Text analysis tools for identification of emerging topics and research gaps in conservation science

Martin J. Westgate[1, 2], Philip S. Barton[1], Jennifer C. Pierson[1] & David B. Lindenmayer[1]

[1] The Fenner School of Environment and Society, The Australian National University, Canberra ACT 0200, Australia

[2] Corresponding author: martin.westgate@anu.edu.au

Keywords: Latent Dirichlet Allocation, Hot topics, Synthesis, Surrogates, Indicators

Running title: Text analysis for research synthesis

Abstract

Keeping track of conceptual and methodological developments is a critical skill for research scientists, but this task is becoming increasingly difficult due to the high rate of academic publication. As a crisis discipline, conservation science is particularly in need of tools that facilitate rapid yet insightful synthesis. Here, we show that how a commonly-used method for text mining – Latent Dirichlet Allocation or 'topic modeling' – can be used in conjunction with statistical tools already familiar to ecologists (cluster analysis, regression, and network analysis) to investigate trends and identify potential research gaps in the scientific literature. We then demonstrate these properties using the literature on ecological surrogates and indicators as a case study. Analysis of topic popularity shows a strong emphasis on the monitoring and management of fragmented ecosystems, while gap analysis suggests a greater role for genetic surrogates and indicators. Our results show that automated text analysis methods need to be used with care, but can provide information that is complementary to that given by systematic reviews and meta-analyses. Text analysis has strong potential for increasing scientists' capacity for rapid and detailed synthesis of conservation science.

Introduction

A key skill for researchers is the ability to understand historical and emerging ideas in their field of specialization, and to synthesize this information to generate novel concepts and methods. Therefore, scientists' capacity to keep track of developments within their research community is fundamental to scientific progress. Although this observation applies across the sciences, tracking research developments is particularly urgent for fast-moving disciplines such as conservation science. This is because their findings have direct implications for evidence-based conservation (Sutherland et al. 2009). Put differently, in a crisis-based discipline, any new developments should ideally be adopted as fast as possible to help prevent further declines and extinctions of threatened species (Soulé 1985). Unfortunately, the quantity of scientific literature currently being published threatens to overwhelm scientists' capacity to keep track of new research (Larsen & von Ins 2010). Consequently, increases in the volume and availability of scientific information need to be matched by increases in the availability of tools for interpreting that content (Boyack & Klavans 2014).

A potentially useful development has been the growth of a suite of statistical methods for investigating patterns and trends in collections of documents (known as 'corpora', singular 'corpus'). Several of these approaches investigate combinations of words within articles (i.e., text analysis) and seek to elucidate the key ideas discussed within a corpus (Griffiths & Steyvers 2004; Grimmer & Stewart 2013; Rusch et al. 2013). Consequently, text analysis has the potential to generate conceptual insights traditionally available only through narrative review, but with the speed and quantitative rigor that characterizes modern scientific investigation (Grimmer & Stewart 2013). However, text analysis is rarely used in ecology and conservation, and so the forms of inference that can be achieved using these methods, and their usefulness for understanding research trajectories, remain poorly articulated. This is unusual given the trend towards greater quantification in ecology and conservation synthesis (Lortie 2014).

In this article, we argue that conservation science is well placed to capitalize on text-analysis tools, as methods for summarizing the results of text mining algorithms have a number of similarities to existing and commonly used methods in ecology (Table 1). We will show how a combination of approaches can be used to guide a broad understanding of trends within academic corpora, using the literature on ecological surrogates and indicators as a case study. The literature on ecological surrogates is particularly suited as a case study of text analysis because it is a large and diverse body of work that has grown dramatically in recent decades (Westgate et al. 2014), thereby presenting a considerable challenge to synthesis. Surrogates are also important from a conservation perspective because they provide the data underpinning nearly all conservation decisions (Collen & Nicholson 2014). Improved understanding and application of surrogates should therefore lead to more efficient ecosystem monitoring and management. Therefore, our case study addresses critical barriers to wider adoption of text analysis, by discussing how complex topics can be synthesized to allow informed decisions regarding research priorities. We conclude by outlining some potential benefits and pitfalls of automated approaches to research synthesis.

Tools for investigating academic corpora

The fundamental problem of text analysis is how to decompose a set of documents into a smaller number of thematic elements (known as 'topics') that can be used to interpret patterns in the corpus. A particularly useful method for this application is Latent Dirichlet Allocation (LDA, sometimes called 'topic modeling'; Blei et al. 2003). In LDA, topics are defined using sets of words that co-occur with unusual frequency, meaning that each topic can be interpreted as a meaningful combination of ideas within the corpus. Moreover, each article is assumed to consist of a number of topics; hence the user can identify the weight assigned to each topic within each article. Because its results can be readily interpreted, LDA has been widely adopted for text analysis in a range of fields including journalism (Rusch et al. 2013), politics (Grimmer & Stewart 2013), and social network analysis (Weng et al. 2010).

At this juncture, it is likely that readers will observe several close parallels between ecological modeling and text analysis. First, the popularity of LDA as a research tool reflects a shift towards model-based multivariate analysis that also can be seen in ecology (Wang et al. 2012). Second, just as methods that are common in ecology and conservation (such as ordination; Legendre & Legendre 2012) could be used for identifying associated words within texts, LDA can be applied to ecological problems such as classification of image time series (Niebles et al. 2008), or the analysis of species assemblages (Valle et al. 2014). Third, similar caveats apply to LDA as to ordination of species occupancy or abundance data. For example, it is common practice to delete rare species from site by species matrices when performing ordinations of species composition. This is to avoid the potentially strong influence of singletons and doubletons on the outcome (Legendre & Legendre 2012). The same process is often advisable for word matrices, in that very rare words can disproportionately influence the algorithm that determines topic composition (Blei et al. 2003). In contrast, very common species only weakly influence clustering of species ordinations, while very common words (known as stop words; e.g. 'the' or 'and') are typically removed during text analysis as they provide little information content (Silva & Ribeiro 2003). As these observations make clear, methods for text analysis are strongly related to those in common use for ecological and conservation problems, a point that we will return to in the discussion.

Although LDA is not the only method capable of text classification, in the remainder of our article, we assume the use of LDA for topic identification. Below, we outline four methods that build on one another to provide complementary forms of information regarding the content of study corpora (see Table 1). Importantly, these methods are intended to facilitate interpretation of the content provided by LDA; they are not tools that can be applied in isolation of a method for topic identification. We then apply these methods to our case study on ecological surrogates and indicators.

Table 1: Methods for interpreting topic content (using topics identified using LDA) that are discussed in this article, and their analogues in ecological modeling

Statistical approach	Text analysis	Ecological modeling
Cluster analysis	Identify clusters of similar topics based on the dominant words they contain (i.e. topics; Blei et al. 2003).	Identify clusters of similar locations based on the species they contain (Legendre & Legendre 2012).
Comparison of frequency distributions	Investigation of the relationship between the number articles containing a topic, and the weight of that topic within each article.	Investigation of the relationship between the number of sites occupied by a species, and the abundance of that species within each site (Gaston et al. 2000).
Linear (mixed) models	Quantify trends in the popularity of a number of topics (Griffiths & Steyvers 2004).	Quantify trends in the abundance of a number of species (Pollock et al. 2012).
Network analysis	Quantify the extent to which pairs of topics tend to occur in similar vs. different texts.	Quantify the strength of associations between pairs of species or individuals (Ings et al. 2009).

Topic similarity

A key problem when using LDA is how to interpret the 'meaning' of each research topic, for which a useful first step is to identify clusters of similar topics. This is achievable because LDA allows extraction of the weight that each word contributes to each topic, which can then be subjected to standard dissimilarity and ordination-based methods (Legendre & Legendre 2012). The value of this method is partly in validation; i.e. questioning whether topics that contain similar words appear similar to the user, based on their understanding of the corpus under investigation. However, it also provides information critical to the interpretation of other trends, such as whether similar topics differ in popularity (see below).

Popularity, growth, and hot topics

Having found a way to classify articles into topics based on the information they contain, one obvious question to ask is: Which topics are most popular? This question can be decomposed into two parts: 1) the total number of articles that have been published on a topic in the period for which data are available; and 2) the extent to which that topic has changed in popularity over that period. The former point gives important information on total research effort within a corpus, while the latter is commonly used to assess which topics are 'hot' (i.e. have experienced positive growth) versus 'cold' (negative growth) within a given research community (Griffiths & Steyvers 2004).

As was the case with topic similarity (see above), assessing topic popularity also uses methods that will be familiar to any ecologist; namely linear regression. In its simplest form, this amounts to a question of how the number of published articles per topic (response variable, denoted by y) changes over time (predictor variable, denoted by x). A useful method for answering this question is to split article counts by topic, and then use mixed models (Bolker et al. 2009) to allow an intercept (i.e. mean number of publications, if the predictor variable is centred) and slope (i.e. rate of change in number of publications) for each topic. For example, the number of citations over time can be investigated using a Poisson mixed model where the expected response is given by the formula:

$$\log(E_{(y|u)}) = \alpha + (\beta + b)x + u$$

Where $E_{(y|u)}$ is the expected response conditional on u, α and β are the fixed intercept and slope (respectively), u and b are the random intercepts and slopes (respectively) that are normally distributed with mean zero, x is the predictor (time), and model variance is given by σ^2_u, σ^2_b. In such a model, topics with positive random intercepts (i.e. $u > \alpha$) can be interpreted as having *higher-than-average numbers* of articles written about them in the period for which data are available. Similarly, topics with positive random slopes would have *higher-than-average growth* in publications during the same period.

Specificity and generality

So far we have discussed interpretation of LDA outputs as a simple data mining exercise. However, it would be an oversimplification to assume that meaningful insights can be generated simply using this approach. A particular issue is that because topics are identified according to sets of co-occurring words, it is possible that some topics may reflect broad themes common to many articles within the corpus (i.e. 'general' topics), rather than describing the key theme of the article in question ('specific' topics). Consequently, it would be useful to be able to calculate some measure of where each topic sits on this axis (i.e. from general to specific).

One method that can be used to assess topic specificity versus generality is to examine the distribution of topic weights within articles. Because LDA can be used to calculate a matrix describing the weight of each topic (columns) within each article (rows), articles can be readily classified by assigning each article to the topic that has the highest weight (i.e. is the maximum for that row). This approach is sensible if one topic receives a much higher weight for a given article than do all the remaining topics, but is problematic if all topics have very similar weights. The details of this process are important because of their implications for interpreting patterns across the whole corpus. In particular, a topic may be rarely 'selected' (i.e. rarely be the highest-weighted topic), but may have moderate weight across a range of articles within the corpus. Therefore, by comparing the mean weight of a topic in 'selected' versus 'unselected' articles, one can make an assessment of the extent to which that topic permeates the literature ('generality'), or in contrast, is restricted to only a subset of articles ('specificity').

Identifying research directions

A final goal that readers might have during a literature review may be to identify future research directions. Although a common and even necessary part of literature review, the idea that we might seek to automate the process of predicting future directions will be strange and even alarming to some readers. Certainly, there are inherent difficulties and ambiguities in this form of prediction. Nonetheless, here we outline some means by which text analysis can be used to facilitate researchers' intuition regarding productive research directions.

Several authors have sought to quantify how ideas permeate research networks. For example, Wang et al. (2013) showed that article citation rates show several predictable attributes, suggesting that scientific impact can be quantified – and therefore predicted – to some broad degree of accuracy. A more useful observation for our purposes would be a theory of how influential ideas emerge from an existing body of literature. One such theory is that scientific progress can be hastened by unifying well understood but disparate concepts (Chen et al. 2009). In practice, such research 'gaps' could be identified as pairs of topics that are unusually separate within the corpus, both in terms of their thematic content, and the articles in which they appear. Such a theory does not preclude the possibility that progress might also occur through spontaneous novel insights ('eureka moments'); but such occurrences are inherently less amenable to prediction, and so can be ignored for our purposes. In this paper, we will refer to our investigation of potential research gaps as 'gap analysis', while acknowledging that this term has a range of alternate meanings, both within and outside of the ecology and conservation literature.

Case study: Ecological surrogates and indicators

To demonstrate how the methods that we have outlined above can be used in practice, below we apply them to a corpus of article abstracts from the scientific literature on ecological surrogates and indicators. Although this is an area of strong research interest to the authors, the insights that we ascertain below derive exclusively from the text analysis methods that we have outlined above. The same process could therefore be applied to any corpus, bearing in mind that interpretation will always be critical to the conclusions that users will draw from their results.

Surrogates and indicators are proxies that are used to draw inference regarding complex ecosystems from a manageable amount of data, and for this reason they are critical for environmental management (Noss 1990; Collen & Nicholson 2014). This body of literature is an interesting application for text analysis due to its sheer size and diversity (up to 11,000 articles; Westgate et al. 2014) which hinders effective synthesis (Lindenmayer & Likens 2011). For example, simple applications of the surrogate concept may test whether particular habitat attributes consistently predict the occurrence or abundance of valued species (Lindenmayer et al. 2014), or whether a species is restricted to (i.e. is indicative of) a particular ecosystem type (De Cáceres et al. 2010). In contrast, complex applications may involve identification of surrogates for broad ecosystem attributes such as resilience (Bennett et al. 2005). These issues represent significant challenges to researchers whose goal is to synthesize knowledge across the full range of methods and applications in surrogate ecology (McGeogh 1998). Using this corpus, we used LDA combined with the methods described above to identify (i) topic similarity, (ii) popularity and growth of topics, (iii) specificity and generality of topics, and (iv) potential research directions.

Methods

We completed a case study by investigating the abstracts of articles that cited a single seminal work on ecological surrogates and indicators (Noss 1990; n= 1160), together with those articles that cited any of the 100 most highly-cited articles that cited Noss' paper (i.e. the 'second generation' citations of Noss' 1990 paper; n= 8674). We addressed our study goals as follows. We first identified 25 topics within this corpus using an LDA model fitted using the 'topicmodels' package (Gruen & Hornik 2011) in the R statistical program (R Core Development Team 2014), and named each topic using our assessment of the top 20 highest-weighted keywords for that topic (Appendix S1). We then investigated each of our goals as follows: (i) We investigated topic similarity by calculating the Euclidean distance between each pair of topics (matrix D1), using a matrix (M1) whose values represented the $\log_{10}$ transformed weight assigned to each word/topic combination. (ii) We investigated topic popularity using mixed models as implemented in lme4 (Bates et al. 2014), and (iii) calculated topic generality using information on the weight assigned to each article/topic combination (M2; the associated distance matrix was named D2). (iv) Gap analysis involved calculating the product of D1 and D2, after scaling each matrix to range between zero and one. A more complete description of our methods is available in the online supporting information (Appendix S1).

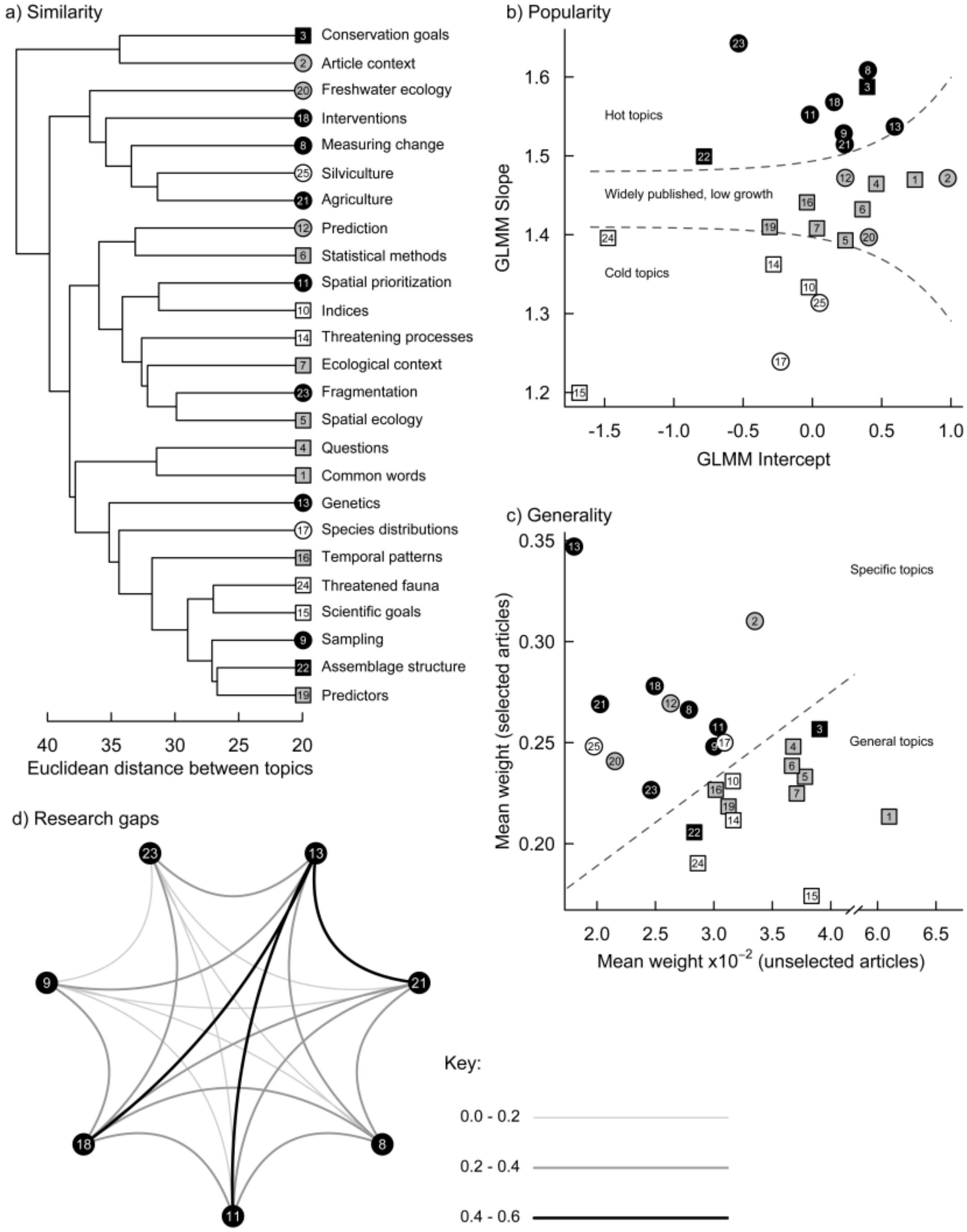
Results

Clustering of topic content using word-based similarity (i) divided our dataset into three broad groups (Fig. 1a). The first group consisted of research into manipulable or dynamic systems (silviculture, agriculture and freshwater ecology), and concepts relevant to the study of those systems (interventions, measuring change). A second group contained topics describing subthemes within the spatial ecology literature (including spatial prioritization and fragmentation), while a final group contained topics describing basic concepts in community ecology (including threatened fauna, assemblage structure, and common predictors of change). While these clusters described meaningful patterns in the dataset, each was matched by an outgroup that contained very broad concepts (e.g. questions, prediction, article context). Finally, three topics formed an outgroup to our community ecology cluster (including genetics, species distributions and temporal patterns), suggesting that these topics had similar goals to community ecology, but used sufficiently different language to be classified as distinct.

Topic popularity analysis (ii) showed that historically popular topics had intermediate growth (Fig. 1b). Fragmentation research had the highest growth rate of any topic in our analysis, which was one of a number of unexpected patterns. In particular, silvicultural research appears to be decreasing in popularity (relative to the mean), despite increases in the conceptually similar field of agriculture. Further, several topics appear to be decreasing relative to our anecdotal assessment of their frequency in the broader ecology literature, namely freshwater ecology, research on spatial priortization, and threatening processes (which included keywords related to urbanization and climate change; Appendix S1). Finally, when taken as a group, 'hot topics' in our study corpus appeared to focus on ways to measure and ameliorate significant threatening processes (e.g. agriculture, fragmentation, interventions), while there was no clear relationship between 'cold' topics.

Topic generality analysis (iii; Fig. 1c) allowed us to distinguish between topics that had high weight in a subset of articles and low weight elsewhere (specific), to topics that were rarely the focus of whole articles, but occurred more evenly throughout the literature (general). Highly 'general' articles tended to include terms that were broadly descriptive of the scientific process. For example, scientific and conservation goals were listed as general topics, as were collections of words listed as 'questions' and 'common words' in our analysis. In contrast, genetics was the most specific topic in our analysis (Fig. 1c). Finally, comparison of results from popularity and generality analysis showed that nine of the 14 low-growth topics identified by popularity analysis (Fig. 1b) were classified as 'general' (Fig. 1c), while only two of the nine 'hot' topics were classified as general. Comparing these findings to topic similarity analysis (Fig. 1a) showed that each cluster of similar topics contained both hot and cold topics, suggesting that subtle differences in topic content can make a large difference to their popularity in the academic literature.

Fig. 1: Classification of topics in the academic literature on surrogates and indicators identified using LDA, showing (clockwise from top left): (a) similarity, (b) popularity, (c) generality, and (d) research gaps. See text and supporting information (Appendix S1) for details of all calculations. Topic colors and shapes in all panels are set according to categories shown in panels (b) and (c). Categories represent coarse groupings defined for example purposes only, and should not be considered as statistically robust.



Finally, analysis of research gaps (iv) showed that several connections between specific, rapidly-growing topics remain poorly investigated. The topics that met these criteria referred to threatening processes (agriculture, fragmentation), management actions to ameliorate the impacts of those processes (prioritization, restoration), or to methods for quantifying ecological responses to either category (sampling, measuring change, and genetics; Fig. 1d). Of these, genetics displayed the greatest degree of separation from the remaining topics, suggesting that genetic approaches for understanding ecosystem changes remain poorly utilized in the surrogate ecology literature. In contrast, work investigating the interface between fragmentation and protected or agricultural areas is already well developed, suggesting lower priority for additional research effort.

Discussion

In this article, we have shown how a suite of tools already familiar to ecologists can be used in conjunction with existing text-analysis methods (LDA) to rapidly summarize the major themes discussed within academic corpora. Further, we have demonstrated these properties using a case study on ecological surrogates and indicators. Our key message is that these methods are easily replicable, quick, and can generate useful insights that would require substantial effort to generate using other forms of review. Below, we discuss our key findings in further detail, as well as some promising directions for expansion of this approach.

Methods for investigating academic corpora

We were impressed by the capacity of our methods to identify trends in subtly differentiated topics. For example, 'temporal patterns' was included in our description of research topics, a finding that reflects current trends in the surrogate ecology literature (Barton et al. 2014). Similarly, we identified fragmentation research as the fastest-growing topic in our corpus (Fig. 1b), a trend that reflects calls for more effective quantification and synthesis of the effects of this process on biodiversity (Ewers et al. 2010). This is encouraging because controversy or inconsistency in terminology can reduce the usefulness of automated approaches such as ours. An example in the ecology literature is the use of identical terminology to mean different things, such as when discussing adaptive management (Westgate et al. 2013) or density dependence (Herrando-Pérez et al. 2012), and these issues probably influenced our case study to some degree. Nonetheless, the fact that several subtle trends were detectable using our approach is highly encouraging for the application of automated methods in conservation biology in the future.

A further application of text analysis is to evaluate hypotheses about different ways that information is communicated and interpreted within research communities. For example, some important findings from our case study were the many relationships between topic similarity, popularity and generality. In particular, hot topics tended to be more specific than cold topics, while clusters of topics that contained similar dominant words differed strongly in popularity. This is potentially concerning, as it could be interpreted as an indicator of publication bias towards narrow concepts. However, there does not seem to be a lack of 'big' ideas in ecology (e.g. McGill 2010), and so a more likely explanation is that new conceptual approaches need to be described in detail before they can be widely understood and adopted. Under this hypothesis, topics become more diffuse throughout the literature with time, meaning that 'hot topics' are those that have high potential, but have yet to be widely adopted. This is supported by the observation that papers describing frequently-used methods are often highly cited (Van Noorden et al. 2014), despite being conceptually narrow. These insights suggest that there is high value in text analysis for elucidating subtle trends in the development of ideas through time.

A key point that we have sought to make in this article is that methods that are commonly used to understand patterns and trends in ecology and conservation can be readily applied to summarizing patterns in research topics identified using LDA (Table 1). This is perhaps most obvious for topic similarity and popularity analysis, which – as mentioned above – are applications of cluster analysis and linear regression, respectively. However, similar analogies exist with generality and gap analysis, as we have defined them in this paper. For example, the approach we use to investigate topic generality is methodologically similar to work on the relationship between abundance and occupancy in ecological communities (Gaston et al. 2000). Further, our identification of research gaps is conceptually similar to the principle of complementarity as applied in spatial prioritization and reserve design (Margules & Pressey 2000), in that it identifies sets of topics that give the greatest cumulative coverage of ideas. Because gap analysis focuses on the relationships between pairs of ideas, the methods that we use to elucidate potential research gaps are also heavily influenced by research into the properties of ecological networks (Ings et al. 2009). Consequently, the concepts we have outlined here should not be unfamiliar to ecologists, albeit in a novel context.

Despite reasons for optimism, a particular difficulty among the methods that we have discussed is deciding which of the research 'gaps' identified by our analysis represent fruitful directions for future research. In fact, some combinations of topics that we identified in our case study (Fig. 1d) may have been avoided by earlier researchers because they are not sensible, rather than by oversight. A further consideration is the potential for the topics identified by gap analysis to refer to areas of strong methodological specialization, in which case researchers' ability to combine insights from these distinct fields of knowledge is likely to be limited. It is worth noting, however, that the practice of combining distinct areas of research is not without precedent as a tool for generating novel insights. A notable example is the maximum entropy formalism, which has broad applications as a statistical inference technique outside of its original field of thermodynamics (Harte 2011). Therefore, while our approach provides a tool to support researchers' insights into the key research fields and trends within their discipline, uncritical use could lead to misguided conclusions (Grimmer & Stewart 2013). Automated text analysis approaches should therefore be used to support or complement (but not replace) detailed evaluation of research options (e.g. Sutherland et al. 2011).

Implications for surrogates and indicators

Our primary goal was to investigate a suite of tools for summarizing the results of a common text mining approach (LDA). However, our case study also led to several discoveries of direct relevance to ecological surrogates and indicators. Of particular importance was our finding that some key research areas have been poorly integrated within the surrogate ecology literature, and these topics therefore represent opportunities for greater collaboration and intellectual development.

The overall message of our case study was the need for more effective tools for biodiversity monitoring in threatened habitats. This is a particularly challenging goal for surrogate ecology, as the efficacy of surrogates for describing variation in other locations, spatial scales, or study taxa has often been limited (Westgate et al. 2014). Fortunately, recent developments show promise for improving this state of affairs. In particular, increased capacity for data sharing is already facilitating assessment of the local-scale impacts of globally-important threatening processes (e.g. Newbold et al. 2015). Further, our gap-analysis showed strong potential for greater use of genetic approaches for quantifying the distribution and trajectory of biodiversity. This integration could be achieved in several ways, but particularly worth noting are studies that incorporate phylogeny into spatial prioritization (Rodrigues et al. 2011), or expanding the use of genetic monitoring methods that use non-invasive sampling (Beja-Pereira et al. 2009). Further development of these tools could to lead to large improvements in our capacity to monitor and manage landscapes for conservation in future.

How corpora are selected for text analysis will have a fundamental influence on the patterns detected by methods such as ours. This may account for the observation that several important research areas from the wider ecology literature – including forestry and species distribution modeling – appear to be declining in popularity within our corpus (Fig. 1b). The use of article abstracts for text analysis has also been criticized for overly limiting the amount of information available to text summary algorithms (Boyack et al. 2013). This may explain the large number of topics in our case study that referred to goals or methods (Fig. 1a), which are likely to be proportionally overrepresented in article abstracts versus full text. Finally, our analysis is only intended as an example of the kinds of results that can be achieved by comparatively simple methods. More rigorous testing would be needed if these methods were intended to guide detailed research synthesis and forecasting. Consequently, potential users of these methods should consider the suitability of text analysis for investigating their particular questions and field of interest.

Finally, we observed that research on the ecology of agricultural environments was a fast-growing topic in our corpus (Fig. 1b), and that gap analysis suggested a high priority for research on their monitoring and management (Fig. 1d). Assessment of the biodiversity value of agricultural regions has become particularly important with the introduction of environmental stewardship programs in a number of countries (Lindenmayer et al. 2012; Scheper et al. 2013). Further, understanding how these systems function has become increasingly important with the introduction of incentive schemes based on carbon sequestration, clean water provision, or pollination services (Whittingham 2011). Consequently, the results of our model-based approach reflect a known shift in ecological science towards understanding and valuing conservation opportunities in non-pristine ecosystems (Mace 2014).

In conclusion, we have demonstrated that a combination of readily available and conceptually straightforward methods can be used to identify meaningful topics within academic corpora. This includes classification of their popularity and generality, as well as identification of rarely-studied combinations of topics that represent gaps in research effort. These insights suggest that greater use of text analysis for ecological synthesis is warranted. Moreover, several methods for aiding the interpretation of results from text mining algorithms are already in common use within the ecology and conservation literature. We argue, therefore, that there are few barriers to further application of text analysis to the ecology and conservation literature, and that this could benefit conservation science.

Acknowledgements

This research was supported by an ARC Laureate Fellowship to D.B.L. Our use of LDA in this article was inspired by work by C. Sievert. P. Lane and J. Wood provided feedback that improved earlier versions of this manuscript.

Supporting information

Further methods and topic keywords (Appendix S1) are available online.

Literature Cited

Barton PS, Westgate MJ, Lane PW, MacGregor C, and Lindenmayer DB. 2014. Robustness of habitat-based surrogates of animal diversity: A multi-taxa comparison over time and after fire. Journal of Applied Ecology **51**:1434-1443.

Bates D, Maechler M, Bolker B, and Walker S. 2014. Lme4: Linear mixed-effects models using eigen and s4. Pages R package version 1.1-6.

Beja-Pereira A, Oliveira R, Alves PC, Schwartz MK, and Luikart G. 2009. Advancing ecological understandings through technological transformations in noninvasive genetics. Molecular Ecology Resources **9**:1279-1301.

Bennett EM, Cumming GS, and Peterson GD. 2005. A systems model approach to determining resilience surrogates for case studies. Ecosystems **8**:945-957.

Blei DM, Ng AY, and Jordan MI. 2003. Latent dirichlet allocation. Journal of Machine Learning Research **3**:993-1022.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, and White J-SS. 2009. Generalized linear mixed models: A practical guide for ecology and evolution. Trends in Ecology & Evolution **24**:127-135.

Boyack KW, and Klavans R. 2014. Creation of a highly detailed, dynamic, global model and map of science. Journal of the Association for Information Science and Technology **65**:670-685.

Boyack KW, Small H, and Klavans R. 2013. Improving the accuracy of co-citation clustering using full text. Journal of the American Society for Information Science and Technology **64**:1759-1767.

Chen C, Chen Y, Horowitz M, Hou H, Liu Z, and Pellegrino D. 2009. Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics **3**:191-209.

Collen B, and Nicholson E. 2014. Taking the measure of change. Science **346**:166-167.

De Cáceres M, Legendre P, and Moretti M. 2010. Improving indicator species analysis by combining groups of sites. Oikos **119**:1674-1684.

Ewers RM, Marsh CJ, and Wearn OR. 2010. Making statistics biologically relevant in fragmented landscapes. Trends in Ecology & Evolution **25**:699-704.

Gaston KJ, Blackburn TM, Greenwood JJD, Gregory RD, Quinn RM, and Lawton JH. 2000. Abundance–occupancy relationships. Journal of Applied Ecology **37**:39-59.

Griffiths TL, and Steyvers M. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences (USA) **101**:5228-5235.

Grimmer J, and Stewart BM. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis:1-31.

Gruen B, and Hornik K. 2011. Topicmodels: An r package for fitting topic models. Journal of Statistical Software **40**:1-30.

Harte J 2011. Maximum entropy and ecology: A theory of abundance, distribution, and energetics. Oxford University Press.

Herrando-Pérez S, Delean S, Brook B, and Bradshaw CA. 2012. Density dependence: An ecological tower of babel. Oecologia **170**:585-603.

Ings TC, et al. 2009. Review: Ecological networks – beyond food webs. Journal of Animal Ecology **78**:253-269.

Larsen P, and von Ins M. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics **84**:575-603.

Legendre P, and Legendre LFJ 2012. Numerical ecology. Elsevier.

Lindenmayer D, and Likens G. 2011. Direct measurement versus surrogate indicator species for evaluating environmental change and biodiversity loss. Ecosystems **14**:47-59.

Lindenmayer DB, Barton PS, Lane PW, Westgate MJ, McBurney L, Blair D, Gibbons P, and Likens GE. 2014. An empirical assessment and comparison of species-based and habitat-based surrogates: A case study of forest vertebrates and large old trees. PLoS ONE **9**:e89807.

Lindenmayer DB, Zammit C, Attwood SJ, Burns E, Shepherd CL, Kay G, and Wood J. 2012. A novel and cost-effective monitoring approach for outcomes in an australian biodiversity conservation incentive program. PLoS ONE **7**:e50872.

Lortie CJ. 2014. Formalized synthesis opportunities for ecology: Systematic reviews and meta-analyses. Oikos **123**:897-902.

Mace GM. 2014. Whose conservation? Science **345**:1558-1560.

Margules CR, and Pressey RL. 2000. Systematic conservation planning. Nature **405**:243-253.

McGeogh MA. 1998. The selection, testing and application of terrestrial insects as bioindicators. Biological Reviews **73**:181-201.

McGill BJ. 2010. Towards a unification of unified theories of biodiversity. Ecology Letters **13**:627-642.

Newbold T, et al. 2015. Global effects of land use on local terrestrial biodiversity. Nature **520**:45-50.

Niebles J, Wang H, and Fei-Fei L. 2008. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**:299-318.

Noss RF. 1990. Indicators for monitoring biodiversity: A hierarchical approach. Conservation Biology **4**:355-364.

Pollock LJ, Morris WK, and Vesk PA. 2012. The role of functional traits in species distributions revealed through a hierarchical model. Ecography **35**:716-725.

R Core Development Team 2014. R: A language and environment for statistical computing, version 3.1.0. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues ASL, et al. 2011. Complete, accurate, mammalian phylogenies aid conservation planning, but not much. Philosophical Transactions of the Royal Society B-Biological Sciences **366**:2652-2660.

Rusch T, Hofmarcher P, Hatzinger R, and Hornik K. 2013. Model trees with topic model preprocessing: An approach for data journalism illustrated with the wikileaks afghanistan war logs. The Annals of Applied Statistics **7**:613-639.

Scheper J, Holzschuh A, Kuussaari M, Potts SG, Rundlöf M, Smith HG, and Kleijn D. 2013. Environmental factors driving the effectiveness of european agri-environmental measures in mitigating pollinator loss – a meta-analysis. Ecology Letters **16**:912-920.

Silva C, and Ribeiro B. 2003. The importance of stop word removal on recall values in text categorization. Pages 1661-1666 vol.1663. Neural Networks, 2003. Proceedings of the International Joint Conference on.

Soulé ME. 1985. What is conservation biology? BioScience **35**:727-734.

Sutherland WJ, et al. 2009. One hundred questions of importance to the conservation of global biological diversity. Conservation Biology **23**:557-567.

Sutherland WJ, Fleishman E, Mascia MB, Pretty J, and Rudd MA. 2011. Methods for collaboratively identifying research priorities and emerging issues in science and policy. Methods in Ecology and Evolution **2**:238-247.

Valle D, Baiser B, Woodall CW, and Chazdon R. 2014. Decomposing biodiversity data using the latent dirichlet allocation model, a probabilistic multivariate statistical method. Ecology Letters **17**:1591-1601.

Van Noorden R, Maher B, and Nuzzo R. 2014. The top 100 papers: Nature explores the most-cited research of all time. Nature **514**:550-553.

Wang D, Song C, and Barabási A-L. 2013. Quantifying long-term scientific impact. Science **342**:127-132.

Wang Y, Naumann U, Wright ST, and Warton DI. 2012. Mvabund– an r package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution **3**:471-474.

Weng J, Lim E-P, Jiang J, and He Q. 2010. Twitterrank: Finding topic-sensitive influential twitterers. Pages 261-270. Proceedings of the third ACM international conference on Web search and data mining. ACM, New York, New York, USA.

Westgate MJ, Barton PS, Lane PW, and Lindenmayer DB. 2014. Global meta-analysis reveals low consistency of biodiversity congruence relationships. Nature Communications **5**:3899.

Westgate MJ, Likens GE, and Lindenmayer DB. 2013. Adaptive management of biological systems: A review. Biological Conservation **158**:128-139.

Whittingham MJ. 2011. The future of agri-environment schemes: Biodiversity gains and ecosystem service delivery? Journal of Applied Ecology **48**:509-513.