), alongside theoretical insights from the sociology of science, we map the co-authorship network that characterizes this applied domain, and investigate the collaborative entanglements within this space. In doing so we pose questions with respect to how patterns of collaboration differ between subjects and how these have changed over time (Newman, 2004). Our analysis provides a dynamic portrait (Newman, 2004) of the fisheries science community, and an avenue through which the social dynamics underpinning fisheries science collaborations, and consequently the production of knowledge within this space may be explored (Bourdieu, 1975; Ding, 2011; Forsyth, 2003; Latour, B, 1993; Liu and Xia, 2015; Martin et al., 2013).

Our study builds upon the important groundwork that has been laid out by previous scholars within this domain (Aksnes and Browman, 2016; Elango and Rajendran, 2012; Jarić et al., 2012; Natale et al., 2012; Nikolic et al., 2011; Oliveira Júnior et al., 2016; Syed et al., 2018a) in a number of ways. First, by focusing our attention on the networks of production, we expand upon existing analysis that has focused on the content of fisheries science (Aksnes and Browman, 2016; Jarić et al., 2012; Natale et al., 2012; Nikolic et al., 2011; Syed et al., 2018a), by characterizing the structure of the community of scientists that produce that output, in a manner that may help us understand its content (Bourdieu, 1975; Forsyth, 2003). Second, as detailed, the work that has previously taken a network approach to the production of fisheries-related knowledge (Elango and Rajendran, 2012; Oliveira Júnior et al., 2016), though illuminating, has been hitherto narrowly bounded either by time or specific knowledge communities. Here our analysis is based upon a dataset comprising 73,240 scientific journal articles, drawn from all 50 journals in the fisheries category as defined by the Science Citation Index Extended (SCIE), and published between 2000 and 2017. This category has been cited as containing the core journals within the field (Aksnes and Browman, 2016). Consequently, the network we construct is expansive, comprising 106,137 authors from 100,175 different affiliations, across a broad spectrum of fisheries science research, related to both capture and culture fisheries. This large network is subsequently analyzed at progressively finer levels of granularity (Ding, 2011), across three planes (spatial, temporal, and topical), in a manner which broadens the bounds of the analysis and provides for a multi-dimensional overview of the field.

### 7.2 Results

The results of our investigation are presented in two parts. First, the macro-level structure of the global fisheries science network is detailed and mapped at the author, country, and institutional levels. Second, moving to a more fine-grained level of analysis, the hidden collaborative groupings within the network—here country clusters and communities of authors, within which the nodes (i.e., countries, authors) are more tightly connected to each other than to the rest of the network (Palla et al., 2005)—are spec-
7.2.1 Topology of the Co-Authorship Network

In line with broader trends (Adams, 2012, 2013; Leydesdorff et al., 2013), and previous work regarding fisheries (Aksnes and Browman, 2016; Jarić et al., 2012), the fisheries science collaboration network is expanding rapidly (Appendix Fig. 7.7). The number of authors participating in the network has increased steadily, whilst the number of collaborative ties via publication has increased almost exponentially, with a rapid increase visible since 2015. This has been fueled, at least in part, by the volumetric rise in fisheries science publications, which has almost doubled since 2000 (Appendix Fig. 7.8). That said, as the network has expanded, the network degree (i.e., the average number of connections possessed by each scientist (Liu and Xia, 2015)) has increased, whilst the average clustering (i.e., the extent to which a scientist’s co-authors also collaborate with each other (Liu and Xia, 2015; Newman, 2004)) has decreased, indicating that collaboration is indeed becoming more extensive. On the other hand, the density (i.e., the degree of connectedness of the network) has decreased, implying that the network has become less structurally cohesive.

In light of the fragmentation and lack of connectivity which has previously been cited as problematic in fisheries science (Jarić et al., 2012; Symes and Hoefnagel, 2010b), and the existing “narrow lenses” that have been detailed as persisting in the field (Syed et al., 2018a), we might tentatively infer that the structural trends exhibited by the fisheries science network are a good thing. Very dense ties can have a homogenizing effect on a network (Bodin and Crona, 2009), whilst high levels of clustering are indicative of fragmentation and division (Lambiotte and Panzarasa, 2009). For example, the more an individual’s collaborators are also connected to one another, the less likely those connections will lead to new collaborations with “dissimilar others”, thereby making exposure to new ideas similarly unlikely (Granovetter, 1973; Lambiotte and Panzarasa, 2009). Thus, that the global fisheries science network exhibits trends in the opposite direction may well suggest that the network is becoming less fragmented (Borrett et al., 2014).

As it has expanded, however, the number of potential connections across the network that have been realized has decreased, and the network has become less structurally cohesive. This pattern could work to limit the spread of ideas across the network (Moody, 2004), with ties in this sense working to enhance knowledge production (Bodin and Crona, 2009). Seen from this angle, this trend may be indicative of a field that is becoming increasingly divided into silos, albeit silos within which there is considerable collaboration. This could have implications in terms of inhibiting knowledge exchange (Borrett et al., 2014), reinforcing lines of division that already exist, or generating new ones. That said, an element of agonistic pluralism is desirable in all fields, certainly in terms of creating space for historically underrepresented ideas (Matulis and
Therefore, cast in a more favorable light, this pattern might suggest that the field is becoming more heterogeneous, in a manner that could provide welcome space for addressing particular problems and the nurturing of new ideas (Borrett et al., 2014), or place-based epistemologies (Escobar, 2004).

7.2.2 Country-Level Giants

Large networks are difficult to visualize, since the nodes (here authors) simply get plotted on top of each other (Moody and Light, 2006). To get a clearer picture of collaboration at the global level, for visualization purposes we therefore aggregated the network of authors at the country level, whereby attribution for each publication was fractionally credited. As has been detailed elsewhere (Aksnes and Browman, 2016; Jarić et al., 2012; Oliveira Júnior et al., 2016), in terms of publication output, the fisheries science network is dominated by authors located in a few geographical regions. A large proportion of the publication volume in this field is produced by a small group of fisheries science powerhouses, comprising a number of traditional fisheries science producers (e.g., US, Canada, Japan, Australia, UK, Norway), who have over the past decades been joined, and in some instances surpassed, by a number of large emerging economies (e.g., China, India, Brazil) (Fig. 7.1 and Appendix Table 7.1).

Although cross-border collaboration in the field has increased over time, the patterns across the field are far from even, and the collaborative landscape—when viewed from the global level—is dominated by Western countries (Fig. 7.2). In line with existing analysis (Jarić et al., 2012), the US, UK, and Canada are the most internationally collaborative countries in the network. As the field has become increasingly collaborative, historical links between European and North American countries have intensified, whilst a number of emerging economies have forged strong links with the US. For example, mirroring the pattern in science more generally (Wagner et al., 2015a), China has emerged as a prominent US collaborator, a relationship that is surpassed only by the collaborative relationship between the US and Canada. Conversely, the traditionally strong—albeit at times unequal—relationship between the US and Japan in this field (Finley, 2011; Hamblin, 2000) has dwindled. The pattern between these two countries has also been detected beyond fisheries science (Wagner et al., 2015b) and may be reflective of a number of factors. These may include shifting geopolitical realities (Hamblin, 2000), Japan’s declining publication output (Aksnes and Browman, 2016), and the isolating effect the Anglophone bias in scientific publication has had on a number of countries, including Japan (King, 2004).

7.2.3 Institutional Dynamics

With an eye to further investigating the level of diversity across the field, we aggregated the network of authors at the institutional level (Figs. S4–S7). Spatially, the
Figure 7.1: Publication percentage per country for the periods 2000–2008 and 2009–2017. Each publication is fractionally credited based on the number of authors and country affiliations. The actual values for the top-25 largest countries can be found in Appendix Table 7.1.
largest institutional cross-border collaborators in the fisheries science network are predominantly located in the Northern Hemisphere, and this pattern remains relatively unchanged. Our analysis does, however, illuminate an increasing typological diversity of institutions across the network. Over time a number of institutes with a more explicit leaning towards the social sciences, albeit erring on the side of economics (e.g., The Institute of Economic Studies, University of Iceland and Socio-Economic Marine Research Institute, NUIG Ireland), have become prominent institutional collaborators at the international level. Alongside this, reflecting the increasingly multi-actor character of fisheries management, these institutions increasingly comprise a mixture of national institutes, universities, private institutes, and non-governmental organizations (NGOs). This trend—whereby, for example, environmental NGOs have become significant producers of scientific knowledge—has also been noted in other fields related to the socio-environment (Holmes, 2011; da Fonseca, 2003).

7.2.4 Hidden Collaborative Groups

Hitherto we have been speaking about the macro-characteristics of the overall network, albeit at different levels. While this provides a good overview of the global fisheries science network, many of the most important characteristics of a network only become apparent when analyzing the hidden groups within that network (Girvan and Newman,
Thus, to get a more nuanced understanding of the field, we decomposed the network into country clusters and communities of authors (Palla et al., 2005).

### 7.2.5 Country Clusters

Fig. 7.3 presents the main country clusters within the collaboration network. As indicated by the colors, the network divides into distinct clusters, with spatial and temporal variation in clustering visible. Three large distinct country clusters are uncovered within the time period 2000–2008, all of which comprise Northern and Southern partners. Four clusters are visible in the 2009–2017 time period, three of which comprise a mixture of Northern and Southern partners, one of Southern partners only. Although the clusters are globally dispersed to varying degrees, regions of spatial clustering are visible in all of them. While the countries within each of the clusters has changed over time, all have maintained this spatialized character. This in keeping with scholarship that has indicated that though the bias towards collaboration within territorial borders (regional, national, and linguistic) has decreased over time, spatial proximity remains an important determinant of research collaboration (Hoekman et al., 2010).

In terms of quantity or quality of collaborative connections, and location within the network, the country clusters are centered on a small group of (mainly Western) countries (Appendix Table 7.2 and 7.3). Many of the fisheries science powerhouses (as detailed in the previous section) have maintained central positions within the clusters, thus placing them in favorable positions with respect to control and dissemination of information. The most geographically expansive cluster is centered on the US. The second on North European countries—with Norway, for example, positioned as the best-connected country within that cluster. A further Europe-centered cluster is also evident, with France and Spain the prominent collaborators in this grouping. The fourth cluster which comprises partners from Africa, Asia, and the Middle East—among them some of the largest aquaculture producers in those regions (FAO, 2018)—is centered on Malaysia and Japan. That said, a number of smaller countries (e.g., Bulgaria, Tunisia, and Cambodia) are positioned favorably on the shortest path between other authors and thus may be playing important roles as knowledge brokers within their clusters (Newman, 2004).

### 7.2.6 Communities of Authors and their Topical Foci

Seeking a finer-grained analysis of the fisheries science landscape, we examined the communities of authors collaborating within the network, across three planes. This involved coupling the social network analysis techniques we have been utilizing thus far with topic modeling, thereby extending the inquiry beyond the spatial and temporal, and adding a topical dimension to the analysis. In excess of 3000 communities of
Figure 7.3: Country clusters ranked 1–4 based on the total number of countries within them for the periods 2000–2008 and 2009–2017.
authors were identified in the fisheries science network, which we ranked according to the number of authors within them. The distribution of the community size across the network is highly skewed, with the largest fifty communities comprising in excess of 80 percent of the authors in the network, whilst the remaining 20 percent is composed largely of sole authors, or groups of two to three authors (Appendix Fig. 7.14). Fig. 7.4 presents the largest fifteen (ranked 1–15) communities within the network, which together comprise almost 60 percent of the network (communities 16–30 can be viewed in Appendix Fig. 7.15). Though the communities are globally dispersed, all display dense points of regional centralization. Across many, rather than having diminished, this spatial clustering has intensified over time.
Figure 7.4: Spatial distribution of authors within communities for the period 2000–2008 and 2009–2017. Communities are ranked (1–15) based on the number of authors within the community, and differentiated by different colors. The size of the nodes represent the number of authors within the same location. Spatial distribution of communities 16–30 can be found in Appendix Fig. 7.15
CHAPTER 7. GLOBAL NETWORK OF FISHERIES SCIENCE

Figure 7.4: Continued with communities rank 11–15
To varying degrees, each of the fisheries science communities has grown in size over time, with the average number of connections each author has increasing (Appendix Fig. 7.16). The density across each of these fifteen communities is low however, indicating that although collaboration has increased, only a small number of the potential connections in each of the communities have been realized. This suggests that, when viewed at the individual level, the communities of authors in the fisheries science network are quite loosely knit. As discussed in relation to the global structure of the network, low levels of cohesion may be reflective of a number of things (e.g., fragmentation and division), which could have implications in terms of inhibiting knowledge exchange. In light of our analysis here, this pattern may indicate that though the network has become more collaborative, scholars are engaging in repeated collaborations, rather than forming new links beyond their existing connections (Leahey, 2016; Leahey and Reikowsky, 2008; Saetnan and Kipling, 2016). Our analysis at the country level would seem to support this finding. With respect to the interlinkages between these communities, the most frequent collaborative links are amongst the European, American, and Oceania communities (Fig. 7.5).

In order to gain a more substantive understanding of the manner in which the authors in the fisheries science network are grouping, we uncovered the topical foci of the network, alongside the community level and temporal variations therein (Fig. 7.6). In total, sixteen latent topics were identified within our corpus (Appendix Fig. 7.17): Age & growth; Aquaculture (growth effects); Aquaculture (health effects); Climate effects; Diet; Gear technology & bycatch; Genetics; Habitats; Immunogenetics; Management; Models (estimation & stock); Physiology; Reproduction; Salmonids; and Shellfish. By and large, these topical foci are relatively consistent with previous analysis of the content of fisheries science (Aksnes and Browman, 2016; Jarić et al., 2012; Syed et al., 2018a). This content has been discussed at length elsewhere, most recently by (Syed et al., 2018a), and is therefore not reported on in detail here. That said, uncovering these foci is instructive, as it allows us to investigate whether the fisheries science communities are clustered around a particular or similar topics (Clauset et al., 2004), how these may have changed, and how they might be related (Moody and Light, 2006).

In terms of community level and temporal variations in topics, while there are commonalities, and each of the communities contain authors that are engaged in the sixteen topics we have identified, variations in this respect are evident. For example, whilst our analysis indicates an almost across the board increase in publication output focused on Management, reflecting the increasing propensity of fisheries scientists in the West to focus their attention on managing human interactions with the natural environment, rather than managing fish per se (Bavington, 2010), the most intense focus on this topic is seen across the European, North American, and Oceania communities. In contrast, a much weaker focus on this topic is seen within the China-, Japan- and Iran-centered communities. Seeking to investigate this further, we calculated the similarities in cumulative topic distributions across the communities (Appendix Fig. 7.18). Our analysis reveals greater topical similarities among the North American, European and Oceania communities. A comparable pattern of similarity is discernible across the
Figure 7.5: Inter-community collaboration for the top-15 largest communities (ranked 1–15). The geographical label for each of the 15 communities indicates where the majority of authors within the community are spatially located.
topical output of a number of communities centered on newer entrants to the network (e.g., European South, Eastern Europe, Brazil, and Iran-centered communities).

Overall, a distinct geography of topics is detectable across the network. In this respect, a clear division of focus across the communities is seen whereby the topics Management, Models (estimation & stock), Gear technology & bycatch, and Habitats are areas of central, and in some instances intensifying focus for a number of the largest (Western-centered) communities in the network. On the other hand, a stronger topical leaning towards aquaculture-related topics (e.g., Aquaculture (effects on growth), Diet, Diseases, and Immunogenetics) is seen across the communities centered on the large emerging economies, many of whom are large aquaculture (and fish feed) producers (FAO, 2018). That said, a focus on aquaculture is also seen amongst European and North American communities that are concentrated in regions with large-scale interests in aquaculture production (e.g., Norway, US, Spain, Eastern Europe (Österblom et al., 2015; FAO, 2018)). For instance, we see a strong topical focus on Salmonids amongst a number of North America and North European communities, which may well be aquaculture related. In this regard, previous analysis has detected an increased focus on aquaculture species such as Atlantic salmon, and rainbow trout (Aksnes and Browman, 2016) (both of which are Salmonids) in fisheries science.
Figure 7.6: The proportion of topics published within each of the 15 largest communities for the periods 2000–2008 and 2009–2017. The number of publications published by each of the 15 communities, and for each time period, are shown in parentheses. For example, (1) Eur-Aus (4244) is the largest community (rank 1), with most of the authors spatially located in Europe and Australia, and 4,244 publications published between 2000–2008. The 4,244 publications cover aspects of Age & Growth for 7.7%, Aquaculture (growth effects) for 6.2%, and so on.
7.3 Discussion

As science has become increasingly internationalized, scholars investigating the shifting spatial structure underlying scientific practice (Hoekman et al., 2010), have posed questions as to whether networks of research collaboration are expanding in every region of the globe (Adams, 2012). Others have suggested that a globalized science may open up research fields in a generative manner, to new perspectives that challenge underlying assumptions, develop new methods, and point to previously unrecognized biases (Frickel et al., 2010). In this sense, scientific networks may be understood as reflecting not only authors, but people, actors, organizations, and things that uphold scientific patterns and beliefs, with different networks having different epistemological and ontological implications (Forsyth, 2003). The picture we have uncovered of fisheries science is in many respects similar to the broader trend in scientific output, which has led some scholars to suggest that the historically dominant ‘Atlantic Axis’ (an axis that has also been dominant in the production of fisheries science) is unlikely to be the main focus of research in the coming decades (Adams, 2012). In a sense, given China’s rapid growth over our timeframe, and its outpacing of the US in terms of total volume of scientific papers published in 2016 (Tollefson, 2018), this does seem likely. That said, it remains to be seen what this shift means for the production, shape, or order of knowledge (Escobar, 2004) in fisheries science.

7.3.1 A Bourdieusian Perspective

Though there seems to be broad acceptance that collaboration is a good thing (Adams, 2013; Katz and Martin, 1997), scholars have cautioned against viewing increased global collaboration as an unquestionable good (Adams, 2012; Katz and Martin, 1997; Leahey, 2016; Xie, 2014). For instance, as research team size and internationalism have become yet another metric against which science is judged (Xie, 2014), trends towards stratification in scientific collaboration patterns at both the institutional and individual levels have been detected (Dahdouh-Guebas et al., 2003; Jones et al., 2008; Xie, 2014). Regarding collaboration at the international level, scholars have argued that the manner in which the emerging geography of science is developing reflects historical patterns of western control and bias (Peters, 2006). Despite this, recent analysis has indicated that empirical work on collaboration tends to be heavily skewed towards the benefits of collaboration (Leahey, 2016; Xie, 2014). This may, we suggest, reflect a tendency of scholars to view scientific fields as largely consensual spaces. Failing to take seriously the role that power plays in shaping these spaces however, makes it very difficult to distinguish between cooperation based on equality and mutuality, and that which is based on domination and subordination (Albert and Kleinman, 2011).

In this sense, sociologists of science have convincingly shown that the structure of scientific knowledge in any field reflects a combination of micro- and macro-sociological factors (Bourdieu, 1975; Cetina, 1999; Forsyth, 2003; Law, 1987; Mol et al., 2002).
Adopting an explicitly Bourdieusian perspective helps us understand the role of power (understood as the capacity to define what legitimate science is) in these processes (Albert and Kleinman, 2011; Bourdieu, 1975, 1991). For example, in directing the topics pursued, methodologies adopted, journals in which research is published, or those we might collaborate with (Bourdieu, 1975). Conceiving of the field in this manner also helps us take seriously the role of consumers (e.g., policy makers, funding agencies, industry and so on) in determining the structure of the scientific field (Albert and Kleinman, 2011; Bourdieu, 1991). In this respect, historians of fisheries science have been astute in highlighting that much of fisheries science has been based on Western ideas about fish (Finley, 2011), and how fisheries might be managed (Bavington, 2010). Historians have also argued that the direction and structure of fisheries research has long been shaped by transitory economic and political forces (Bavington, 2010; Finley, 2011; Smith, 1994). This, it has been suggested, has provided the opportunity and encouragement for the development of research programs in certain areas over others, working to sideline longer-term economic, social and scientific goals, and limit the development of the scientific field in the process (Smith, 1994).

### 7.3.2 Democratizing Fisheries Science?

Considering these issues with respect to our analysis here, it is worth noting that although a spirit of internationalism has always animated the field (Hamblin, 2000; Rozwadowski, 2004), the bulk of fisheries science has long been produced by states with significant fishing interests around the globe (Finley, 2011; Smith, 1994). In this regard, while the geography of fisheries science (Adams, 2012) may have expanded, this pattern has not. The largest fisheries research nations (Aksnes and Browman, 2016), including the new entrants, are countries with highly industrialized fishing fleets or significant aquaculture interests (FAO, 2018; Kroodsma et al., 2018). Thus, whilst the arrival of new entrants might in one sense be seen as a shift towards an increasingly democratized global network of science (Xie, 2014), it may well work to marginalize some actors further (Jones et al., 2008; Xie, 2014). For instance, less “developed” countries with significant interests in fisheries in terms of food security, and livelihoods (Oliveira Júnior et al., 2016).

A number of regions remain marginal in this system despite increasing volumes of fisheries-related knowledge produced by authors in Asia, and Latin America (much of which is aquaculture related). For example, despite strong relative growth rates over the past decades (Aksnes and Browman, 2016), output from the Middle East is negligible when viewed at the macro-level. Similarly, standing as a stark reflection of the inequalities in output between developed and developing countries in this field (as in others) (Jarić et al., 2012), the African continent remains without any large hubs of production. Though the network has become more geographically extensive, this extension appears to be mirroring the shifting patterns of fisheries production, and the growing contribution of aquaculture to the global production of fisheries, much
of which is produced in Asia (FAO, 2018), rather than necessarily mirroring a shift towards an increasingly democratized global network of science (Forsyth, 2003; Xie, 2014).

In light of the existing inequality in publication output, it seems reasonable to suggest that the collaboration patterns we have depicted may serve to amplify inequalities within this domain. For instance, according to our analysis, large emerging economies and developed countries are the largest cross-border collaborators. Further, although prominent North-North and North-South collaborations are visible within the fisheries science network, South-South collaboration remains peripheral when evaluated from a global perspective (Leydesdorff et al., 2013). As is the case in other fields, much of the publication output being produced by developed countries displays an increasingly international character, whilst large swathes of the research being published in emerging economies remains entirely domestic (Adams, 2013; IOC-UNESCO, 2017) (see also Appendix Fig. 7.9). Given that internationally collaborated scientific papers are more likely to be published and cited, and are therefore more visible (Adams, 2012; Katz and Martin, 1997), these patterns could further side-line work by authors from countries who are already marginalized within this research system. These patterns may also work to reinforce dominant ways of thinking in the field towards perspectives from the Northern hemisphere (Forsyth, 2003), which have previously been cited as problematic within this domain (Francis, 1980).

7.3.3 Systems of Regionalization

The fisheries science landscape we have uncovered depicts a more regionalized than globalised system of knowledge production. In this regard, existing research has shown that scientific collaboration at the international level is shaped by the dynamic interplay of geographical, political, economic, historical, cultural and linguistic factors (Adams et al., 2014; Dahdouh-Guebas et al., 2003; Hoekman et al., 2010; Katz, 1994; Katz and Martin, 1997; Saetnan and Kipling, 2016). Our analysis suggests that a complex mix of these is at play in the fisheries science network. In line with work in other fields (Hoekman et al., 2010; Katz, 1994; Leahey, 2016; Parreira et al., 2017), even if collaboration has become increasingly internationalized, spatial proximity remains an important feature of the collaborative entanglements within the field, and this is seen across the country clusters (Fig. 7.3) and communities of authors (Fig. 7.4) uncovered within the network. This feature of collaboration is in itself reflective of an array of overlapping factors. Among these are regional political groupings, e.g. trade blocs (Parreira et al., 2017), funding mechanisms or opportunities that remain at the national and regional levels (Hoekman et al., 2010), colonial ties (Adams et al., 2014), and no doubt overlapping fishing interests—proximate and distant.

Considering this latter point further, it has been suggested that different scientific fields might have specific “spatial requirements” due to their research topics (Hoekman et al., 2010). For example, collaborative proximity may be due to environmental similarities
among countries. It may therefore make sense that researchers focused on similar geographical areas or biomes would work together (Parreira et al., 2017). As regards to fisheries, this seems reasonable given that many countries share closely overlapping fishing grounds and thus proximate fisheries interests, across shared ecoregions. Indeed, historians have shown that the requirements of the marine environment—for instance, the de-territorializing impulse of fish and the sea (Bear, 2013)—have historically been amongst the drivers of internationalization in the field (Hamblin, 2000; Rozwadowski, 2004). However, our analysis suggests that distant fishing interests also breed collaboration, as do distant colonial ties. For instance, with respect to the country clusters, the fishing interests of France and Spain which extend along the Eastern Tropical Atlantic, and Western Indian Ocean (Campling, 2012), or those of the US that extend into Pacific region (Hamblin, 2000; Havice, 2018), might reasonably be highlighted. As a further example, the concentrated research links that France has with its former colonies in North-West and West Africa, have been well documented (Adams et al., 2014).

### 7.3.4 Collaboration Styles

As indicated, sociologists of science have also stressed the role of micro-level characteristics in shaping the structure of scientific fields (Bourdieu, 1975). In this regard, an additional driver of collaboration highlighted in the literature is preferential attachment at the individual level (Wagner and Leydesdorff, 2005). For example, studies have indicated that authors have a tendency to collaborate with “like-minded others” (Leahey, 2016), which may lead to a particular style of collaboration. In this respect, whilst we have detected an increasingly collaborative field, albeit a regionalized one, our analysis of the fisheries science network has identified structural characteristics that indicate the style of collaboration authors are engaging in is a thin one. This pattern may reflect the tendency of scholars to work within their own networks, rather than forming new links beyond those (Saetnan and Kipling, 2016), and may be driven by an array of factors. For instance, scholars might engage in repeat collaborations—which offers returns in terms of trust building, and increased certainty—in an effort to mitigate against the cost of collaboration. They might also select collaborators from within their own specialty areas, who share areas of expertise, methodological or theoretical perspectives (Leahey, 2016).

In a sense, given the increasingly specialized nature of science (Casadevall and Fang, 2014), including fisheries science (Mather et al., 2008), we might expect this. Indeed, existing research has highlighted that specialization and collaboration in science are not unrelated (Leahey and Reikowsky, 2008). On the one hand, it is precisely this specialization that is driving the need for collaboration (Casadevall and Fang, 2014; Leahey and Reikowsky, 2008). On the other, specialization has been found to inform collaboration strategies, with scientists often having a tendency to engage in within-specialty collaboration rather than complementary collaboration that spans
boundaries (Leahey and Reikowsky, 2008). A Bourdieusian perspective helps us to understand the role that power may play in directing these choices, and thus the structure of the scientific field in a broad sense (Bourdieu, 1975). This might, for example, explain why research has found that the steps to interdisciplinary science over the past three decades have actually been very small, oftentimes drawing on neighboring fields and only modestly increasing the connections to areas further afield (Porter and Rafols, 2009). The danger with such a strategy, however, is that it can become a reinforcing style of collaboration (Leahey, 2016), which may have potential costs in terms of the production of novel information, and hindering exposure to heterogeneous ideas (Blondel et al., 2008). Given that much advancement in fisheries science has been cited as coming from the branches of the discipline rather than the roots (Francis, 1980), this pattern may work to limit the development of the field in a direction that may equip it to address some of the ongoing challenges in the field.

### 7.3.5 The Topical Landscape of Fisheries Science

A distinct geography of topics has been detected across the field. Though likely reflective of a combination of the macro- and micro-sociological characteristics we have discussed, this geography is further suggestive of the political and economic influences directing the research priorities in this field, and the continuing dominance of specific ideas about fish and fisheries within this space. Unsurprisingly, given that Western fisheries management has been built on, and remains based upon, calculations of maximum sustainable yield and the allocation of quotas (Campling et al., 2012; Finley, 2011; Nielsen and Holm, 2007; Smith, 1994; Winder, 2018), a number of the largest communities in the network remain heavily focused on Models (estimation & stock). Fisheries scientists themselves have suggested, that here more attention has been paid to developing sophisticated ways of fitting analytic models, than to the actual assumptions underpinning these models (Francis, 1980). This in itself, is not unrelated to the demands on fisheries scientists to provide numbers for policy. Similarly, reflecting the heavy spotlight on discarding in fisheries over the past two decades or so (Alverson et al., 1994; Borges, 2015; Kelleher, 2005), Gear technology & bycatch is a further area of strong topical focus for the largest community of authors within the network. Further, that a large proportion of the publication output of fisheries science is increasingly commandeared by aquaculture-related topics is not unrelated to the rapid investment and consequent expansion in production this area has seen (Aksnes and Browman, 2016; Winder, 2018). As capture fisheries have continued to diminish, this area has been given increasing priority by both fisheries managers (Bavington, 2010; Winder, 2018), and governments as a growth strategy under the rubric of ‘blue growth’ (Barbesgaard, 2018; Hadjimichael, 2018; Winder, 2018; Winder and Le Heron, 2017), and this is reflected not only in the topical foci we have uncovered in the fisheries science network, but across the entire structure of the network.
7.4 Limitations and Ways Forward

There are several limitations of our study that may adversely influence our findings. The more technical methodological limitations have been detailed in the methods section of the paper, but here we discuss the more general limitations and suggest possible avenues for further research that might address these. In terms of our corpus, whilst this comprises high ranking journals in fisheries science, we acknowledge that it does not capture the entire spectrum of work being done in the area of fisheries. The fisheries category in Web of Science (WOS) is delimited to SCIE, and thus skewed towards work coming from the natural sciences. Though there are few dedicated fisheries journals in the social sciences or arts and humanities, a number do exist beyond this category (e.g., Marine Policy, Ocean and Coastal Management). It is also important to highlight also that important fisheries related work is published beyond these specialized outlets. For example, work which is oriented towards the social sciences or humanities may be published in journals that have a more general focus, or a stronger leaning towards those sciences (e.g., Journal of Agrarian Change). Further, by virtue of being confined to scientific journal articles, our analysis does not include books or grey literature. It has been suggested, in this respect, that a significant proportion of fisheries-related research is published by national institutes, and may not be visible in the scientific literature (Aksnes and Browman, 2016). In conjunction, as we have touched upon, the results of our analysis are based entirely on English publications, and thus reflect an Anglophone bias (King, 2004). For instance, with respect to our findings, Russia’s absence, as the fourth largest producer of capture fisheries in the world (FAO, 2018) is suggestive here. Though this is likely reflective of a number of factors (Bornmann et al., 2015), this is at least in part a reflection of the linguistic bias of our dataset, and thus an underrepresentation. This is likely the case for other linguistic communities across the field. However, in this regard, it is not irrelevant that this bias is reflected across science, and thus likely plays a role in structuring the shape of science (King, 2004), including scientific collaborations.

Boundaries have to be drawn in all research. However, shifting the boundaries we have applied here, and drawing on an expanded dataset that captures a wider spread of fisheries-related work, might prove a fruitful avenue for further research and may overcome a number of the limitations we have detailed with respect to this study. Doing so would go a long way to capturing a broader spectrum of fisheries-related knowledge, beyond the traditional domain of ‘fisheries science’. We also suggest that a more in-depth analysis regarding the styles and drivers of collaboration, and in terms of collaboration inequalities in this field are avenues of inquiry that are worthy of pursuit. In this regard, our analysis could be expanded further by focusing on the funding sources underpinning fisheries science. For instance, studies have indicated that one factor that works to bias research activities to core regions, particularly when funding is intended to serve the interests of national or regional research performing entities, is unequal funding opportunities (Hoekman et al., 2010). Investigating this with respect to fisheries science, drawing on an extended corpus, is a specific next step to developing this
7.5 Conclusion

Broad-based collaboration, it is argued, is crucial to solving the challenges ongoing with respect to fisheries. In light of this ‘collaboration imperative’, we have mapped and examined the landscape of scientific collaboration across the field of fisheries science. Overall, our analysis has presented a shifting field that has become increasingly collaborative, though less cohesive, with a number of key players maintaining hegemonic positions within the network. By and large, the most productive (and collaborative) countries in terms of fisheries science are those which have large industrialized fisheries-related interests, many of them global in nature. Although the collaboration network has become more extensive, it has also become more intensive in places, with a clear spatial pattern evident in the structure of scientific collaborations across the field. In this respect, the fisheries science landscape is one whereby the centers of knowledge production and the connections between them display trends more akin to regionalization than globalization.

Some of the characteristics of the network suggest that authors across the field may be engaging in a repeat, rather than broad style of collaboration, which may work as a reinforcing mechanism with respect to the knowledge that is produced by the field. This pattern is likely to limit the potential gains of collaboration, and could have consequences in terms of pushing the boundaries of fisheries science in new and fruitful ways, in a manner which may help address some of the ongoing challenges within the field. Though likely shaped by an array of both micro- and macro-sociological factors, the patterns of collaboration, and the geography of topics uncovered across the field betray a number of political-economic influences, which merit reflection by both policy makers and scientists alike.

7.6 Materials and Methods

In order to map the landscape of fisheries science collaborations, we employed a combination of social network analysis techniques and topic modeling. Similar approaches have previously been used to characterize scientific fields, both in terms of their collaborative structures (Moody, 2004; Newman, 2001, 2004) and topical foci (Syed et al., 2018a; Syed and Weber, 2018). Our analysis of the network was conducted at progressively finer levels of granularity (Ding, 2011), progressing from the overall macro-structure to the hidden subgroups within the network. In doing so, we examined at each level of granularity both the structure (links) and individual properties (nodes) of the network (McPherson et al., 2001). Further, given that social networks are not static
over time, but rather evolve (Bodin and Crona, 2009), alongside analyzing the entire network, we examined our dataset at two-time intervals: 2000–2008 and 2009–2017.

7.6.1 Data Collection

Fisheries science publications were selected based on the fisheries category as defined by the Science Citation Index Extended (SCIE). This category spans a list of 50 journals covering all aspects of fisheries science, technology, and industry. All 50 journals (Appendix Table 7.4) were included, and all articles published between 2000 and 2017 were selected. The Scopus developer API was subsequently utilized to extract article data such as abstracts, authors, and affiliations. Specifically, the Scopus Abstract Retrieval API provides all (meta-) data associated with a particular article. The Scopus unique identifier for authors and affiliations was used to disambiguate authors and affiliations with identical names, and to merge the same author with different names. For affiliations without an affiliation ID, a surrogate key was constructed by concatenating all parts of the affiliation address. A filtering process was used to exclude non-English articles and those that did not constitute a research article (such as errata) or contained no abstract. A total of 73,240 articles were deemed fit for further analysis, with a total of 106,137 authors and 100,175 affiliations. The Google Geocoding API was used to convert affiliation addresses into geographic coordinates (latitude-longitude).

7.6.2 Social Network Analysis

The co-author networks were constructed by linking two authors (i.e., nodes) on the basis of co-authorship (i.e., edges). The frequency of collaborations between two authors defined the weight of the edge spanning the two nodes. The resulting network was subsequently analyzed utilizing social network analysis, which provides an array of statistics (Appendix Table 7.5) for doing so (Leydesdorff and Wagner, 2008; Newman, 2001).

7.6.3 Hidden Groups

Most real networks contain groups within which the nodes are more tightly connected to each other than the rest of the network, oftentimes referred to as clusters or communities (Palla et al., 2005). These groupings might be connected in various ways (e.g., topic, location, history etc.), with studies indicating that links are often homophilous (McPherson et al., 2001). Uncovering these a priori unknown groups allows for the identification of functional units within a system (Blondel et al., 2008), alongside their structural properties (Newman, 2012a), which can vary widely, and
may—with respect to our interests here—be consequential in terms of knowledge creation (Granovetter, 1973; Lambiotte and Panzarasa, 2009). Thus, utilizing community detection techniques in an effort to get a more nuanced understanding of the network, we decomposed the network into country clusters and communities of authors. To better characterize these hidden groups, we utilized centrality measures which are reflective of the importance and effectiveness of particular nodes within the network (Bodin and Crona, 2009; Freeman, 1978; Newman, 2012a).

### 7.6.4 Community Detection

To detect community structures, we used the Louvain algorithm (Blondel et al., 2008), extended with a time parameter to allow for community detection at various resolutions (Lambiotte et al., 2014). The inclusion of a time parameter increases community stability and aims to ameliorate community size bias (Fortunato and Barthelemy, 2007). In a comparative study (Lancichinetti and Fortunato, 2009), the Louvain community detection algorithm was found to have ‘excellent performance’ on several classes of benchmark graphs (Girvan and Newman, 2002; Lancichinetti and Fortunato, 2009), although benchmark performance may not necessarily align with broader real-world situations (Newman, 2012a). We performed a grid-search (Appendix Fig. 7.19) on the parameter space of the resolution parameter (from 0.1 to 2.0 in steps of 0.1) and due to the heuristic nature of the Louvain algorithm, conducted 10 different random initializations for each grid-search. In doing so, we aimed to find communities with high modularity (Newman, 2003; Newman and Girvan, 2004), a measure that quantifies the quality of the detected communities. A similar process was performed to detect communities or clusters of countries (Appendix Fig. 7.20 and 7.21). The inter-community collaboration was measured by the (weighted) edges traversing communities within an induced community graph, where communities are represented as nodes themselves. The frequency of domestic and international collaborations between countries was quantified by creating an adjacency matrix of co-authors and their country affiliations. A similar process was performed to quantify collaborations between affiliations. The domestic collaborations were restricted to co-authors with the same country affiliation.

Detecting communities in networks (e.g., social, biological, citation, metabolic networks) can generally be classified into discovering non-overlapping communities where each node belongs to a single community (Blondel et al., 2008; Clauset et al., 2004; Decelle et al., 2011a,b; Fortunato, 2010; Hofman and Wiggins, 2008; Newman and Girvan, 2004; Newman and Leicht, 2007; Nowicki and Snijders, 2001), or overlapping communities, where nodes can belong to several communities (Ahn et al., 2010; Airoldi et al., 2008; Ball et al., 2011; Derényi et al., 2005; Gopalan and Blei, 2013; Gregory, 2010; Lancichinetti et al., 2011; Viamontes Esquivel and Rosvall, 2011). Increasingly, real-world networks can be characterized as overlapping (Palla et al., 2005), and the most general formulation of a community detection algorithm should ideally
include both overlapping and non-overlapping communities (Ball et al., 2011). A major drawback, however, of overlapping community detection algorithms is that the number of communities within a network needs to be known in advance (Ball et al., 2011). Typically, this number is unknown, although recent studies have attempted to apply Bayesian inference and Monte Carlo methods to estimate this number (Newman and Reinert, 2016; Riolo et al., 2017). However, a successful application of such methods highly depends on the choice of an appropriate prior probability. Community detection algorithms based on modularity maximization (a quality index for partitioning networks into communities) circumvent this drawback, but might result in a bias of the community sizes it uncovers (Ball et al., 2011; Bickel and Chen, 2009; Fortunato and Barthelemy, 2007). Typically it fails to find very small communities. The Louvain algorithm (Blondel et al., 2008), used in this study uses such modularity maximization, and the number of communities as well as the division into communities are performed automatically. However, the Louvain algorithm treats communities as disjoint (non-overlapping), forming a technical methodological limitation with respect to our study. Thus, explicitly identifying nodes that bridge communities could be an interesting directive for future research.

### 7.6.5 Topic Modeling

To uncover latent topics, the topic model Latent Dirichlet Allocation (LDA) (Blei, 2012; Blei et al., 2003) was used. All pre-processing steps to suitably prepare documents for statistical topical inference (Hoffman et al., 2010) are described in our previous work (Syed et al., 2018a), which are highly optimized for the fisheries domain (Syed and Spruit, 2017, 2018a). The number of topics uncovered was determined by performing a grid search on the parameter space of the various LDA hyper-parameters (Appendix Fig. 7.22), and the quality of topics calculated by a topic coherence score (Röder et al., 2015). For readability, topics were labeled by fisheries domain experts via close inspection of the topic's top words (Appendix Fig. 7.17), titles (Appendix Table 7.6), abstracts, and a visual representation through multidimensional scaling (Appendix Fig. 7.23) (Chuang et al., 2012; Sievert and Shirley, 2014). The Hellinger distance (Hellinger, 1909) was used to calculate the similarities between cumulative topic distributions of communities (Appendix Fig. 7.18).
Appendix
Figure 7.7: Social network analysis metrics obtained from the full network of 106,173 authors from 73,240 publications from 2000–2017. Explanation of metrics provided in Table 7.5
Figure 7.8: The frequency counts of publications and journals included in the dataset ranging from 2000–2017.

Figure 7.9: The collaboration frequency counts of domestic and international country collaborations for the time frame 2000–2008 and 2009–2017. Only the top-25 largest collaborating countries are shown, sorted clockwise.
Figure 7.10: The collaboration frequency counts of only the international collaborations amongst affiliations for the time frame 2000–2008. Only the top-10 percent highest frequency links for the top-25 largest collaborating affiliations are shown.
2009–2017

Figure 7.11: Similar to Fig. 7.10 but for the time frame 2009–2017.
Figure 7.12: The collaboration frequency counts of the domestic as well as the international collaborations amongst affiliations for the time frame 2000–2008. Only the top-10 percent highest frequency links for the top-25 largest collaborating affiliations are shown.
Figure 7.13: Similar to Fig. 7.12 but for the time frame 2009–2017.
Figure 7.14: Community size distribution (top) for the top-50 largest communities (by number of authors) sorted by rank; rank 1 being the largest, rank 2 the second largest, and so on. The bottom figure displays the community size cumulative percentage in relation to all the authors (nodes) in the network. For example, the top-10 (i.e., rank 1-10) communities account for over 40% of the total number of authors in the network.
Figure 7.15: Spatial distribution of communities ranked 16–30 differentiated by color, with the size of the node representing the number of authors.
Figure 7.15: Continued with communities ranked 26-30
Figure 7.16: Social network metrics for the top-15 largest communities sorted by rank for the two time intervals (2000–2008 and 2009–2017).
Figure 7.17: Topic-word distributions of the top-10 high-probability words for the 16 uncovered latent fisheries science topics. Each topic is labeled (top) with a logical topic description that best captures the semantics of the top words.
Figure 7.18: The Hellinger distance between cumulative topic distributions of the 15 largest communities. The Hellinger distance is a symmetrical distance measure of two probability distributions. Smaller distances indicate that two cumulative topic distributions are more similar, meaning that the two communities publish more similar work in terms of latent topics.
### CHAPTER 7. GLOBAL NETWORK OF FISHERIES SCIENCE

#### Figure 7.19:
Modularity scores obtained by performing a grid search on the resolution parameter, ranging from 0.1 to 2.0 as shown on the top, and conducting ten random initializations (run-1 to run-10 shown on the left side) to create author communities (i.e., clusters) for the time frame 2000–2008. A resolution parameter of 1.2 with the second random initializations (run-2) provided author clusters with the highest modularity value.

#### Figure 7.20:
Modularity scores obtained by performing a grid search on the resolution parameter (ranging from 0.1 to 2.0) and conducting ten random initializations (run-1 to run-10) to create country communities (i.e., clusters) for the time frame 2000–2008. Resolution 1.1 with run-3 provides country clusters with the highest modularity value.
Figure 7.21: Similar to Fig. 7.20 but for the time frame 2009–2017. Resolution 1.0 with run-1 provides country clusters with the highest modularity value.
Figure 7.22: Coherence scores for all created LDA models by performing a grid search on the number of topics or k-parameter (shown on top), and varying the different LDA hyper-parameters (shown on left). $p =$ number of epochs over the corpus, $\text{max } =$ convergence iteration parameter for the expectation step (EM-algorithm), $\alpha =$ symmetrical (sym.) or asymmetrical (asym.) Dirichlet prior distribution on the topic probabilities within documents (i.e. document-topic proportions). $\eta =$ symmetrical (sym.) or asymmetrical (asym.) Dirichlet prior distribution on the word probabilities within topics (i.e. topic-word proportions). A higher coherence score can be viewed as a better LDA model.
**Figure 7.23:** Inter-topic distance map that shows a two-dimensional representation (via multidimensional scaling) of the 16 uncovered fisheries science topics with labels. The distance between the nodes represents the topic similarity with respect to the distributions of words (i.e. latent Dirichlet allocation’s output). The surface of the nodes indicates the topic prevalence within the corpus, with bigger nodes representing topics being more prominent within the document collection (all nodes add up to 100%).
Table 7.1: The number of absolute and relative publications published between 2000–2008 and 2009–2017 for the 25 countries with the highest publication output. Differences in percentages and absolute values are presented in the total columns. A single publication is fractionally credited based on the number of authors and country affiliations.

<table>
<thead>
<tr>
<th>country</th>
<th>2000-2008 publications</th>
<th>percentage</th>
<th>2009-2017 publications</th>
<th>percentage</th>
<th>total publications</th>
<th>percentage diff.</th>
<th>publication diff.</th>
<th>total publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>8,497</td>
<td>27.2</td>
<td>8,950</td>
<td>21.4</td>
<td>16,994</td>
<td>-5.8</td>
<td>453</td>
<td>16,994</td>
</tr>
<tr>
<td>Japan</td>
<td>2,674</td>
<td>8.6</td>
<td>1,959</td>
<td>4.6</td>
<td>5,348</td>
<td>-3.9</td>
<td>-715</td>
<td>5,348</td>
</tr>
<tr>
<td>Canada</td>
<td>2,220</td>
<td>7.1</td>
<td>2,014</td>
<td>4.8</td>
<td>4,440</td>
<td>-2.3</td>
<td>-206</td>
<td>4,440</td>
</tr>
<tr>
<td>Australia</td>
<td>2,025</td>
<td>6.5</td>
<td>2,344</td>
<td>5.6</td>
<td>4,050</td>
<td>-0.9</td>
<td>319</td>
<td>4,050</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>1,548</td>
<td>5.0</td>
<td>1,223</td>
<td>2.9</td>
<td>3,096</td>
<td>-2.0</td>
<td>-325</td>
<td>3,096</td>
</tr>
<tr>
<td>Norway</td>
<td>1,423</td>
<td>4.6</td>
<td>1,468</td>
<td>3.5</td>
<td>2,845</td>
<td>-1.0</td>
<td>46</td>
<td>2,845</td>
</tr>
<tr>
<td>Spain</td>
<td>1,108</td>
<td>3.6</td>
<td>1,560</td>
<td>3.7</td>
<td>2,216</td>
<td>0.2</td>
<td>452</td>
<td>2,216</td>
</tr>
<tr>
<td>France</td>
<td>949</td>
<td>3.0</td>
<td>894</td>
<td>2.1</td>
<td>1,898</td>
<td>-0.9</td>
<td>-55</td>
<td>1,898</td>
</tr>
<tr>
<td>China</td>
<td>857</td>
<td>2.7</td>
<td>5,157</td>
<td>12.3</td>
<td>4301</td>
<td>9.6</td>
<td>4301</td>
<td>1,714</td>
</tr>
<tr>
<td>New Zealand</td>
<td>769</td>
<td>2.5</td>
<td>657</td>
<td>1.5</td>
<td>1,538</td>
<td>-0.9</td>
<td>-112</td>
<td>1,538</td>
</tr>
<tr>
<td>India</td>
<td>611</td>
<td>2.0</td>
<td>1,504</td>
<td>3.6</td>
<td>893</td>
<td>1.6</td>
<td>323</td>
<td>1,222</td>
</tr>
<tr>
<td>Mexico</td>
<td>563</td>
<td>1.8</td>
<td>886</td>
<td>2.1</td>
<td>1,126</td>
<td>0.3</td>
<td>323</td>
<td>1,126</td>
</tr>
<tr>
<td>Germany</td>
<td>472</td>
<td>1.5</td>
<td>590</td>
<td>1.4</td>
<td>944</td>
<td>-0.1</td>
<td>118</td>
<td>944</td>
</tr>
<tr>
<td>Italy</td>
<td>462</td>
<td>1.5</td>
<td>667</td>
<td>1.5</td>
<td>924</td>
<td>0.1</td>
<td>205</td>
<td>924</td>
</tr>
<tr>
<td>Taiwan</td>
<td>455</td>
<td>1.5</td>
<td>470</td>
<td>1.2</td>
<td>909</td>
<td>-0.3</td>
<td>15</td>
<td>909</td>
</tr>
<tr>
<td>Portugal</td>
<td>415</td>
<td>1.3</td>
<td>564</td>
<td>1.3</td>
<td>831</td>
<td>0.0</td>
<td>148</td>
<td>831</td>
</tr>
<tr>
<td>Brazil</td>
<td>403</td>
<td>1.3</td>
<td>1,673</td>
<td>4.0</td>
<td>806</td>
<td>2.7</td>
<td>1270</td>
<td>806</td>
</tr>
<tr>
<td>South Korea</td>
<td>373</td>
<td>1.2</td>
<td>699</td>
<td>1.6</td>
<td>746</td>
<td>0.5</td>
<td>327</td>
<td>746</td>
</tr>
<tr>
<td>Denmark</td>
<td>344</td>
<td>1.1</td>
<td>434</td>
<td>1.0</td>
<td>689</td>
<td>-0.1</td>
<td>89</td>
<td>689</td>
</tr>
<tr>
<td>Turkey</td>
<td>311</td>
<td>1.0</td>
<td>868</td>
<td>2.0</td>
<td>658</td>
<td>1.1</td>
<td>558</td>
<td>621</td>
</tr>
<tr>
<td>Finland</td>
<td>287</td>
<td>0.9</td>
<td>204</td>
<td>0.4</td>
<td>573</td>
<td>-0.4</td>
<td>-83</td>
<td>573</td>
</tr>
<tr>
<td>Greece</td>
<td>270</td>
<td>0.9</td>
<td>248</td>
<td>0.5</td>
<td>539</td>
<td>-0.3</td>
<td>-22</td>
<td>539</td>
</tr>
<tr>
<td>Poland</td>
<td>268</td>
<td>0.9</td>
<td>371</td>
<td>0.8</td>
<td>537</td>
<td>0.0</td>
<td>103</td>
<td>537</td>
</tr>
<tr>
<td>Sweden</td>
<td>256</td>
<td>0.8</td>
<td>270</td>
<td>0.6</td>
<td>513</td>
<td>-0.2</td>
<td>13</td>
<td>513</td>
</tr>
<tr>
<td>Netherlands</td>
<td>255</td>
<td>0.8</td>
<td>267</td>
<td>0.6</td>
<td>510</td>
<td>-0.2</td>
<td>12</td>
<td>510</td>
</tr>
<tr>
<td>TOTAL TOP 25</td>
<td>27,814</td>
<td>89.1</td>
<td>35,941</td>
<td>85.9</td>
<td>55,628</td>
<td>-3.2</td>
<td>8127</td>
<td>55,628</td>
</tr>
</tbody>
</table>
Table 7.2: The centrality measures for the four country clusters (ranked 1-4 based on the numbers of countries within them) within the period 2000–2008. Only the top 5 countries with the highest centrality value for each of the four country clusters is shown. Description of centrality measures can be found in Table 7.5

<table>
<thead>
<tr>
<th>Rank</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>United States</td>
<td>0.81</td>
<td>United States</td>
<td>0.68</td>
<td>United States</td>
<td>0.5</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>0.43</td>
<td>Spain</td>
<td>0.83</td>
<td>France</td>
<td>0.57</td>
<td>Spain</td>
<td>0.5</td>
<td>France</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>Greenland</td>
<td>0.84</td>
<td>United Kingdom</td>
<td>0.81</td>
<td>United Kingdom</td>
<td>0.6</td>
<td>Norway</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>0.7</td>
<td>Jamaica</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.69</td>
<td>Australia</td>
<td>0.64</td>
<td>Canada</td>
<td>0.44</td>
<td>Japan</td>
<td>0.76</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>0.69</td>
<td>France</td>
<td>0.66</td>
<td>Belgium</td>
<td>0.49</td>
<td>Italy</td>
<td>0.67</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>India</td>
<td>0.57</td>
<td>Portugal</td>
<td>0.49</td>
<td>Italy</td>
<td>0.65</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>France</td>
<td>0.57</td>
<td>Spain</td>
<td>0.55</td>
<td>Spain</td>
<td>0.67</td>
<td>Jamaica</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.76</td>
<td>United States</td>
<td>0.56</td>
<td>Australia</td>
<td>0.46</td>
<td>United States</td>
<td>0.56</td>
<td>Australia</td>
</tr>
<tr>
<td>2</td>
<td>0.69</td>
<td>France</td>
<td>0.49</td>
<td>Belgium</td>
<td>0.44</td>
<td>Canada</td>
<td>0.44</td>
<td>Canada</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>India</td>
<td>0.49</td>
<td>Portugal</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Japan</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>France</td>
<td>0.55</td>
<td>Spain</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Japan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.81</td>
<td>United States</td>
<td>0.68</td>
<td>United States</td>
<td>0.68</td>
<td>United States</td>
<td>0.5</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>0.83</td>
<td>France</td>
<td>0.57</td>
<td>Spain</td>
<td>0.57</td>
<td>Spain</td>
<td>0.5</td>
<td>France</td>
</tr>
<tr>
<td>3</td>
<td>0.84</td>
<td>Greenland</td>
<td>0.81</td>
<td>United Kingdom</td>
<td>0.81</td>
<td>United Kingdom</td>
<td>0.6</td>
<td>Norway</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>0.7</td>
<td>Jamaica</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
<th>Betweenness Centrality value</th>
<th>Betweenness Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>United States</td>
<td>0.81</td>
<td>United States</td>
<td>0.68</td>
<td>United States</td>
<td>0.5</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>0.43</td>
<td>Spain</td>
<td>0.83</td>
<td>France</td>
<td>0.57</td>
<td>Spain</td>
<td>0.5</td>
<td>France</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>Greenland</td>
<td>0.84</td>
<td>United Kingdom</td>
<td>0.81</td>
<td>United Kingdom</td>
<td>0.6</td>
<td>Norway</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>0.7</td>
<td>Jamaica</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
<th>Closeness Centrality value</th>
<th>Closeness Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.69</td>
<td>Australia</td>
<td>0.64</td>
<td>Canada</td>
<td>0.44</td>
<td>Japan</td>
<td>0.76</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>0.69</td>
<td>France</td>
<td>0.66</td>
<td>Belgium</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Canada</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>India</td>
<td>0.49</td>
<td>Portugal</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Canada</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>France</td>
<td>0.55</td>
<td>Spain</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
<th>Degree Centrality value</th>
<th>Degree Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.76</td>
<td>United States</td>
<td>0.56</td>
<td>Australia</td>
<td>0.46</td>
<td>United States</td>
<td>0.56</td>
<td>Australia</td>
</tr>
<tr>
<td>2</td>
<td>0.69</td>
<td>France</td>
<td>0.44</td>
<td>Canada</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Canada</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>India</td>
<td>0.43</td>
<td>Portugal</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Canada</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>France</td>
<td>0.43</td>
<td>Spain</td>
<td>0.43</td>
<td>Japan</td>
<td>0.43</td>
<td>Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
<th>Eigenvector Centrality value</th>
<th>Eigenvector Centrality country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.81</td>
<td>United States</td>
<td>0.68</td>
<td>United States</td>
<td>0.68</td>
<td>United States</td>
<td>0.5</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>0.83</td>
<td>France</td>
<td>0.57</td>
<td>Spain</td>
<td>0.57</td>
<td>Spain</td>
<td>0.5</td>
<td>France</td>
</tr>
<tr>
<td>3</td>
<td>0.84</td>
<td>Greenland</td>
<td>0.81</td>
<td>United Kingdom</td>
<td>0.81</td>
<td>United Kingdom</td>
<td>0.6</td>
<td>Norway</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>1</td>
<td>Jamaica</td>
<td>0.7</td>
<td>Jamaica</td>
</tr>
</tbody>
</table>

199
Table 7.3: The centrality measures for the four country clusters (ranked 1–4 based on the numbers of countries within them) within the period 2009–2017. Only the top-5 countries with the highest centrality value for each of the four country cluster is shown. Description of centrality measures can be found in Table 7.5

<table>
<thead>
<tr>
<th>Rank</th>
<th>Betweenness Centrality</th>
<th>Closeness centrality</th>
<th>Degree centrality</th>
<th>Eigenvector centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>value</td>
<td>country</td>
<td>value</td>
<td>country</td>
</tr>
<tr>
<td>1</td>
<td>0.26</td>
<td>United States</td>
<td>0.88</td>
<td>United States</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>Canada</td>
<td>0.73</td>
<td>Australia</td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>Mexico</td>
<td>0.65</td>
<td>Canada</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>Australia</td>
<td>0.63</td>
<td>Mexico</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>Barbados</td>
<td>0.62</td>
<td>South Africa</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>Tunisia</td>
<td>0.79</td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>United Kingdom</td>
<td>0.77</td>
<td>United Kingdom</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>Peru</td>
<td>0.69</td>
<td>Spain</td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>France</td>
<td>0.68</td>
<td>Italy</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>Monaco</td>
<td>0.68</td>
<td>Portugal</td>
</tr>
<tr>
<td>3</td>
<td>0.44</td>
<td>Bulgaria</td>
<td>0.79</td>
<td>Belgium</td>
</tr>
<tr>
<td></td>
<td>0.29</td>
<td>Belgium</td>
<td>0.78</td>
<td>Germany</td>
</tr>
<tr>
<td></td>
<td>0.19</td>
<td>Lithuania</td>
<td>0.75</td>
<td>Norway</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>Russian Federation</td>
<td>0.72</td>
<td>Netherlands</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>Benin</td>
<td>0.71</td>
<td>Russian Federation</td>
</tr>
<tr>
<td>4</td>
<td>0.41</td>
<td>Cambodia</td>
<td>0.81</td>
<td>Malaysia</td>
</tr>
<tr>
<td></td>
<td>0.29</td>
<td>Thailand</td>
<td>0.74</td>
<td>Japan</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>Iran, Islamic Republic of</td>
<td>0.74</td>
<td>Thailand</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>Egypt</td>
<td>0.72</td>
<td>India</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>Philippines</td>
<td>0.71</td>
<td>Bangladesh</td>
</tr>
</tbody>
</table>
Table 7.4: The complete list of journals covered by the fisheries category as defined by the Science Citation Index Extended (SCIE) 2016–2017. This category spans a list of 50 journals covering all aspects of fisheries science, technology, and industry. All 50 journals were included in the dataset. IF = Impact Factor, N = Number of Publications.

<table>
<thead>
<tr>
<th>RANK</th>
<th>JOURNAL NAME</th>
<th>IF</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fish and Fisheries</td>
<td>9.013</td>
<td>525</td>
</tr>
<tr>
<td>2</td>
<td>Reviews in Aquaculture</td>
<td>4.618</td>
<td>195</td>
</tr>
<tr>
<td>3</td>
<td>Reviews in Fish Biology and Fisheries</td>
<td>3.575</td>
<td>588</td>
</tr>
<tr>
<td>4</td>
<td>Fish &amp; Shellfish Immunology</td>
<td>3.148</td>
<td>4530</td>
</tr>
<tr>
<td>5</td>
<td>Fisheries</td>
<td>3.000</td>
<td>503</td>
</tr>
<tr>
<td>6</td>
<td>Aquaculture Environment Interactions</td>
<td>2.905</td>
<td>161</td>
</tr>
<tr>
<td>7</td>
<td>ICES Journal Of Marine Science</td>
<td>2.760</td>
<td>3350</td>
</tr>
<tr>
<td>8</td>
<td>Aquaculture</td>
<td>2.570</td>
<td>8551</td>
</tr>
<tr>
<td>9</td>
<td>Reviews in Fisheries Science &amp; Aquaculture</td>
<td>2.545</td>
<td>321</td>
</tr>
<tr>
<td>10</td>
<td>Canadian Journal of Fisheries and Aquatic Sciences</td>
<td>2.466</td>
<td>3446</td>
</tr>
<tr>
<td>11</td>
<td>Fisheries Research</td>
<td>2.185</td>
<td>3446</td>
</tr>
<tr>
<td>12</td>
<td>Journal of Fish Diseases</td>
<td>2.138</td>
<td>1408</td>
</tr>
<tr>
<td>13</td>
<td>Ecology of Freshwater Fish</td>
<td>2.054</td>
<td>938</td>
</tr>
<tr>
<td>14</td>
<td>Marine Resource Economics</td>
<td>1.911</td>
<td>378</td>
</tr>
<tr>
<td>15</td>
<td>Marine and Freshwater Research</td>
<td>1.757</td>
<td>2202</td>
</tr>
<tr>
<td>16</td>
<td>Aquaculture Nutrition</td>
<td>1.665</td>
<td>1403</td>
</tr>
<tr>
<td>17</td>
<td>Fish Physiology and Biochemistry</td>
<td>1.647</td>
<td>1848</td>
</tr>
<tr>
<td>18</td>
<td>Fisheries Oceanography</td>
<td>1.578</td>
<td>714</td>
</tr>
<tr>
<td>19</td>
<td>Aquacultural Engineering</td>
<td>1.559</td>
<td>786</td>
</tr>
<tr>
<td>20</td>
<td>Diseases of Aquatic Organisms</td>
<td>1.549</td>
<td>2528</td>
</tr>
<tr>
<td>21</td>
<td>Journal of Fish Biology</td>
<td>1.519</td>
<td>5419</td>
</tr>
<tr>
<td>22</td>
<td>Transactions of the American Fisheries Society</td>
<td>1.502</td>
<td>2266</td>
</tr>
<tr>
<td>23</td>
<td>Aquaculture Research</td>
<td>1.461</td>
<td>3873</td>
</tr>
<tr>
<td>24</td>
<td>CCAMLR Science</td>
<td>1.429</td>
<td>156</td>
</tr>
<tr>
<td>25</td>
<td>Fisheries Management and Ecology</td>
<td>1.327</td>
<td>837</td>
</tr>
<tr>
<td>26</td>
<td>Knowledge and Management of Aquatic Ecosystems</td>
<td>1.217</td>
<td>342</td>
</tr>
<tr>
<td>27</td>
<td>North American Journal of Fisheries Management</td>
<td>1.201</td>
<td>2359</td>
</tr>
<tr>
<td>28</td>
<td>Marine and Coastal Fisheries</td>
<td>1.177</td>
<td>291</td>
</tr>
<tr>
<td>29</td>
<td>Aquaculture International</td>
<td>1.095</td>
<td>1383</td>
</tr>
<tr>
<td>30</td>
<td>Journal of the World Aquaculture Society</td>
<td>1.015</td>
<td>1254</td>
</tr>
<tr>
<td>31</td>
<td>New Zealand Journal of Marine and Freshwater Research</td>
<td>0.938</td>
<td>1003</td>
</tr>
<tr>
<td>32</td>
<td>Journal of Aquatic Animal Health</td>
<td>0.906</td>
<td>615</td>
</tr>
<tr>
<td>33</td>
<td>Fishery Bulletin</td>
<td>0.879</td>
<td>785</td>
</tr>
<tr>
<td>34</td>
<td>Journal of Applied Ichthyology</td>
<td>0.845</td>
<td>2785</td>
</tr>
<tr>
<td>35</td>
<td>Fisheries Science</td>
<td>0.839</td>
<td>2861</td>
</tr>
</tbody>
</table>
### Table 7.4: Continued.

<table>
<thead>
<tr>
<th>RANK</th>
<th>JOURNAL NAME</th>
<th>IF</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>Journal of Shellfish Research</td>
<td>0.721</td>
<td>1,946</td>
</tr>
<tr>
<td>37</td>
<td>North American Journal of Aquaculture</td>
<td>0.715</td>
<td>1,035</td>
</tr>
<tr>
<td>38</td>
<td>Fish Pathology</td>
<td>0.673</td>
<td>415</td>
</tr>
<tr>
<td>39</td>
<td>Acta Ichthyologica et Piscatoria</td>
<td>0.670</td>
<td>538</td>
</tr>
<tr>
<td>40</td>
<td>Latin American Journal of Aquatic Research</td>
<td>0.594</td>
<td>583</td>
</tr>
<tr>
<td>41</td>
<td>California Cooperative Oceanic Fisheries Investigations Reports</td>
<td>0.586</td>
<td>177</td>
</tr>
<tr>
<td>42</td>
<td>Turkish Journal of Fisheries and Aquatic Sciences</td>
<td>0.484</td>
<td>825</td>
</tr>
<tr>
<td>43</td>
<td>Aquatic Living Resources</td>
<td>0.448</td>
<td>710</td>
</tr>
<tr>
<td>44</td>
<td>Bulletin of the European Association of Fish Pathologists</td>
<td>0.431</td>
<td>630</td>
</tr>
<tr>
<td>45</td>
<td>Israeli Journal of Aquaculture-Bamidgeh</td>
<td>0.348</td>
<td>631</td>
</tr>
<tr>
<td>46</td>
<td>Boletim do Instituto de Pesca</td>
<td>0.295</td>
<td>232</td>
</tr>
<tr>
<td>47</td>
<td>Iranian Journal of Fisheries Sciences</td>
<td>0.285</td>
<td>516</td>
</tr>
<tr>
<td>48</td>
<td>Indian Journal of Fisheries</td>
<td>0.235</td>
<td>481</td>
</tr>
<tr>
<td>49</td>
<td>California Fish and Game</td>
<td>0.219</td>
<td>231</td>
</tr>
<tr>
<td>50</td>
<td>Nippon Suisan Gakkaishi</td>
<td>0.090</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>73,240</strong></td>
</tr>
</tbody>
</table>
Table 7.5: Explanation and description of the used social network analysis or graph theory metrics.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Description</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>The actual number of connections divided by the total number of possible connections.</td>
<td>Shows the level of connectedness of the network.</td>
</tr>
<tr>
<td>Degree</td>
<td>The number of connections attached to each node.</td>
<td>Shows the average number of connections possessed by each scientist.</td>
</tr>
<tr>
<td>Weighted degree</td>
<td>The number of connections attached to each node, taking into account the weight of the connection</td>
<td>Shows the average weighted number of connections possessed by each scientist.</td>
</tr>
<tr>
<td>Max cliques</td>
<td>The maximal complete subgraph of a given graph. In other words, the largest group of nodes where all the nodes are connected to one another.</td>
<td>Allows to identify intense collaborative sub-networks, where everyone within the sub-network has co-authored with everyone else (either through a single or multiple publications)</td>
</tr>
<tr>
<td>Average Clustering</td>
<td>The extent to which a scientist's co-authors also collaborate with each other</td>
<td>Allows to make inferences with respect to the likely exchange of new ideas across the network.</td>
</tr>
<tr>
<td>Degree Centrality</td>
<td>Measures the number of links to other nodes a particular node has.</td>
<td>Allows for the identification of central nodes within the network, in terms of the number of connections they have.</td>
</tr>
<tr>
<td>Closeness Centrality</td>
<td>Measures the distance of a node to all other nodes in the network</td>
<td>Allows for the identification of nodes that are most likely to receive information quickly in the network.</td>
</tr>
<tr>
<td>Eigenvector Centrality</td>
<td>A centrality measure that has been adjusted on the assumption that the centrality of a node cannot be assessed in isolation from the centrality of all the other nodes to which it is connected.</td>
<td>Allows for the identification of nodes that are well-connected to others who are also well connected.</td>
</tr>
<tr>
<td>Betweenness Centrality</td>
<td>Measures the extent to which a particular node lies between the other nodes in the network</td>
<td>Allows for the identification of nodes that may otherwise look unimportant, but that play important intermediary roles in the network in terms of information flow (e.g. brokers).</td>
</tr>
</tbody>
</table>
Table 7.6: Publications with the highest topic proportion from a single topic; indicating that the publication mostly covers aspects of that particular topic. For each of the 16 topics, the top-2 publications are shown.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Year</th>
<th>Title</th>
<th>Journal</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2017</td>
<td>Mortality of pond-cultured Litopenaeus vannamei associated with Acute Hepatopancreatic Necrosis Disease (AHPND) and White Spot Syndrome Virus (WSSV) infection in the Philippines</td>
<td>Fish Pathology</td>
<td>0.84</td>
</tr>
<tr>
<td>1</td>
<td>2017</td>
<td>Detection of viruses in the exotic shrimp Penaeus vannamei Boone, 1931 cultured in India</td>
<td>Indian Journal of Fisheries</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>2002</td>
<td>Duration of synchronous egg cleavage cycles at different temperatures in Siberian sturgeon (Acipenser baerii)</td>
<td>Journal of Applied Ichthyology</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>2002</td>
<td>Semen from rainbow trout produced using cryopreserved spermatozoa is more suitable for cryopreservation</td>
<td>Journal of Fish Biology</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>2015</td>
<td>Effects of reef attributes on fish assemblage similarity between artificial and natural reefs</td>
<td>ICES Journal of Marine Science</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>2010</td>
<td>Local v. microhabitat influences on the fish fauna of tidal pools in north-east Brazil</td>
<td>Journal of Fish Biology</td>
<td>0.87</td>
</tr>
<tr>
<td>4</td>
<td>2003</td>
<td>Movements of Atlantic salmon migrating upstream through a fish-pass complex in Scotland</td>
<td>Ecology of Freshwater Fish</td>
<td>0.89</td>
</tr>
<tr>
<td>4</td>
<td>2015</td>
<td>Farmed Atlantic salmon Salmo salar L. parr may reduce early survival of wild fish</td>
<td>Journal of Fish Biology</td>
<td>0.88</td>
</tr>
<tr>
<td>5</td>
<td>2012</td>
<td>The incomplete history of mitochondrial lineages between two rockfishes, Sebastes longispinis and Sebastes hubbsi (Scorpaeniformes: Scorpaenidae)</td>
<td>Journal of Fish Biology</td>
<td>0.86</td>
</tr>
<tr>
<td>5</td>
<td>2015</td>
<td>Analysis of sailfish (Istiophorus platypterus) population structure in the North Pacific Ocean</td>
<td>Fisheries Research</td>
<td>0.85</td>
</tr>
<tr>
<td>6</td>
<td>2005</td>
<td>A perspective on bottom water temperature anomalies in Long Island Sound during the 1999 lobster mortality event</td>
<td>Journal of Shellfish Research</td>
<td>0.93</td>
</tr>
<tr>
<td>6</td>
<td>2012</td>
<td>Atlantic water temperature and climate in the Barents Sea, 2000-2009</td>
<td>ICES Journal of Marine Science</td>
<td>0.81</td>
</tr>
<tr>
<td>7</td>
<td>2011</td>
<td>Catch-at-age assessment in the face of time-varying selectivity</td>
<td>ICES Journal of Marine Science</td>
<td>0.97</td>
</tr>
<tr>
<td>7</td>
<td>2015</td>
<td>Autocorrelated error in stock assessment estimates: Implications for management strategy evaluation</td>
<td>Fisheries Research</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>2013</td>
<td>Age determination and morphological study using otoliths in Cyprinus carpio Linnaeus, 1758 in the Southern Caspian Sea</td>
<td>Iranian Journal of Fisheries Sciences</td>
<td>0.83</td>
</tr>
<tr>
<td>8</td>
<td>2006</td>
<td>Age and growth of the hollowsnout grenadier, Caelorinchus caelorhinchus (Risso, 1810), in the Aegean Sea</td>
<td>Journal of Applied Ichthyology</td>
<td>0.82</td>
</tr>
<tr>
<td>Topic</td>
<td>Year</td>
<td>Title</td>
<td>Journal</td>
<td>%</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>9</td>
<td>2016</td>
<td>Optimal dietary carbohydrate to lipid ratio for bullfrog Rana (Lithobates) catesbeiana</td>
<td>Aquaculture Research</td>
<td>0.92</td>
</tr>
<tr>
<td>9</td>
<td>2015</td>
<td>Optimum Level of Dietary n-3 Highly Unsaturated Fatty Acids for Juvenile Sea Cucumber, Apostichopus japonicus</td>
<td>Journal of the World Aquaculture Society</td>
<td>0.90</td>
</tr>
<tr>
<td>10</td>
<td>2011</td>
<td>Combined effects of cycled starvation and feeding frequency on growth and oxygen consumption of gilthead sea bream, Sparus aurata</td>
<td>Journal of the World Aquaculture Society</td>
<td>0.88</td>
</tr>
<tr>
<td>10</td>
<td>2014</td>
<td>Changes in digestive enzyme activities of red porgy Pagrus pagrus during a fasting-refeeding experiment</td>
<td>Fish Physiology and Biochemistry</td>
<td>0.88</td>
</tr>
<tr>
<td>11</td>
<td>2017</td>
<td>Citral and linalool chemotypes of Lippia alba essential oil as anesthetics for fish: a detailed physiological analysis of side effects during anesthetic recovery in silver catfish (Rhamdia quelen)</td>
<td>Fish Physiology and Biochemistry</td>
<td>0.83</td>
</tr>
<tr>
<td>11</td>
<td>2012</td>
<td>Transportation of silver catfish, Rhamdia quelen, in water with eugenol and the essential oil of Lippia alba</td>
<td>Fish Physiology and Biochemistry</td>
<td>0.80</td>
</tr>
<tr>
<td>12</td>
<td>2013</td>
<td>Molecular characterization, expression analysis, and biological effects of interleukin-8 in grass carp Ctenopharyngodon idellus</td>
<td>Fish and Shellfish Immunology</td>
<td>0.92</td>
</tr>
<tr>
<td>12</td>
<td>2011</td>
<td>Myeloid differentiation factor 88 gene is involved in antiviral immunity in grass carp Ctenopharyngodon idella</td>
<td>Journal of Fish Biology</td>
<td>0.89</td>
</tr>
<tr>
<td>13</td>
<td>2003</td>
<td>An effective minimum concentration of un-ionized ammonia nitrogen for controlling Prymnesium parvum</td>
<td>North American Journal of Aquaculture</td>
<td>0.77</td>
</tr>
<tr>
<td>13</td>
<td>2011</td>
<td>Spatial distribution of diatom and pigment sedimentary records in surface sediments in Doubtful Sound, Fiordland, New Zealand</td>
<td>New Zealand Journal of Marine and Freshwater Research</td>
<td>0.75</td>
</tr>
<tr>
<td>14</td>
<td>2003</td>
<td>Determination of substrate preferences of tench, Tinca tinca (L.), under controlled experimental conditions</td>
<td>Journal of Applied Ichthyology</td>
<td>0.65</td>
</tr>
<tr>
<td>14</td>
<td>2004</td>
<td>Prey preference of Carcinus aestuarii: Possible implications with the control of an invasive mytilid and Manila clam culture in a northern Adriatic lagoon</td>
<td>Aquaculture</td>
<td>0.64</td>
</tr>
<tr>
<td>15</td>
<td>2012</td>
<td>Species identification in seamount fish aggregations using moored underwater video</td>
<td>ICES Journal of Marine Science Reviews in Fish Biology and Fisheries</td>
<td>0.92</td>
</tr>
<tr>
<td>15</td>
<td>2015</td>
<td>Sharks caught by the Brazilian tuna longline fleet: an overview</td>
<td>Fish and Fisheries</td>
<td>0.84</td>
</tr>
<tr>
<td>16</td>
<td>2012</td>
<td>Rights-based fisheries governance: From fishing rights to human rights</td>
<td>Fish and Fisheries</td>
<td>0.98</td>
</tr>
<tr>
<td>16</td>
<td>2005</td>
<td>Autopsy your dead...and living: A proposal for fisheries science, fisheries management and fisheries</td>
<td>Fish and Fisheries</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Conclusions

Applied data scientists and other scholars employ probabilistic topic models, such as Latent Dirichlet Allocation (LDA), to explore large corpora that would be impossible to examine manually. The popularity of these unsupervised machine learning techniques is further driven by the available open source libraries—Mallet (McCallum, 2002), Gensim (Rehurek and Sojka, 2010) and Stanford TMT (Ramage and Rosen, 2009)—and the relatively easy-to-obtain data. In this respect, however, there are aspects of LDA that can affect the quality of the topic model output—the latent topics. For example, the number of topics parameter, the choice for an appropriate Dirichlet prior, the various pre-processing steps, and the choice of data are all aspects that can affect the quality of the derived latent topics. On the other end, the raw latent topics provide little information to some of the domain-related questions of the corpus under study. Motivated by (i) trying to understand how to optimize aspects of LDA to obtain high-quality latent topics, and (ii) trying to use these topics to create valuable and useful knowledge for the domain of fisheries science, this thesis posed the following main research question:

MRQ — How can we improve the knowledge discovery process from textual data through latent topical perspectives?

To formulate the answer, we took a systematic and iterative knowledge discovery process called Knowledge Discovery in Databases (KDD) as a blueprint for effectively discovering knowledge from textual data. Successfully and optimally conducting each step within KDD (see 1.1), contributed to improving the final output: knowledge. Thus, Chapters 2–7 within this thesis were aimed at optimally performing KDD steps, which, taken together, are aimed at improving the overall knowledge discovery process for latent topics from fisheries scientific publications. In most chapters, we employed the highly studied and popular topic model Latent Dirichlet Allocation (see 1.2.1).

Chapter 2 studied two manifestations of scientific data, the full-text publication and abstract data. The data selection phase is the first step within KDD, and selecting
the data that leads to high-quality topics is an essential one. Additionally, Chapter 2 provided insights into the pre-processing steps (the second KDD step) and how they affected the latent topics. Since LDA is a Bayesian probabilistic topic model, prior knowledge can be encoded into the model. Chapter 3 explored all combinations of prior Dirichlet distributions and their effects on the quality of latent topics.

To assess the quality of latent topics, we used coherence measures, which, to date, are the most promising and near-human accurate measures to use. Additionally, we explored other evaluation approaches in Chapter 4. Combined, Chapters 2, 3 and 4 provide methodological optimizations for LDA with the objective to uncover high-quality topics.

Chapters 5–7 were aimed at interpreting the latent topics with the aim to discover knowledge; the final output of the KDD process. In Chapters 5 and 6, we explored how to construct knowledge from raw latent topics and sub-topics to shed light on the four pillars (ecologic, social, economic, and institution) of fisheries sustainability. In Chapter 7, we additionally combine latent topics with spatial and temporal data from fisheries science collaborations.

Chapters 2–7 are all aimed at improving the knowledge discovery process, and a mapping of each chapter to the relevant KDD process can be found in Fig. 1.6. Moreover, for Chapters 2–7 we posed six distinct research questions, and each contributed to providing answers to the main research question. In the following sections, we reiterate the six research questions and draw six conclusions. Collectively, they provide an answer to the main research question.

**RQ1 — What types of textual data result in high-quality latent topics?**

When uncovering latent topics from scientific publications, typically one can commence by choosing either abstract or full-text data. The two variants are generally treated separately from a data retrieval perspective, with abstract data arguably being the easiest to obtain. Although researchers are increasingly using computer-aided content analysis techniques, such as LDA, no study has examined whether abstract or full-text data produces higher quality latent topics of the underlying content. In Chapter 2, we quantitatively (i.e., through topic coherence) and qualitatively (i.e., through human ranking) assessed the quality of latent topics derived from abstract and full-text data from fisheries science publications. We constructed two datasets containing 4,400 and 15,000 articles from various fisheries peer-reviewed journals. Each dataset included both the abstract and full-text variants of the same articles (see Table 1.1).

The first dataset of 4,400 articles showed a significant difference in the quality of the derived latent topics, with the full-text topics exhibiting greater coherence scores and higher human topic rankings. Though the data was pre-processed identically, the lower coherence scores obtained from abstract data were mainly caused by so-called noise terms being present within the topic's top words. Such terms are not related to the biological, ecological, or socio-ecological meanings of those topics, but can be seen as
semantically incorrect terms: using, used, use, within, total, two, among, and within. Such noise terms require proper attention when dealing with abstract data, for example, through an enhanced cleaning (pre-processing) phase, part-of-speech filtering, or a domain-specific stop word list.

The latent topics from the dataset of 15,000 articles revealed very similar topic coherence and human topic ranking results for both abstract and full-text data. The latent topics from full-text articles were not significantly better in most instances, albeit that they obtained somewhat higher coherence scores. The presence of noise terms appeared to be mitigated by the higher number of publications and larger vocabulary sizes.

Through a domain expert’s evaluation, full-text topics obtained from the smaller dataset (4,400 articles) displayed a more fine-grained granularity of the latent meanings of those topics—a finding not present within the larger dataset (15,000 articles). Although we identified similar and detailed topics in both datasets, there remains an inconsistency between some uncovered topics, with specific topics being present in abstract data, and absent in full-text data (and vice versa). Although we conducted random initializations and aimed at uncovering the high-quality latent topics, these discrepancies nevertheless exist, and care should be given regarding the true underlying topical structure.

![Graphs showing coherence scores for abstract and full-text data](image)

**Figure 8.1:** A: Full-text data produces higher-quality latent topics. B: full-text and abstract data produce topics of comparable quality.

**Conclusion I**

For relatively small datasets, **full-text data produces higher-quality latent topics**, whereas, for larger datasets, full-text and abstract data produce topics of comparable quality. Smaller datasets should place more emphasis on the pre-processing
steps; specifically, care should be given domain-specific stop words that can potentially reduce the quality of latent topics. Additionally, for smaller datasets, **full-text data results in more detailed topics**, where typically topics have a higher granularity.

**RQ2 — How does the hyper-parameterization of a topic model algorithm affect the quality of latent topics?**

For LDA, a Bayesian probabilistic topic model, two important hyper-parameters exist that encode prior knowledge into the model. These are the Dirichlet distributions for the probabilities of words within topics, and the probabilities of topics within documents (see Fig. 1.3). In Chapter 3, we studied how symmetrical and asymmetrical (learned from data) parameterizations of the Dirichlet distributions affect the quality of the latent topics. We created 2,000 different LDA models to examine six different combinations of priors on two datasets containing 4,400 abstracts and 8,000 full-text data. To evaluate the quality of topics, we utilized topic coherence scores (as a proxy for topic quality) and human topic ranking.

When looking at the prior Dirichlet distribution of topic probabilities within documents, a natural assumption to make is that specific topics occur more frequently (e.g., popular or more researched topics), and other topics less frequently (e.g., niche research domains). This assumption is encoded by an asymmetrical Dirichlet prior. Conversely, a symmetrical prior (commonly the default choice) would contradict this assumption, as it encodes an equal probability for all topics to be present within documents. Our empirical analysis indeed confirms that an asymmetrical prior of topic probabilities within documents results in topics with greater and statistically significant coherence scores. However, this effect is most significant when using abstract data, and this effect appears to hold (to a far lesser extent) for full-text data only for LDA models with a high number of topics (K parameter). Additionally, human topic ranking showed more high-quality topics when using an asymmetrical prior.

Concerning the prior Dirichlet distribution of word probabilities within topics, we preferably want topics to be different (i.e., distinct) from each other to avoid conflicts between them. A symmetrical prior distribution allows topics to be as different as need be, as it is not (a priori) influenced by the word use statistics of all the documents. Also, a symmetrical prior distribution considers the power-law usage of words (i.e., some words occur in many of the documents). Thus, a symmetrical prior is naturally assumed to be the preferred choice. However, our results show no real benefits when applying a symmetrical or asymmetrical prior distribution to both datasets. Yet, in very few cases, a symmetrical distribution shows slight, but still very marginal, overall improved coherence and human topic ranking results for both datasets.

Additionally, some concerns raised within the conclusion of RQ1 related to incorrect terms and discrepancies between latent topics (latent topics present in one dataset and absent in another, or vice versa) seem to hold and are reconfirmed within this study.
Figure 8.2: Varying Dirichlet priors result in different quality topics on abstract data (A), and very similar quality topics on full-text data (B). See Chapter 3 for full details.

Conclusion II

For abstract datasets, an asymmetrical Dirichlet distribution of topic probabilities within documents can significantly increase the quality of latent topics. For full-text datasets, a similar, although far less significant observation holds for LDA models with $\geq 30$ topics. A symmetrical or asymmetrical Dirichlet distribution of word probabilities within topics have a negligible effect on the quality of latent topics.

RQ3 — Can we assess the quality of latent topics using a semi-automatically constructed list of semantically related words?

The latent topics uncovered with Latent Dirichlet Allocation are expressed by probability distributions over the (fixed) vocabulary. Typically, when sorted, the words with high probability (e.g., top 10) reveal the semantic meaning of the latent topic. Such high probability words can be viewed as semantically related words, as they more frequently co-occur within the same linguistic context. Quantitative (coherence scores) and qualitative (human topic ranking) measures are used to evaluate the quality of the co-occurring words, and thus the quality of the uncovered latent topics. In Chapter 4, we studied an alternative approach to assess the quality of latent topics.

Semantically related words can additionally be formalized by a semantic lexicon, where the words are related by a hyponym-hypernym association, including transitive relationships. For example, the words “salmon”, “trout”, and “chars” are all hyponym words that share the hypernym salmonidae. Likewise, the words “crab”, “lobster”, and
“crayfish” are related by the hypernym crustacean. We studied the construction of such semantic lexicons (from web content) by employing a semi-automatic process, called “bootstrapping”, and finding semantically related words with extraction patterns (noun phrases that share the same verb). The bootstrapping approach used a number of seed words to steer the lexicon into the correct linguistic context. Before adding new words to the lexicon, two different scoring functions were used based on word frequency, and collocation statistics—words that occur near each other. A (fisheries) domain expert evaluated the quality of the bootstrapped lexicon.

Extraction patterns that contain strong domain-specific verbs—such as the verbs “fishing” and “regulate” for the domain of fisheries—generally achieved the highest accuracy lexicon. Thus, strong domain-related verbs, which more frequently occur within a particular domain and less frequently in a non-related domain, are better linguistic cues for correct lexicon words. The use of collocation statistics, in contrast to word frequencies, provided higher accuracy lexicons only in a small number of cases. When bootstrapping large lexicons (i.e., 100 words), the differences in scoring functions diminished, and the accuracy converged to similar values.

We experimented with different algorithmic approaches that can, potentially, be utilized to evaluate the quality of latent topics. We have demonstrated several algorithmic optimizations to create lexicons of higher quality but have not performed the actual quality evaluation of latent topics. For future research, the use of semantic similarities—a measure of conceptual distance between two words or sentences—and lexical databases are ideal candidates for such endeavors (Baddeley, 1966; Li et al., 2006; Resnik, 1999). By exploiting extraction patterns, we have limited lexicon entries of semantically related words to nouns only, which may result in limitations when evaluating latent topics.

![Figure 8.3](image_url)

(a) frequency scoring  
(b) collocation scoring

**Figure 8.3**: New lexicon entries from strong fisheries-domain related verbs (see blue line for instance) resulted in higher accuracy lexicons. See Chapter 4 for full details.
CHAPTER 8. CONCLUSIONS

**Conclusion III**

Bootstrapping semantic lexicons with extraction patterns and strong domain related verbs results in higher accuracy lexicons. Scoring new lexicon entries with collocation statistics shows slightly improved accuracy for smaller lexicons. While (high-quality) semantic lexicons of hyponym-hypernym associations provide a way to assess the quality of latent topics, similarity measures and lexical databases are two approaches to evaluate this.

**RQ4 — How can we construct knowledge from latent topics derived from large collections of documents?**

Chapters 2 and 3 studied methodological optimizations of Latent Dirichlet Allocation that enable the creation of high-quality topics. In Chapter 5, we utilized lessons learned and employed a large-scale topic model analysis for the domain of fisheries science. We constructed a unique dataset of over 46,000 scientific publications from 21 top-tier fisheries journals published between 1990–2016. For fisheries sustainability, it is essential that the ecological, social, economic, and institutional elements are appropriately considered, in other words, in an equal and balanced manner. Within these four pillars of sustainability, it is argued that the ecological considerations have attracted the most attention, leaving the social, economic, and institutional (i.e., human dimensions) relatively neglected. Chapter 5 aimed to investigate whether latent topics can be utilized to assess and quantify the presence of the ecological and human dimension within fisheries scientific publications. Specifically, we investigated whether fisheries science research is diversified enough to capture the full complexity of the system, or if it is focused on a few selected components of this system.

Our analysis revealed a topical deconstruction of 25 broad topics to describe the entire fisheries science corpus. From the 25 uncovered topics, 24 relate to biological considerations, and a single topic (i.e., fisheries management) covered aspects of the human dimension. It is evident that the research focus in fisheries during the last 26 years has not entirely captured the complexity of the fisheries domain, especially of the human dimension component. To enable a dynamic portrait for the field of fisheries science, we analyzed topical trends for different time intervals. The topics related to fisheries management, conservation, and models show the most substantial proportional increase between 1990–2016. Negative trends were observed for the topics related to non-fish species, biochemistry, and primary production. Although fisheries management was identified as the only topic addressing human dimension considerations, it was simultaneously the topic that showed the highest proportional increase (+5.2%) and was the third most prevalent topic (6.13%). The increasing prevalence and interest in the topic of fisheries management might indicate the strengthening of the connection between fisheries science and management processes, in the light of the growing concern about the status of fish stocks worldwide.

When exploring latent topics, an extensive post-analysis phase can reveal a detailed
and macroscopic view of the corpus under study. It furthermore enables answering questions that through manual analysis would be extremely challenging. For fisheries sustainability, it is essential that the biological and human dimensions are addressed in an equal and balanced manner. Our large-scale topic analysis of fisheries science shows a substantial imbalance in the underlying topical content, focusing heavily on the biological dimension and possibly impeding fisheries sustainability.

Conclusion IV

The interpretation of latent topics in raw form reveal very little insight regarding the domain under study. However, with an extensive post-analysis phase, the latent topics can reveal new and useful knowledge that would have been impossible to obtain through manual analysis. This post-analysis phase can include: (i) the labeling of topic-word distributions, (ii) the aggregation and segregation of document-topic distributions (by time, journal and overarching theme), (iii) the visualization of latent topics, and (iv) using regression methods to obtain trends.

RQ5 — How can we construct knowledge from latent sub-topics derived from large collections of documents?

In Chapter 5, we uncovered main or general topics from fisheries science publications with the aim to quantify and shed light on the four pillars of sustainability. Main topics provide a broad summary of the corpus and can be viewed as a first-level decomposition of the corpus into latent topics. Additionally, in Chapter 6, we zoomed in on the main topics to explore their underlying sub-topics (i.e., second-level analysis) from a dataset comprising 22,000 fisheries science publications from 13 top-tier fisheries journals covering aspects of fishery models. Within fisheries science, modeling and simulation are among the most frequently used methods. Fishery models come in a multitude of shapes and varieties, each addressing specific aspects of the fishery system. For fisheries sustainability, and from the viewpoint of fishery models, it is also essential to shed light on the various aspects they address (ecological, social, economic, or institutional). Therefore, in Chapter 5, we aimed to construct new insights from the sub-topics found within the broader topics covering aspects of fishery models.

From the corpus of 22,000 publications, we uncovered two main modeling topics: (i) estimation models (a topic that contains the ideas of catch, effort, and abundance estimation) and (ii) stock assessment models (a topic on the assessment of the current state of fishery and future projections of fish stock responses and management effects). The topic of estimation models revealed 14 underlying sub-topics, and the topic stock assessment revealed 15 underlying sub-topics. The sub-topics address primarily ecological aspects of the fishery system (i.e., biological dimension), with only a few sub-topics addressing the human dimension—management effects and management tools. Both modeling topics and their underlying sub-topics, therefore, primarily focus on the biological aspects of fisheries, a finding that indicates that fishery models
might not adequately address the four pillars of sustainability.

A first-level topic analysis produces a set of main or general topics that can subsequently be used to filter the corpus for a zoomed-in or second-level topic analysis. In Chapter 6, we used the broad topics to identify documents explicitly addressing modeling aspects, to then uncover their underlying sub-topics. Our analysis revealed interpretable sub-topics that can help to explore the corpus in more detail. However, it is essential to understand the implications of such an approach. For example, filtering the corpus by main topics was accomplished by selecting documents where the main topic constituted the most significant part of the document. More concretely, the main topic proportion was greater than any of the remaining topics. Alternative approaches include adopting a threshold value, where documents are filtered if they constitute at least some percent of the main topic, for instance 50%. In any case, one needs to take topic co-occurrence into account, as other topics still make up for the remaining part of the filtered documents. Technically, with LDA, the word-to-topic assignment can also be used as an alternative approach to filter documents by main topics, as each word in a document gets assigned a topic (see also step 2(b) in the generative process in 1.2.1). However, during our analysis, doing so resulted in less interpretable topics and was therefore not used.

**Conclusion V**

A first-level topic analysis uncovers main topics, which can subsequently be used to perform a second-level topic analysis with the purpose of exploring the main topics’ underlying sub-topics. In doing so, we uncovered interpretable sub-topics that show a more detailed decomposition of the main topic. However, care must be given to the filtering approach, as filtered documents contain co-occurring topics that might not necessarily cover aspects desirable in the subsequent level of analysis.

**RQ6 — How can we utilize knowledge derived from latent topics for a subsequent knowledge discovery process?**

In Chapters 5 and 6, we have used latent topics and sub-topics uncovered from fisheries science publications to conduct a large-scale assessment of fisheries sustainability. Specifically, we have aimed to shed light on the four pillars of sustainability within general fisheries science topics, as well as sub-topics for the field of fishery models. In both cases, our analysis revealed a strong bias towards the ecological considerations, relatively neglecting the human dimension. In Chapter 7, through an iterative knowledge discovery in databases process (KDD), we utilized latent topics obtained from 73,240 fisheries publications to enrich a social network analysis of 106,137 authors from 100,175 affiliations.

Our social network analysis of co-authorships provided insights into the structure of
fisheries science collaborations, where degree, distance, and centrality measures are used to quantify elements of the network. Additionally, to extend our analysis, we have combined properties of each author, such as the publication data, affiliation data (e.g., country or institution), spatial data (such as longitude and latitude information), and hidden community data (through community detection algorithms). In doing so, we have created a social network enriched with insights into temporal, spatial and by using the latent topics, a topical decomposition of the publications. We have shown how latent topics, combined with techniques from social network analysis, can not only provide an analysis of the latent scientific output (as provided in Chapters 5 and 6) but can also provide answers to an array of questions regarding the collaboration patterns within fisheries.

We find that while the fisheries science network is becoming more extensive, it is simultaneously becoming more intensive, with a clear division of focus between the traditional powerhouses (e.g., US, Canada, Japan, Australia, UK, Norway) and new entrants (e.g., China, Brazil, India). The uncovered network exhibits clusters and links which, though likely shaped by an array of overlapping factors, reveal a number of political-economic patterns, which merit reflection by both fisheries scientists and policymakers alike.

Our analysis has presented a shifting field that has become increasingly collaborative, though less cohesive, with a number of key players maintaining hegemonic positions within the network. The most productive and collaborative countries are those which have large industrialized fisheries-related interests. Although the collaboration network has become more extensive, it has also become more intensive in places, with a clear spatial pattern evident in the structure of scientific collaborations across the field. In this respect, the fisheries science landscape is one whereby the centers of knowledge production and the connections between them display trends more akin to regionalization than globalization.

The authorship network suggest that authors across the field may be engaging in a repeat, rather than broad style of collaboration, which may work as a reinforcing mechanism with respect to the knowledge that is produced by the field. This pattern is likely to limit the potential gains of collaboration, and could have consequences in terms of pushing the boundaries of fisheries science in new and fruitful ways, in a manner which may help address some of the ongoing challenges within the field.

**Conclusion VI**

By employing probabilistic topic models to derive knowledge from fisheries science publications, and by utilizing the latent topics in combination with a social network analysis of fisheries science collaboration, we have shown that such dual analysis reveals, besides insights into the corpus under study, new and un-
explored insights for the field of fisheries science and the production of knowledge. Such an investigation was achievable since (i) the two types of analysis were highly-related—since authors produce publications—and (ii) meta-data on authors and their affiliation was available through external data sources.

8.1 Scientific Contributions

In the previous sections, we re-introduced our main research question and the six formulated research questions that, collectively, provide answers to improving the knowledge discovery process of latent topics from fisheries scientific articles. This section presents a summary of the scientific and some societal contributions.

For researchers wanting to employ latent Dirichlet allocation to uncover latent topics from scientific publications, the following observations can help to accomplish this goal more effectively:

- The more extended and elaborate format of full-text articles has been shown to produce topics of higher granularity. Additionally, for relatively small datasets—around 5,000 in our case—full-text articles produce more high-quality topics. For larger datasets—around 15,000—abstract and full-text data produce topics of very similar quality.

- For abstract data, the use of an asymmetrical Dirichlet prior for topics in documents outperforms symmetrical Dirichlet priors concerning topic quality. For full-text data, asymmetrical priors can sometimes increase topic-quality, although this increase is not significant most of the time. Asymmetrical or symmetrical Dirichlet priors for words in topics have a negligible effect on the quality of topics.

- The use of topic coherence measures as a proxy for topic quality has been revealed to generate interpretable topics, and higher topic coherence scores are associated with higher human topic ranking scores.

- For relatively small datasets, more care should be given to domain-specific stop words, as most standard stop word lists do not adequately filter out non-relevant words, which negatively impacts the quality of the topics.

- When the interpretability of topics is considered highly crucial, some form of normalization is still desired. As per our analysis, we advise applying lemmatization techniques in the pre-processing phase. With lemmatization, and in contrast to stemming techniques, verbs and nouns can still be distinguished and, generally speaking, it has an advantageous effect on the readability of the topic-word distributions.
– Local minima (i.e., model convergence to a sub-optimal solution) persist to be a significant drawback of probabilistic topic models, including LDA. As per our analysis, running multiple random initializations still produced some discrepancies between the uncovered topics. Possible solutions to circumvent this drawback are given in the next section.

– A first-level topic analysis produced topics that can subsequently be used to filter for documents containing a particular topic, whereby a second-level topic analysis produced their underlying sub-topics. As per our analysis, this produced interpretable sub-topics for fishery models. However, a second-level topic analysis comes with some degree of topic co-occurrence, as other topics still make up some part of the filtered documents.

For the domain of fisheries science, we highlight the following contributions:

– The topics uncovered from 46,000 fisheries publications revealed highly imbalanced sustainability considerations, with a significant focus on the biological dimension, largely neglecting the human dimension. Out of the 25 uncovered topics, the human dimension was only expressed through a single topic: fisheries management. However, fisheries management showed the most substantial proportional increase between 1990–2016 and now constitutes the third most prevalent topic.

– The sub-topics uncovered from 22,236 fisheries modeling publications show a strong focus on modeling the biological aspects of the fishery system, with only a few sub-topics covering aspects of the human dimension.

– The fisheries science collaboration network of 106,137 authors has become increasingly collaborative, though less cohesive, with a number of key players maintaining hegemonic positions within the network. By and large, the most productive (and collaborative) countries in terms of fisheries science are those which have large industrialized fisheries-related interests, many of them global in nature.

– The fisheries science collaboration landscape is one whereby the centers of knowledge production and the connections between them display trends more akin to regionalization than globalization. The collaboration network suggest that authors across the field may be engaging in a repeat, rather than broad style of collaboration.

### 8.1.1 LDA Workflow

Topic models are a popular unsupervised machine learning technique to understand and explore large collections of documents. This section of the thesis describes the
steps involved (see Fig. 8.4) to go from raw data to an effective topic model analysis. The steps are embedded into the various phases of the KDD process (Fig. 1.1) that we have used throughout this thesis. We furthermore provide the full workflow in Python code for performing LDA on full-text articles which can be found at: github.com/shaheen-syed/LDA.

**Figure 8.4:** Steps to perform a topic model analysis on scientific articles. Code can be found at github.com/shaheen-syed/LDA

**EXTRACTION:** The extraction phase involves the process of obtaining a set of documents from a repository, such as Scopus or the Web of Science, or it can involve the steps of scraping a publisher’s website to retrieve full-text articles (typically in PDF.
format). Scopus generally provides publication abstracts, including all the meta-data (journal, authors, affiliations, publication date) through various APIs. The upside of using an API is that publication content is easily obtained for a large number of documents simultaneously, however, these APIs often do not provide full-text for all the publications or journals of interest. In these cases, scraping a publisher’s websites can be an alternative solution. This process involves building many handcrafted crawlers, as each publisher lists their publications in a different manner on their website. Download limits should always be respected when building such scripts. Another option would be to manually download articles, although such approaches might not be feasible if the document collection of interest contains thousands or tens of thousands of articles. To enable a comparison of topics by time, or a comparison of topics by journals, it is important to store this information alongside the content of the document.

Relevant code:

```
• extraction.extract_publications()
```

**SELECTION:** This step involves the selection or filtering for relevant publications. For example, the LDA analysis might only look at English language publications, only take into account publication that constitute a proper research articles (excluding errata, editorials, letters, comments etc.). Additionally, filtering for missing data is important when publication meta-data is missing.

**PRE-PROCESSING:** The pre-processing phase can be seen as the process of going from a document source to an interpretable representation for the topic model algorithm. This phase is typically different for full-text and abstract data. One of the main differences is that abstract data is often provided in a clean format, whereas full-text is commonly obtained by converting a PDF document into its plain text representation.

Within this phase, an important part is to filter out the content that is not important from a topic model's point-of-view, rather than from a human’s point-of-view. Abstract data usually comes in a clean format of around 300–400 words, and little additional text is added to it; typically the copyright statement is the only text that should be removed. In contrast, full-text articles can contain a lot of additional text that has been added by the publisher. This is article meta-data and boilerplate. It is important that such additional text is removed, and various methods to do so exist. Examples include: deleting the first cover page; deleting the first n-bits of the content; using regular expressions or other pattern matching techniques to find and remove additional text, or more advanced methods (e.g., (Kohlschütter et al., 2010)). For full-text articles, a choice can be made to also exclude the reference list or acknowledgment section of the publication.

Latent Dirichlet allocation, as well as other probabilistic topic models, are bag-of-words (BOW) models. Therefore, the words within the documents need to be tokenized; the process of obtaining individual words (also known as unigrams) from sentences.
For English text, splitting words on white spaces would be the easiest example. Besides obtaining unigrams, it is also important to find multi-word expressions (Manning and Schütze, 1999), such as two-word (bi-grams) or multi-word (n-grams) combinations. Named entity recognition (NER)—a technique from natural language processing (NLP)—can, for instance, be used to find multi-word expressions related to names, nationalities, companies, locations, and objects within the documents. The inclusion of bi-grams and entities allows for a richer bag-of-words representation than a standard unigram representation. Documents from languages with implicit word boundaries may require a more advanced type of tokenization (Goldwater et al., 2006).

Although all tokens within a document serve an important grammatical or syntactical function, for topic modeling they are not all equally important. Words need to be filtered for numbers, punctuation marks, and single-character words as they bear no topical meaning. Furthermore, stop words (e.g., the, is, a, which) are words that have no specific meaning from a topical point-of-view, and such words need to be removed as well. For English, and a number of other languages, there exist fixed lists of stop words that can easily be used (many NLP packages such as NLTK and Spacy include them). However, it is important to also create a domain-specific (also referred to as corpus-specific) list of stop words and filter for those words. Such domain-specific stop words can also become apparent in the evaluation phase. If this is the case, going back to the pre-processing phase and excluding them would be a good approach. Another approach to removing stop words is to use TF-IDF (Salton, 1968) and include or exclude words within a certain threshold. Contrary to our analysis, a study by Schofield et al. (2017) has indicated that removing stop words have no substantial effect on model likelihood, topic coherence, or classification accuracy.

For grammatical reasons, different word forms or derivationally related words can have a similar meaning and, ideally, for a topic model analysis, such terms need to be grouped (i.e., they need to be normalized). Stemming and lemmatization are two NLP techniques to reduce inflectional and derivational forms of words to a common base form. Stemming heuristically cuts off derivational affixes to achieve some normalization, albeit crude in most cases. Stemming loses the ability to relate stemmed words back to their original part-of-speech, such as verbs or nouns, and decreases the interpretability of topics in later stages (Evangelopoulos et al., 2012). Lemmatization is a more sophisticated normalization method that uses a vocabulary and morphological analysis to reduce words to their base form, called lemma. For increased topic interpretability, we recommend lemmatization over stemming. Additionally, uppercase and lowercase words can be grouped for further normalization. The process of normalization is particularly critical for languages with a richer morphology (Taghva et al., 2005). Failing to do can cause the vocabulary to be overly large, which can slow down posterior inference, and can lead to topics of poor quality (Boyd-Graber et al., 2014).

Relevant code:
ONDONING: The transformation phase includes the creation of a dictionary of words and preparing the data for the topic model software or package. The dictionary of words (also referred to as the vocabulary) is typically a list of unique words represented as integers. For example, ‘fish’ is 1, ‘population’ is 2 and so on. The length of the dictionary is thus the number of unique words within the corpus; normalization reduces the length of the dictionary, and speeds up the inference time. The next step is to represent the documents as bag-of-words features. Doing this for all the documents creates a matrix (i.e., table) with the rows being the individual documents, the columns being the words within the dictionary, and the cells being the frequency of that word within the document. This is one representation of bag-of-words features and other less sparse representations exist as well, see for example (Boyd-Graber et al., 2014). If not performed during the pre-processing phase, words that occur only once, and words that occur in roughly 90% of the documents can be eliminated as they serve no discriminative topical significance. Especially omitting high frequently occurring words prevents such words from dominating all topics. Removing high and low-frequency words (i.e., pruning) within a matrix representation is generally much faster to perform.

Several LDA tools are available and each of them requires a slightly different transformation step to make the data suitable for topic analysis. However, in essence, they require a conversion from words to bag-of-words representation, to some matrix representation of the full corpus. Several LDA packages exist that might be worth exploring: Gensim (Rehurek and Sojka, 2010), Mallet (McCallum, 2002), Stanford Topic Modeling Toolbox (Ramage and Rosen, 2009), Yahoo! LDA (Narayanamurthy, 2011), and Mr. LDA (Zhai et al., 2012).

Relevant code:

```
• transformation.transform_for_lda()
```

DATAMINING: The data mining phase involves fitting or training the LDA model. It also involves a careful analysis of the hyper-parameters and the creation of different LDA models. Similarly to the transformation phase, the use of a certain LDA module or software tool determines what parameters and hyper-parameters can be adjusted.

Since calculating a closed form solution of the LDA model is intractable (Blei et al., 2003; Blei, 2012), approximate posterior inference is used to create the distributions of words in topics, and topics in documents. To avoid local minima, both in the case of variational (Blei and Jordan, 2006; Teh et al., 2006; Wang et al., 2011) and sampling-based (Newman et al., 2007; Porteous et al., 2008) inference techniques, the initialization of the model is an important consideration in the data mining phase. Thus,
regardless the initialization, and regardless the inference method used, to guard for problems of local minima, multiple starting points should be used to improve the stability of the inferred latent variables (Boyd-Graber et al., 2014).

Running the inference is the most important step in the data mining phase. It results in the discovery of the latent variables (words in topics, and topics in documents). The convergence of the model (e.g., the likelihood) should be closely monitored. The time typically depends on initialization, the number of documents, model complexity, and inference technique. A straightforward approach to optimize for the various hyper-parameters would be to perform a grid-search and inferring LDA models for combinations of them. Such hyper-parameters include the number of epochs or passes over the corpus, the number of iterations for convergence, the number of topics, the types of Dirichlet priors, and the starting points.

Relevant code:

- `datamining.execute_lda()`

**EVALUATION:** The evaluation phase includes a careful analysis and inspection of the latent variables from the various created LDA models. Since LDA is an unsupervised machine learning technique, extra care should be given during this post-analysis phase; in contrast to, for example, supervised methods where typically a labeled gold-standard dataset exist.

Measures such as predictive likelihood on held-out data (Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan and Mimno, 2009) have been proposed to evaluate the quality of generated topics. However, such measures correlate negatively with human interpretability (Chang et al., 2009), making topics with high predictive likelihood less coherent from a human perspective. High-quality or coherent latent topics are of particular importance when they are used to browse document collections or understand the trends and development within a particular research field. As a result, researchers have proposed topic coherence measures, which are a quantitative approach to automatically uncover the coherence of topics (Aletras and Stevenson, 2013; Stevens et al., 2012; Newman et al., 2010a; Röder et al., 2015). Topics are considered to be coherent if all or most of the words (e.g., a topic's top-N words) are related. Topic coherence measures aim to find measures that correlate highly with human topic evaluation, such as topic ranking data obtained by, for example, word and topic intrusion tests (Chang et al., 2009). Human topic ranking data are often considered the gold standard and, consequently, a measure that correlates well is a good indicator for topic interpretability.

Exploring the topics by a human evaluator is considered the best approach. However, since this involves inspecting all the different models, this approach might not be feasible. Topic coherence measures can quantitatively calculate a proxy for topic quality, and per our analysis, topics with high coherence were considered interpretable by do-
main experts. Combing coherence measures with a manual inspection is thus a good approach to find the LDA model that result in meaningful and interpretable topics. In short, three questions should be answered satisfactory (Boyd-Graber et al., 2014):

1. Are topics meaningful, interpretable, coherent and useful?

2. Are topics within documents meaningful, appropriate and useful?

3. Do the topics facilitate a better understanding of the underlying corpus?

The evaluation phase can also result in topics that are very similar (i.e., identical topics), topics that should ideally be merged or split (i.e., chained or mixed topics), topics that are un-interpretable (i.e. nonsensical), or topics that contain unimportant, too specific, or too general words. In those cases, it would be wise to revisit the pre-processing phase and repeat the analysis.

Relevant code:

- evaluation.calculate_coherence()
- evaluation.plot_coherence()
- evaluation.output_lda_topics()

**INTERPRETATION:** The interpretation phase, although closely related to the evaluation phase, includes a more fine-grained understanding of the latent variables. The main goal of the interpretation phase is to go beyond the latent variables and understand the latent variables in the context of the domain under study. This phase is highly depending on the research question that we would want to have answered. What topics are present, how are they distributed over time, and how are they related to other topics are possible ways to explore the output of the LDA analysis. Similarly to the evaluation phase, aiming for a deeper understanding of the topics might also result in flaws in the analysis. For example, a visualization of the topics that places two very distinct topics in close proximity, high probability of a topic in a document that does not cover aspects of that topic, or topics that should not co-occur together are indicators of flaws or areas of improvements. In such cases, it would be wise to revisit the pre-processing phase and to re-run the analysis with, for instance, different model parameters or pre-processing steps.

Relevant code:

- interpretation.infer_document_topic_distribution()
- interpretation.get_document_title_per_topic()
CHAPTER 8. CONCLUSIONS

- interpretation.plot_topics_over_time()
- interpretation.plot_topics_over_time_stacked()
- interpretation.plot_topic_co_occurrence()

8.2 Limitations

The specific limitations are addressed in the individual chapters, though it is worth reiterating some of the most important ones concerning the Latent Dirichlet Allocation method and the chosen domain of fisheries science.

8.2.1 Latent Dirichlet Allocation

**Bag-of-words assumption:** LDA can be considered a bag-of-words (BOW) model where the order of words is neglected. Specifically, for the field of natural language generation, this assumption would not work. However, the assumption is realistic if the purpose is to uncover the semantic structure of a corpus (Blei, 2012). Also, treating the individual words as document features would lose the meaning of compound words (e.g., fisheries management, rainbow trout). We have mitigated this by including two-word expressions (bigrams), and multi-word expressions (n-grams) through named entity recognition methods. To generate more realistic-looking documents within the generative process, one can sample words that are conditioned on the previously sampled words, thus capturing short and long-term dependencies. Such approaches (e.g., (Griffiths et al., 2005; Wallach, 2006b)) relax the BOW assumption and can result in topics that are more meaningful.

**Document exchangeability:** LDA further assumes document exchangeability (the order in which documents are analyzed) is unimportant, yet all documents are analyzed at the end of the LDA analysis. This assumption is problematic when topics have changed considerably in the way they are described, especially when analyzing documents covering a very long time span. The dynamic topic model (Blei and Lafferty, 2006) is one approach to capture evolving topics by creating topics for different time slices, where each slice is conditioned on the previous time slice. We might surmise that specific fisheries topics were described differently in the early 1990s than they are today and, if so, the standard LDA model used in this thesis does not explicitly capture this phenomenon.

**Number of topics:** LDA assumes that the number of topics, that is the value of K, is known a priori. Similar assumptions are typically found within other unsupervised clustering methods, such as the K-nearest neighbor or Gaussian Mixture Models. We have used a grid-search approach whereby LDA models were created by varying the
K-parameter. The best model was obtained by calculating the model quality and including an inspection by domain experts. For LDA, this is a common and useful technique. However, Bayesian non-parametric models exist, such as the hierarchical topic model that automatically infers the number of topics from the data (Whye Teh et al., 2004). Such hierarchical topic models show effective and superior performance over the standard LDA model but come with additional computational complexity.

**Posterior inference:** To infer the posterior probability of the hidden variables given the observed documents, we have utilized an online learning method (Hoffman et al., 2010) based on variational inference (Jordan et al., 1999). The benefit of online learning methods is that the corpus need not be in-memory, as is the case with batch methods. However, besides variational inference, other methods such as Markov Chain Monte Carlo (MCMC) based methods exist, such as Gibbs sampling (Griffiths and Steyvers, 2004). It is argued that variational inference is computationally faster, and that Gibbs sampling is in principle more accurate (Porteous et al., 2008). However, in a comparative study, different inference techniques show similarly accurate results when the hyper-parameter of the Dirichlet priors (studied in Chapter 3) are optimized (Asuncion et al., 2012).

**Model initialization:** One of the drawbacks of probabilistic topic models, including LDA, is that they uncover topics regardless of whether they are naturally there (Blei and Lafferty, 2009), and that the model can yield different topics at different initializations (Chuang et al., 2014). Adequately pre-processing and optimizing for model quality, for instance through coherence scores, can usually produce a set of interpretable and meaningful topics, as we have shown in the various chapters. However, we have identified topics present within one model, and absent in another, even when optimizing for topic quality across different random initializations (see Chapter 3). The discrepancies between uncovered topics are inherently caused by approximating the posterior distribution, as a closed form solution is intractable, and where solutions might converge to a sub-optimal solution (i.e., local optima). Typically, starting the model with different random initializations is one way to go, which we experimented with in Chapters 2, 3, 6 and 7, although such an approach was not feasible for the extensive corpus of 46,000 full-text publications studied in Chapter 5. This drawback makes LDA less stable, reproducible, and reliable, and solutions to circumvent or mitigate these problems are scarcely studied. Optimizing for starting points (Roberts et al., 2016), using different similarity metrics for model comparison (Koltcov et al., 2014), and adapting approaches from community detection algorithms for networks (Lancichinetti et al., 2015) are a few approaches to consider.

**Topic labeling:** For readability and interpretability of the results, we have chosen to attach semantically meaningful labels to the probability distributions over words (i.e. the topics). However, the labeling of the topics is a very subjective endeavor. In all cases, we have used fisheries domain experts to perform the labeling task and have provided, next to the topic’s top words, a selection of document titles strongly associated with the topics, abstracts, and several visualizations to aid in correctly determining topic labels. Nevertheless, manual topic labeling, even though considered
the gold standard in topic labeling (Lau et al., 2011), is limited by the subjectivity inherent in human interpretation (Urquhart, 2001), and an analysis of the topics by other domain experts could yield different results.

**Topic quality:** To evaluate the quality of latent topics, one can fit several topic models to a training set of documents and calculate a model fit, such as perplexity or the log-likelihood, on a test set of the data (Scott and Balldridge, 2013). The model that best fits the test set would be considered a better model. However, topic models are used by humans to interpret and explore the documents, and there is no technical reason that the best-fitted model would best help in performing this task (Blei, 2012; Boyd-Graber et al., 2014). In fact, research has shown that such measures negatively correlate with human interpretation (Chang et al., 2009). Within this thesis, we have adopted the $C_v$ coherence measure (Röder et al., 2015). It quantitatively measures the quality of the latent topics from the perspective of human interpretation to achieve near-human accuracy (Boyd-Graber et al., 2014), combined with a qualitative assessment by domain experts. The $C_v$ coherence measure has been shown to outperform all other available coherence measures and can thus be viewed as an adequate measure for topic quality. However, using a reference corpus, such as Wikipedia, can potentially improve the estimate of word frequencies and word co-occurrences that are part of coherence measures, and can potentially increase the quantification of topic quality (Yang et al., 2017).

**Research domain:** All our LDA computational experiments have been performed on documents related to the fisheries domain. While the reasons for this domain are explained in Chapter 1, the results presented in this thesis might not necessarily be generalizable to other forms of documents. However, optimization of Dirichlet priors (Chapter 3) has been shown to provide similar results as previously studied for documents related to news and patent data. Also, given that we studied documents from a domain-specific field (i.e., fisheries), our results might be more generalizable to documents from other domain-specific sciences compared to documents found in more general-purpose journals such as *Nature*, *Science*, *PNAS*, and *PLOS ONE*.

### 8.2.2 Fisheries Domain

**English language:** Throughout this thesis, we have explored latent topics derived from scientific publications that were entirely based on English-speaking journals. The analysis of latent topics addressing the four pillars of fisheries’ sustainability, and the analysis of spatial and temporal characteristics of authors and publication output all reflect an Anglophone bias. In doing so, we have disregarded all fisheries science output published in non-English languages, such as Spanish, Portuguese, and Chinese. Even though English is considered the lingua franca in science (Montgomery, 2103), research has shown that, for instance, in 2014, 35.6% of 75,513 publications on biodiversity conservation were published in non-English languages (Amano et al., 2016).
As a consequence, the results presented in this thesis do not capture all the global scientific output of fisheries.

**Journal selection:** In terms of our corpus, while this contains high-ranking journals in fisheries science, we acknowledge that it does not capture the entire spectrum of work being done in the area of fisheries. By focusing on peer-reviewed fisheries journals, we have, for instance, ignored all scientific output published in the gray literature, such as governmental or institutional reports. Additionally, by heavily focusing on fisheries journals, some work related to fisheries, in particular, that which is oriented towards the social sciences, may be published in journals that have a more general focus, or a stronger leaning towards those sciences. As a consequence, our work does not reflect all output regarding fisheries science. However, we have aimed to include, especially in Chapters 4, 5 and 6, a large number of journals explicitly addressing fisheries science in their aims and scopes and those considered to be high-impact journals in the field.

### 8.3 Future Work

Some future work to consider can be directly related to some of the limitations described in Section 8.2. When using LDA, relaxing the bag-of-words assumption (Griffiths et al., 2005; Wallach, 2006b) and the document exchangeability assumption (Blei and Lafferty, 2006) are useful areas to further explore. Also, the use of non-parametric topic models (Whye Teh et al., 2004) can reveal additional and possibly more accurate results. For the domain of fisheries science, including non-English publications and incorporating journals not specifically targeted at fisheries science are approaches to extend the work presented in this thesis. The field of probabilistic topic modeling is an active area of development, with many new approaches presented during recent years. We want to briefly highlight three interesting and promising alternatives that are useful when exploring and studying the scientific output of a particular domain. We conclude with an overview of extensions of LDA that highlight the potential of probabilistic topic model research.

**Labeled LDA (L-LDA):** Scientific publications are typically annotated with author-defined keywords, and frequently also by journal-assigned keywords. However, the association between a specific keyword and the relevant part of the text is not provided. Labeled LDA (Ramage et al., 2009) is an extension of LDA that learns to find such associations and can be used to zoom-in or filter for specific content.

**Author-topic model (ATM):** The author-topic model (Rosen-Zvi et al., 2004) is an extension of LDA that combines author information into the learning process. It infers a topical distribution per author, and thus within the context of scientific publications, it reveals what topics a specific author publishes. Authors who have similar topic distributions, measured by, for instance, KL-divergence (Kullback and Leibler, 1951) or
the Hellinger distance (Hellinger, 1909), can be viewed as publishing similar research. Additionally, entropy measures can, for instance, show if authors publish more narrowly or broadly oriented work. We have experimented with the ATM, and a working implementation can be found in the Gensim (Rehurek and Sojka, 2010). However, as per our analysis, the current implementation fails to handle datasets with roughly more than 10,000 authors.

**Correlated topic model (CTM):** The correlated topic model (Blei and Lafferty, 2007) allows topics to correlate with each other, as a document about the topic “cars” is more likely to also be about “emissions” than it is about diseases. The standard LDA model uses a Dirichlet distribution which implicitly assumes independence between topics. The use of a more flexible distribution to model topics in documents, such as the logistic normal, allows for variability between the components (Blei and Lafferty, 2009). The correlated topic model has been shown to fit the data better and provides a more realistic model of the latent topical structure in documents.

**Other topic models to consider:** Many other types of topic models exist which show potential directives for future research into probabilistic topic models. For example the relational topic model (Chang and Blei, 2010), the spherical topic model (Reisinger et al., 2010), the sparse topic model (Wang and Blei, 2009), the bursty topic model (Doyle and Elkan, 2009), the supervised topic model (Blei and McAuliffe, 2007), the biterm topic model (Yan et al., 2013), topic modeling with network regularization (Mei et al., 2008), the Pachinko allocation model (Wei and McCallum, 2006), the Markov topic model (Wang et al., 2009), the polylingual topic model (Mimno et al., 2009), the cross-collection topic model (Paul, 2009), the differential topic model (Chen et al., 2015), and topic modeling over time (Wang and McCallum, 2006).

### 8.4 Personal Reflections

Before I started my PhD, I had mainly been working with people very close to my own discipline: computer science. We spoke the same language and we were wearing similar glasses (scientific makeup). When I began my PhD journey, this became vastly different as I suddenly plunged into a pool of fisheries scientists, marine conservationists, political scientists, social scientists, and people from many other disciplines. To make things even more challenging, being part of the European Training Network SAF21 (Social Science Aspects of Fisheries for the 21st Century), I was surrounded by people from different countries and cultures. How to communicate, navigate, and collaborate with my new colleagues, seniors, project members, beneficiaries and host institutions revealed not to be a trivial task, and one that I had to learn. I made some substantial steps, but there are many more hurdles to overcome. In this final chapter, I want to highlight three important lessons that I learned and that are worth mentioning explicitly.
“Know your audience” became an important one for me. Oftentimes, I wanted to present the underlying fine-grained mechanisms of the algorithms I was studying, but the better choice was to explain what it does, not how it does it. Also, writing for fisheries journals was a skill I had to master, sometimes the hard way with the peer review process. Within this thesis, I have tried to leave out some of the overly unnecessary technical jargon. However, there is still much to learn in this respect and, thankfully, throughout this PhD journey, the feedback I have received from peers outside of my own discipline has always been constructive and helpful.

“Do what you love to do”. A PhD can be a stressful undertaking, and the Internet is filled with articles, blogs, and fora stating that PhD researchers suffer from mental health problems such as chronic anxiety and clinical depression. Doing research, analyzing, and just coding away is what I love to do, and this is something I was lucky enough to do most of the time. Perhaps apart from the paper writing and review process, the PhD journey felt like a breeze through life. Okay, perhaps a breeze most of the time.

“Three years is challenging”. The modern version of a streamlined three-year PhD program which is typically found within EU-funded projects is a very challenging one. Especially, when it is combined with training camps, network meetings, secondments, and the tedious and long publication process. Thankfully, being part of a European Training Network gave me the opportunity to live, study and work in diverse and multicultural environments in the Netherlands (Utrecht University), Norway (UiT – The Arctic University of Norway) and the UK (Manchester Metropolitan University). It also gave me the opportunity to be surrounded by interesting and inspiring colleagues, while exploring many other parts of the world (Denmark, Spain, Portugal, Iceland, US, Japan, Italy, and Germany) has made the journey worthwhile and highly enlightening.


BIBLIOGRAPHY


FAO. The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all. 2016.


C. Finley. *All the Fish in the Sea: Maximum Sustainable Yield and the Failure of Fisheries Management*. The University of Chicago Press, Chicago, IL, USA, 2011.


BIBLIOGRAPHY


E. Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen
Veränderlichen. *Journal für die reine und angewandte Mathematik (Crelle's Journal)*,

R. Helms and K. Buijsrogge. Knowledge Network Analysis: A Technique to Analyze
Knowledge Management Bottlenecks in Organizations. In *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, pages 410–414. IEEE,

C. C. Hicks, A. Levine, A. Agrawal, X. Basurto, S. J. Breslow, C. Carothers, S. Charnley,
S. Coulthard, N. Dolsak, J. Donatuto, C. Garcia-Quijano, M. B. Mascia, K. Norman,
M. R. Poe, T. Satterfield, K. St. Martin, and P. S. Levin. Engage key social concepts

M. Hilbert and P. Lopez. The World's Technological Capacity to Store, Communicate,
ence.1200970.

H. Hill and M. Lackups. Journal Publication Trends Regarding Cetaceans Found in
Both Wild and Captive Environments: What do we Study and Where do we Publish?

J. Hoekman, K. Frenken, and R. J. Tijssen. Research collaboration at a distance: Chang-

In *NIPS’10 Proceedings of the 23rd International Conference on Neural Information Pro-
Associates Inc.


international ACM SIGIR conference on Research and development in information re-
10.1145/312624.312649.

D. Hoggarth, C. Mees, and C. O’Neill. *A guide to fisheries stock assessment using the

D. D. Hoggarth, S. Abeyasekera, R. I. R. Arthur, J. R. J. Beddington, R. R. W. Burn,
for fishery management : a framework guide to the stock assessment tools of the


ICES. SIHD, 2017.


H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. 2012. doi: 10.1016/B978-0-12-381479-1.00001-0.


P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603, 2010. doi: 10.1007/s11192-010-0202-z.


G. M. Winder. Introduction: Fisheries, Quota Management, Quota Transfer and Bioeconomic Rationalization. In G. M. Winder, editor, *Fisheries, Quota Management*


All Published Work by Shaheen Syed


Papers under submission


Summary

It is estimated that the world’s data will increase to roughly 160 billion terabytes by 2025, with most of that data occurring in an unstructured form. Today, we have already reached the point where more data is being produced than can be physically stored. To ingest all this data and to construct valuable knowledge from it, new computational tools and algorithms are needed, especially since manual probing of the data is slow, expensive, and subjective.

For unstructured data, such as text in documents, an ongoing field of research is probabilistic topic models. Topic models are techniques to automatically uncover the hidden or latent topics present within a collection of documents. For instance, it allows us to infer the topical content of thousands or millions of documents without prior labeling or annotating of these documents. This unsupervised nature makes probabilistic topic models a useful tool for applied data scientists to interpret and examine large volumes of documents for extracting new and valuable knowledge.

The popularity of topic models is further driven by the available open source libraries that handle most of the complexity. However, applied data scientists regularly treat topic models as black boxes without thoroughly exploring their underlying assumptions and (hyper-)parameter values. This black box approach arises from the inherently complex statistical nature of topic models and, with improper use, their results can be interpreted as questionable. With this in mind, this thesis aims to provide answers regarding how to optimally and efficiently employ probabilistic topic models to large collections of documents. In this pursuit, we have chosen the domain of fisheries science as our testbed and fisheries scientific publications as our source of textual data. We have taken a systematic and iterative knowledge discovery process called knowledge discovery in databases (KDD) as a blueprint for effectively discovering knowledge from textual data. The main research question in this thesis, therefore, is:

How can we improve the knowledge discovery process from textual data through latent topical perspectives?

The first three chapters of this thesis seek to understand how different types of textual
data, pre-processing steps, and hyper-parameter settings of probabilistic topic models affect the quality of the derived latent topics. Specifically, we study the effects that using abstract and full-text data, and the use of different prior distributions have on the quality of the latent topics. Additionally, we explore alternative approaches for evaluating the quality of latent topics through semantic lexicons. In doing so, we contribute to the methodological analysis and optimization of topic models, providing a starting point for researchers who want to apply topic models with scientific rigor to scientific publications.

The remaining three chapters are aimed at the interpretation of the latent topics and how such (raw) latent topics can be turned into useful (fisheries) domain knowledge. We explore ways to go beyond exploring the latent structures of the documents and interpret the results in a broader context. That is, we aim to construct new knowledge and shed light on the ecological, social, economic, and institutional considerations within fisheries for increased fisheries sustainability. In other words, by applying topic models to fisheries science publications, we study the domain through a new computational lens and expand on traditional approaches to assess fisheries sustainability.

Throughout this thesis, and within each chapter, specific phases of the KDD process are covered. Combined, they provide guidelines on how to optimize the knowledge discovery process with the aim to understand the latent topical content of scientific publications better.
Samenvatting

Tegen het jaar 2025 zal alle data ongeveer 160 miljard terabytes omvatten en zal de meeste data in ongestructureerde vorm zijn. Op dit moment hebben we al het punt bereikt waar meer data wordt geproduceerd dan fysiek opgeslagen kan worden. Om al deze data te verwerken én om er bruikbare kennis uit te ontleden zijn nieuwe computationele methoden en algoritmes nodig, vooral omdat het handmatig verwerken van al deze data traag, duur en subjectief is.

Voor ongestructureerde data, zoals tekst in documenten, zijn probabilistische topic modellen een actief onderzoeksgebied. Probabilistische topic modellen zijn methoden om automatisch de onderliggende of verhulde onderwerpen of thema’s (hier topics genoemd) bloot te leggen van grote verzamelingen documenten. Via probabilistische topic modellen is het bijvoorbeeld mogelijk om de actuele inhoud van duizenden of miljoenen documenten af te leiden zonder deze documenten vooraf te labelen of van aantekeningen te voorzien. Deze automatische methodiek maakt probabilistische topic modellen een handig hulpmiddel voor onderzoekers om uit grote hoeveelheden documenten nieuwe en waardevolle kennis te extraheren.

De populariteit van topic modellen wordt deels gevoed door de vele beschikbare open-source bibliotheken die de meeste onderliggende complexiteit ervan wegnemen. Echter, de werking van topic modellen wordt door onderzoekers regelmatig als een ‘black box’ beschouwd, en de onderliggende aannames en parameters van topic modellen worden niet altijd grondig verkend. Deze ‘black box’ beschouwing kan deels verklaard worden door de complexe statistische aard die inherent is bij probabilistische topic modellen, én bij verkeerd gebruik kunnen de resultaten ervan als twijfelachtig worden geïnterpreteerd. Met dit in acht genomen probeert dit proefschrift antwoorden te vinden om probabilistische topic modellen zo optimaal en effectief mogelijk toe te passen. In dit streven hebben we het domein van de visserij wetenschap gekozen als test casus én zijn de wetenschappelijke publicaties in dit domein gekozen als bron van tekstuele data. Daarnaast is er een kennisontdekkingsproces genaamd “Knowledge Discovery in Databases” (KDD) gebruikt om als blauwdruk te fungeren voor het effectief omzetten van tekstuele data naar kennis. De hoofdonderzoeksvraag van dit
SAMENVATTING

Proefschrift is daarom:

Hoe kunnen we het kennisontdekkingsproces van tekstuele gegevens verbeteren door de onderliggende topics in de teksten bloot te leggen?

In de eerste drie hoofdstukken van dit proefschrift proberen we inzicht te krijgen in hoe verschillende soorten tekstuele data, voorbewerkingsstappen en hyper-parameters een effect kunnen hebben op de verhulde topics uit grote verzamelingen documenten. We bestuderen in het bijzonder de effecten die het gebruik van abstracte of volledige tekst én verschillende ‘prior’ distributies hebben op de kwaliteit van de topics. Daarnaast verkennen we alternatieve benaderingen voor het evalueren van de kwaliteit van verhulde topics via semantische lexicons. Hiermee dragen we bij aan de methodologische analyse en optimalisatie van topic modellen én bieden we een startpunt voor onderzoekers die topic modellen met wetenschappelijke nauwgezetheid willen toepassen op wetenschappelijke publicaties.

De overige drie hoofdstukken zijn gericht op de interpretatie van de topics en hoe deze kunnen worden omgezet in nuttige (visserij) domeinkennis. We onderzoeken dus manieren om de latente structuren, de verhulde topics, van documenten te interpreteren en de resultaten ervan in een bredere context te plaatsen. Dat wil zeggen, we willen nieuwe kennis opdoen en licht werpen op de ecologische, sociale, economische en institutionele vraagstukken binnen de visserij wetenschap, specifiek om de duurzaamheid van de visserij te vergroten. Met andere woorden, door het toepassen van topic modellen op wetenschappelijke publicaties van de visserij bestuderen we het wetenschappelijke domein via een nieuwe computacionele lens, én breiden we hiermee de traditionele benadering voor het beoordelen van duurzaamheid in dit domein uit.

In dit proefschrift en binnen elk hoofdstuk worden specifieke fasen van het kennisontdekkingsproces behandeld. Gezamenlijk bieden de hoofdstukken antwoorden en richtlijnen voor het optimaliseren van het kennisontdekkingsproces met als doel een beter inzicht en begrip te ontwikkelen voor het extraheren van latente topics uit wetenschappelijke publicaties.
Curriculum Vitae

Shaheen Syed was born on February 9th, 1985, in Rotterdam, the Netherlands. He obtained his bachelor’s degree in Informatics from the Rotterdam University of Applied Sciences in 2011 and his master’s degree (cum laude) in Business Informatics from Utrecht University in 2013. His master’s thesis explored regression models for the prediction of open source project failure. After his studies, he worked in the industry for three years as a data analyst and software developer.

In 2016, Shaheen started his EU Horizon2020 Marie Skłodowska-Curie funded PhD research at the Manchester Metropolitan University while being enrolled at the Department for Information and Computing Sciences at Utrecht University. During his PhD, he was a visiting researcher at the University of Tromsø in Norway where he developed machine learning applications for an EU Horizon2020 project, and he organized and chaired a machine learning workshop for an intergovernmental organization named the International Council for the Exploration of the Sea (ICES).