Conservation Biology

Essay

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

MARTIN J. WESTGATE,* PHILIP S. BARTON, JENNIFER C. PIERSON, AND DAVID B. LINDENMAYER

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

Abstract: *Keeping track of conceptual and methodological developments is a critical skill for research scientists, but this task is 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. We show how a common text-mining method (latent Dirichlet allocation, or topic modeling) and statistical tests familiar to ecologists (cluster analysis, regression, and network analysis) can be used to investigate trends and identify potential research gaps in the scientific literature. We tested these methods on the literature on ecological surrogates and indicators. Analysis of topic popularity within this corpus showed a strong emphasis on monitoring and management of fragmented ecosystems, while analysis of research gaps suggested 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, increasing scientists' capacity for research synthesis.*

Keywords: hot topics, indicators, latent Dirichlet allocation, synthesis, surrogates

Herramientas de Análisis de Texto para la Identificación de Temas Emergentes y Vacíos en la Investigación de la Ciencia de la Conservación

Resumen: *Dar seguimiento a los desarrollos conceptuales y metodológicos es una habilidad crítica para los investigadores, pero es una tarea cada vez más difícil debido a la tasa alta de publicaciones académicas. Como una disciplina de crisis, la ciencia de la conservación está particularmente necesitada de herramientas que faciliten una síntesis rápida y a la vez profunda. Mostramos cómo un método común de búsqueda y procesamiento de textos (asignación Dirichlet latente, o modelado de temas) y las pruebas estadísticas conocidas por los ecólogos pueden ser usados para investigar tendencias e identificar vacíos potenciales de la investigación en la literatura científica. Probamos estos métodos en la literatura sobre los sustitutos y los indicadores ecológicos. El análisis de la popularidad de temas dentro de este recopilatorio mostró un fuerte énfasis en el monitoreo y manejo de los ecosistemas fragmentados, mientras que el análisis de los vacíos en la investigación sugirió un papel mayor para los sustitutos e indicadores genéticos. Nuestros resultados muestran que los métodos de análisis de texto necesitan ser usados con cuidado pero también que pueden proporcionar información que es complementaria a aquella dada por las revisiones sistemáticas y los meta-análisis, lo que incrementa la capacidad de los científicos para sintetizar la investigación.*

Palabras Clave: asignación Dirichlet latente, indicadores, síntesis, sustitutos, temas sobresalientes

Introduction

The ability to understand historical and emerging ideas is a critical skill for research scientists, as is the capacity 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 conservation

*email martin.westgate@anu.edu.au
Paper submitted December 11, 2014; revised manuscript accepted July 23, 2015.

DOI: 10.1111/cobi.12605

scientists because their findings have direct implications for evidence-based conservation (Sutherland et al. 2009). 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). Several of these approaches identify combinations of words within articles (i.e., text analysis) and so help 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 greater speed and quantitative rigor (Grimmer & Stewart 2013). However, text analysis is rarely used in ecology and conservation, which is unusual given its usefulness for understanding research trajectories and given the ongoing trend in ecology and conservation toward greater quantification (Lortie 2014).

We believe conservation science is well placed to capitalize on text-analysis tools because methods for summarizing the results of text-mining algorithms are similar (or even identical) to existing and commonly used ecological methods (Table 1). We considered how a combination of approaches could be used to understand patterns and trends within academic corpora by examining the literature on ecological surrogates and indicators as a case study. This field is particularly suited to text analysis because it is a large and diverse body of work that has recently grown dramatically (Westgate et al. 2014), thereby presenting a considerable challenge to quantitative synthesis. Surrogates are also important from a conservation perspective because they provide the data underpinning nearly all conservation decisions (Collen & Nicholson 2014), and so that improved understanding and application of surrogates should lead to more efficient ecosystem monitoring and management. Therefore, we addressed critical barriers to the wider adoption of text analysis by considering how complex topics can be synthesized to facilitate informed decisions regarding research priorities.

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 (i.e., topics) that can be used to interpret patterns in the corpus. With latent Dirichlet allocation (LDA or topic modeling [Blei et al. 2003]), topics are defined using sets of words that co-occur with unusual frequency, and so 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 fields such as journalism (Rusch et al. 2013), politics (Grimmer & Stewart 2013), and social network analysis (Weng et al. 2010).

Although text analysis is rare in ecology and conservation, there are several close parallels between LDA and existing ecological modeling approaches. First, the popularity of LDA as a research tool reflects a shift toward model-based multivariate analysis that is also evident in ecology (Wang et al. 2012). Second, just as methods that are common in ecology and conservation (such as ordination [Legendre & Legendre 2012]) can be used to identify associated words within texts, LDA can be applied to ecological problems such as classification of image time series (Niebles et al. 2008) or 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 (i.e., stop words, e.g., *the* or *and*) are typically removed during text analysis because they provide little information content (Silva & Ribeiro 2003). As these parallels make clear, methods for text analysis are strongly related to those used in ecology and conservation.

Although LDA is not the only text-classification method, here we assumed the use of LDA for topic identification. We examined 4 methods (analysis of topic similarity, generality, popularity, and research gaps) that build on one another to provide complementary forms of information regarding the content of study corpora (Table 1). These methods facilitate interpretation of the content provided by LDA and cannot be applied in isolation of a method for topic identification.

Topic Similarity

A key problem with 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

Table 1. Methods for examining content in academic corpora (using topics identified with latent Dirichlet allocation), and their analogues in ecological modeling.

Statistical approach	*Text analysis*	*Ecological modeling*
Cluster analysis	identify clusters of similar topics based on the words they contain (Blei et al. 2003)	identify clusters of similar locations based on the species they contain (Legendre & Legendre 2012)
Comparison of frequency distributions	investigate relationship between the number of articles assigned to each topic and the weight of that topic within each article	investigate 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 extent to which pairs of topics tend to occur in similar vs. different texts	quantify strength of associations between pairs of species or individuals (Ings et al. 2009)

to standard dissimilarity and ordination-based methods (Legendre & Legendre 2012). The value of this method is partly in validation, for example, in determining whether topics that contain similar words appear similar to the user based on their understanding of the corpus under investigation. It also provides information critical to the interpretation of other trends, such as whether similar topics differ in popularity.

Popularity, Growth, and Hot Topics

Determining which topics are most popular can be achieved either by calculating the total number of articles that have been published on a topic over a particular period or by investigating changes in topic popularity over that period. The former provides information on total research effort within a corpus, while the latter is commonly used to assess which topics are hot (i.e., show positive growth) versus cold (negative growth) within a given research community (Griffiths & Steyvers 2004).

As for topic similarity, methods to assess topic popularity are also familiar to ecologists, namely linear regression. In its simplest form, topic popularity can be investigated by quantifying the change over time (predictor variable x) in the number of published articles per topic (response variable y). A useful way to do this is to separate article counts by topic and then use mixed models (Bolker et al. 2009) to fit a unique intercept (i.e., mean number of publications if the predictor variable is centered) 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

$$\log(E_{(y|u)}) = \alpha + (\beta + b)x + u, \quad (1)$$

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_u^2, σ_b^2. In such a model, topics with positive random intercepts (i.e., $u > 0$) can be interpreted as having higher-than-average numbers of articles written about them in the period. Similarly, topics with positive random slopes have higher-than-average growth in publications during the same period.

Specificity and Generality

So far we have discussed LDA as a data-mining exercise, but it is an oversimplification to assume that all topics are directly comparable. A particular problem is that because topics are identified according to sets of co-occurring words, 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 is useful to be able to calculate some measure of where each topic sits on the spectrum from general to specific.

To assess topic specificity versus generality one can 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., the maximum for that row). This approach is sensible if one topic receives a much higher weight for a given article than all the remaining topics, but is problematic if all topics have 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 is restricted to a subset of articles (specificity).

Identifying Research Directions

One goal of a literature review may be to identify future research directions. Although a common and even necessary part of literature review, the idea of automating the process of predicting future directions may be alarming to some. Certainly, there are inherent difficulties and ambiguities in this form of prediction. Nonetheless, 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 have distinct quantitative attributes, suggesting that scientific impact can be quantified and therefore predicted. A more useful observation for our purposes would be a theory on 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. This theory does not preclude the possibility that progress might also occur through spontaneous novel insights, but such insights are inherently less amenable to prediction and so can be ignored for our purposes. Here, we refer to our investigation of these phenomena as research gap analysis.

Ecological Surrogates and Indicators

To demonstrate how the methods outlined above can be used in practice, we applied them to a corpus of article titles and abstracts from the scientific literature on ecological surrogates and indicators. Although this is an area of strong research interest to us, the insights we ascertained derived exclusively from the text analysis methods described 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 draw from their results.

Surrogates and indicators are proxies used to draw inference regarding complex ecosystems from a manageable amount of data; thus, they are critical to environmental management (Noss 1990; Collen & Nicholson 2014). This body of literature also presents a test of text-analysis methods due to its large 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 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 with the goal of synthesizing knowledge across the full range of methods and applications in surrogate ecology (McGeogh 1998). To examine this corpus, we used LDA combined with the methods described above to quantitatively investigate topic similarity, popularity, generality, and potential research directions.

Methods

We examined the abstracts of articles that cited a single seminal work on ecological surrogates and indicators (Noss 1990; $n = 1160$) and articles that cited any of the 100 most highly cited articles that cited Noss (1990) (i.e., the second-generation citations; $n = 8674$). We first identified 25 topics within this corpus with an LDA model fitted using the topicmodels package (Gruen & Hornik 2011) in R (R Core Development Team 2014) and named each topic based on our assessment of the 20 highest weighted keywords for that topic (Supporting Information).

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 and topic combination. We used mixed models to investigate topic popularity (eq. 1) as implemented in lme4 (Bates et al. 2014) and calculated topic generality with information on the weight assigned to each article and topic combination (M2; the associated distance matrix was named D2). Research gap analysis involved calculating the product of D1 and D2 after scaling each matrix to a range of 0–1. A detailed description of our methods is available in Supporting Information.

Results

Clustering of topic content via word-based similarity divided our data set 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). The second group contained topics describing subthemes within the spatial ecology literature (including spatial prioritization and fragmentation). The final group contained topics describing basic concepts in community ecology such as threatened fauna, assemblage structure, and common predictors of change. Although these clusters described meaningful patterns in the data set, 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

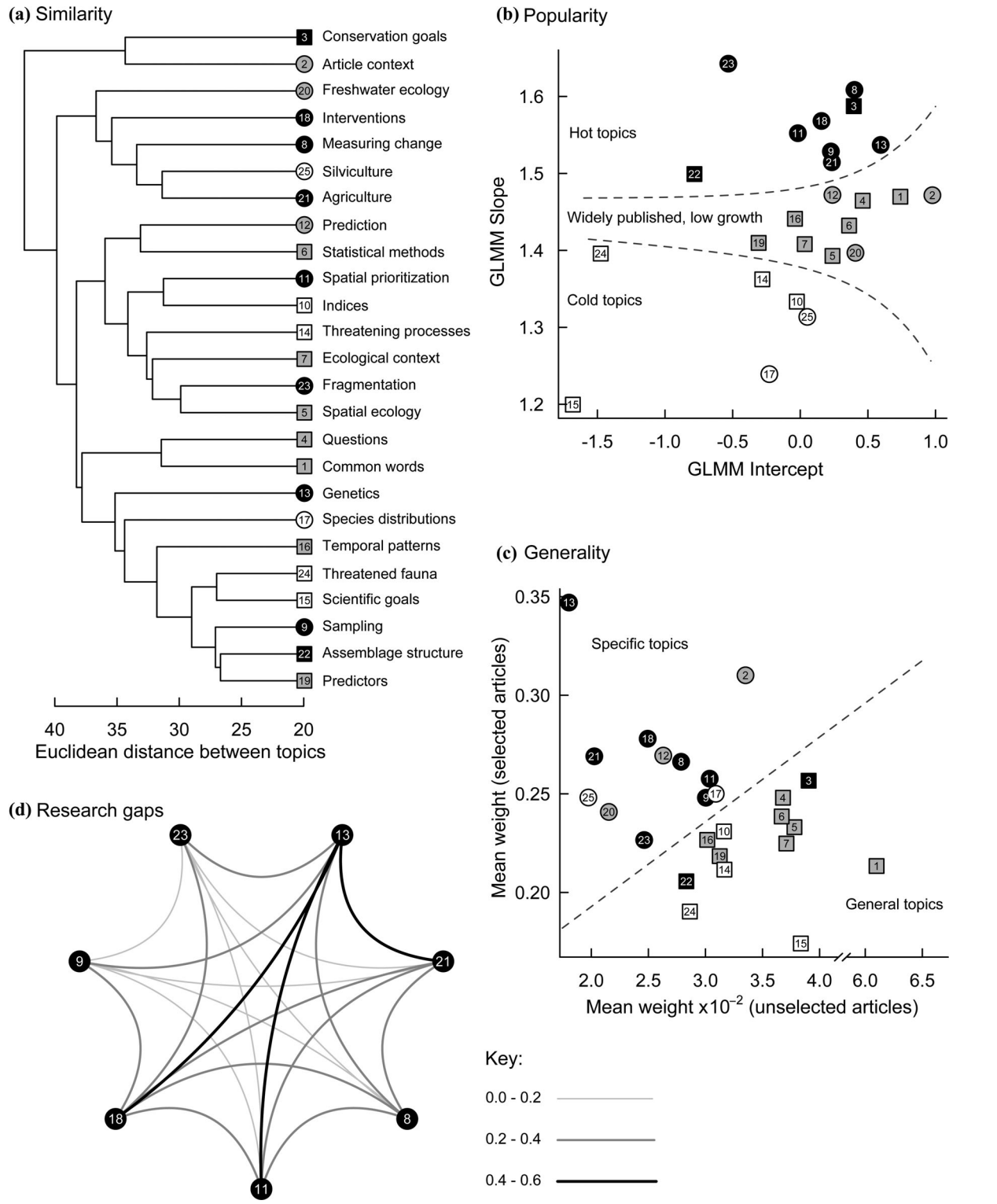

Figure 1. (a) Similarity, (b) popularity, (c) generality, and (d) research gap distance for topics in the academic literature on surrogates and indicators identified with latent Dirichlet allocation (LDA). See text and supporting information for details of all calculations. Point numbers show the order of topics as returned by LDA. In all panels, shadings are designated according to their category in panel (b), and point shapes are designated according to their category in panel (c). Categories represent coarse groupings defined for example purposes only and should not be considered statistically robust. Key refers to research gap distances in panel (d).

(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 showed that historically popular topics had intermediate growth (Fig. 1b). Fragmentation research had the highest growth rate of any topic, which was one of a number of unexpected patterns. In particular, silvicultural research decreased in popularity (relative to the mean), despite increases in the conceptually similar field of agriculture. Several topics decreased relative to our anecdotal assessment of their frequency in the broader ecology literature, namely freshwater ecology, spatial priortization, and threatening processes (which included keywords related to urbanization and climate change [Supporting Information]). Taken as a group, hot topics focused on ways to measure and ameliorate significant threatening processes (e.g., agriculture, fragmentation, interventions), whereas there was no clear association between cold topics.

Topic generality analysis (Fig. 1c) allowed us to distinguish between topics that had high weight in a subset of articles but low weight elsewhere (specific) and 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 the topics listed as questions and common words in our analysis. In contrast, genetics was the most specific topic (Fig. 1c). Comparison of results from popularity and generality analyses showed that 9 of the 14 low-growth topics identified by popularity analysis (Fig. 1b) were classified as general (Fig. 1c), while only 2 of the 9 hot topics were classified as general. Comparing these findings with results of 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.

Analysis of research gaps 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), and 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 are underused 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

We found that 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. Our key message is that these methods are easily replicable, can be executed quickly, and can generate useful insights that would require substantial effort to generate with other forms of review.

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). These issues probably influenced our case study to some degree. Nonetheless, that several subtle trends were detectable using our approach is highly encouraging for the application of automated methods in conservation biology.

A further application of text analysis is to evaluate hypotheses about different ways information is communicated and interpreted within research communities. For example, some important findings from our case study were the many relationships among 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 problematic because it could be interpreted as an indicator of publication bias toward narrow concepts. However, there does not seem to be a lack of big ideas in ecology (e.g., McGill 2010), 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 have high potential but are 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. This finding suggests there

is high value in text analysis for elucidating subtle trends in the development of ideas through time.

Our key insight is that methods that are commonly used to understand patterns and trends in ecology and conservation can be readily used to summarize patterns in research topics identified using LDA (Table 1). This is perhaps most obvious for topic similarity and popularity analysis, which are applications of cluster analysis and linear regression, respectively. However, similar analogies exist with generality and research gap analysis, as we have defined them here. For example, the approach we used 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 research gap analysis focuses on the relationships between pairs of ideas, the method we used to identify gaps is 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 we considered is deciding which of the research gaps we identified represent fruitful directions for future research. Some combinations of topics we identified (Fig. 1d) may have been avoided by earlier researchers because they were not sensible, rather than because they were overlooked. A further consideration is the potential for the topics identified by research 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

In our investigation of LDA, we also made several discoveries of direct relevance to ecological surrogates and indicators. In particular, we found 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.

Through our case study, we identified a need for more effective tools for biodiversity monitoring in threatened habitats. This is a particularly challenging goal for surrogate ecology because the efficacy of surrogates for describing processes 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. 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, research 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) and expanding the use of genetic monitoring methods that use noninvasive 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.

A key caveat of our approach is that the choice of method for selecting study corpora will fundamentally affect the patterns detected by the methods that we have used. This may account for our observation that several important research areas from the wider ecology literature—including forestry and species distribution modeling—appeared to be declining in popularity (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.

Finally, we observed that research on the ecology of agricultural environments was a fast-growing topic (Fig. 1b), and our 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 some countries (Lindenmayer et al. 2012; Scheper et al. 2013). 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 toward understanding and valuing conservation opportunities in non-pristine ecosystems (Mace 2014).

We found 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.

Acknowledgments

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. The authors are solely responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Barton PS, Westgate MJ, Lane PW, MacGregor C, 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, Walker S. 2014. Lme4: linear mixed-effects models using Eigen and s4. R package version **1:**1–6.

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

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

Blei DM, Ng AY, 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, White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution **24:**127–135.

Boyack KW, 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, 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, Pellegrino D. 2009. Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics **3:**191–209.

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

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

Ewers RM, Marsh CJ, 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, Lawton JH. 2000. Abundance–occupancy relationships. Journal of Applied Ecology **37:**39–59.

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

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

Gruen B, 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, Oxford.

Herrando-Pérez S, Delean S, Brook B, 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, 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, Legendre LFJ. 2012. Numerical ecology, 3rd English edn. Elsevier, Amsterdam.

Lindenmayer D, 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, 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, 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, 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, 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, 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.

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, 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, 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, Ribeiro B. 2003. The importance of stop word removal on recall values in text categorization. Proceedings of the International Joint Conference on Neural Networks, 2003 **3:**1661–1666.

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, 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, 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, Nuzzo R. 2014. The top 100 papers: nature explores the most-cited research of all time. Nature **514:**550–553.

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

Wang Y, Naumann U, Wright ST, 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, 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.

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

Westgate MJ, Likens GE, 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.