Do synthesis centers produce novel, potentially transformative research?

Research publication diversity as an indicator of novelty and transformative capacity

Stacy Rebich-Hespanha[1], Ismael Rafols[2], John N. Parker[3], Edward J. Hackett[3], Joao P. Hespanha[4], The Sensible Science Working Group, Assessing Synthesis and Synthesis Centers[1,5]

1. National Center for Ecological Analysis and Synthesis (NCEAS), UC Santa Barbara, CA
2. Ingenio (CSIC-UPV), Polytechnic University of València, Spain
3. Arizona State University, Tempe, AZ
4. Dept. of Electrical and Computer Engineering, UC Santa Barbara, CA
5. National Evolutionary Synthesis Center (NESCent), Duke University, UNC Chapel Hill, NC State University

evaluating synthetic publications

Synthesis is an emerging method for producing transformative research, and centers to promote synthesis are on the rise in the US and around the world. New analytic tools and techniques are needed to assess the originality and transformative potential of synthesis. We propose that research outputs produced within synthesis centers will exhibit distinctive qualities that distinguish them from other publications in their fields. To explore this possibility, we conducted a topical analysis of titles, abstracts, and keywords for approximately 400,000 articles published in 108 leading journals from the fields of Ecology, Evolutionary Biology, Biodiversity Conservation, Forestry, and Fisheries. We then described each document as a proportional combination of the discovered topics, and used the Rao-Stirling heuristics to estimate, for each document, various measures that illuminate contrasting aspects of diversity (i.e. variety, balance, and disparity). We then compare diversity metrics for the synthesis center documents with those for all other documents in our corpus to evaluate whether and how the measured diversity of synthesis center publications differs from that of other publications in the relevant fields.

data sources

Ecology
Biodiversity Conservation
Evolutionary Biology
Fisheries
Forestry

ISI journal abstracts (N = 398,378)

Data included all abstract-containing records published between 1996 and late August 2013 for four top general science journals (*Nature*, *PLOS ONE*, *PNAS*, and *Science*) and the top 20 journals (based on Eigenfactor[2]) for each of five ISI subject categories (94 journals total).

An additional 14 journals that were key publication venues for the synthesis centers (marked with * in the journal list to right) were added to the corpus to allow for inclusion of a greater proportion of synthesis center documents. The corpus contained 1668 (70%) of the 2390 documents published during that period based on work completed at the two synthesis centers (NCEAS = 1298 of 1837, NESCent = 370 of 553).

Natural language fields from these records (title, abstract, and keywords) were used to generate the LDA topic model.

Agric. For. Meteorol.
Agrofor. Syst.
Am. J. Bot.*
Am. J. Phys. Anthropol.
Am. Midl. Nat.
Am. Mus. Novit.
Am. Nat.
Anim. Conserv.
Ann. For. Sci.
Annu. Rev. Ecol. Evol. Syst.
Annu. Rev. Ecol. Syst.
Appl. Veg. Sci.
Aquac. Nutr.
Aquac. Res.
Aquaculture
Behav. Ecol.*
Biodivers. Conserv.
Bioinformatics*
Biol. Conserv.
Biol. Invasions
Biol. J. Linnean Soc.
Biol. Lett.
Bioscience*
BMC Evol. Biol.
Bull. Amer. Mus. Nat. Hist.
Can. J. Fish. Aquat. Sci.
Can. J. For. Res.-Rev. Can. Rech. For.
Conserv. Biol.
Conserv. Genet.
Conserv. Genet. Resour.
Dis. Aquat. Org.
Divers. Distrib.
Ecography
Ecol. Appl.
Ecol. Econ.*
Ecol. Freshw. Fish
Ecol. Lett.
Ecol. Model.
Ecology
Ecosystems*
Eur. J. For. Res.
Evolution
Fish Oceanogr.
Fish Res.
Fish Shellfish Immunol.
Fish. Fish.
Fish. Sci.
Fisheries Manag. Ecol.
For. Ecol. Manage.
For. Sci.
Forest Policy Econ.
Forestry
Front. Ecol. Environ.*
Genetics*
Glob. Change Biol.
Heredity
Holzforschung
ICES J. Mar. Sci.
Int. J. Plant Sci.*
Int. J. Wildland Fire
Integr. Comp. Biol.*
J. Anim. Ecol.*
J. Appl. Ecol.
J. Appl. Ichthyol.
J. Appl. Ichthyol.-Z. Angew. Ichthyol.
J. Biogeogr.
J. Ecol.
J. Evol. Biol.
J. Fish Biol.
J. Fish Dis.
J. For.
J. Hum. Evol.
J. Mol. Evol.
J. Nat. Conserv.
J. Nat. Hist.
J. Shellfish Res.
J. Veg. Sci.
J. Vertebr. Paleontol.*
Mar. Ecol.-Prog. Ser.
Mar. Freshw. Res.
Mol. Biol. Evol.
Mol. Ecol.
Mol. Ecol. Resour.
Mol. Phylogenet. Evol.
North Am. J. Fish Manage.
Northeast. Nat
Oecologia
Oikos
Paleobiology
Plant Ecol.
PLoS Genet.*
Polar Biol.
Proc. R. Soc. B-Biol. Sci.
Proc. R. Soc. Lond. Ser. B-Biol. Sci.
Restor. Ecol.*
Rev. Chil. Hist. Nat.
Silva. Fenn.
Southw. Natural.
Syst. Biodivers.
Syst. Biol.
Taxon
Theor. Popul. Biol.
Trans. Am. Fish. Soc.
Tree Genet. Genomes
Tree Physiol.
Trees-Struct. Funct.
Trends Ecol. Evol.
Wood Sci. Technol.

LDA topic modeling

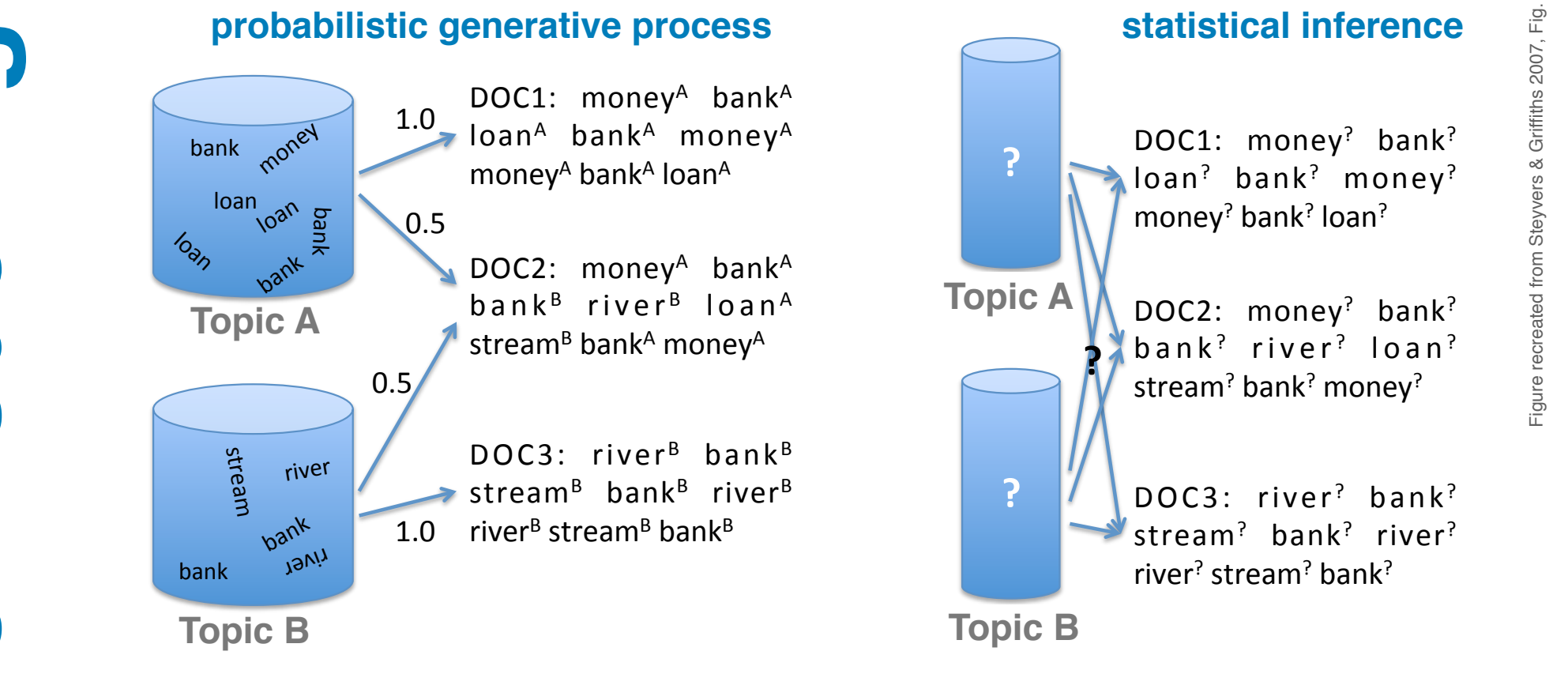
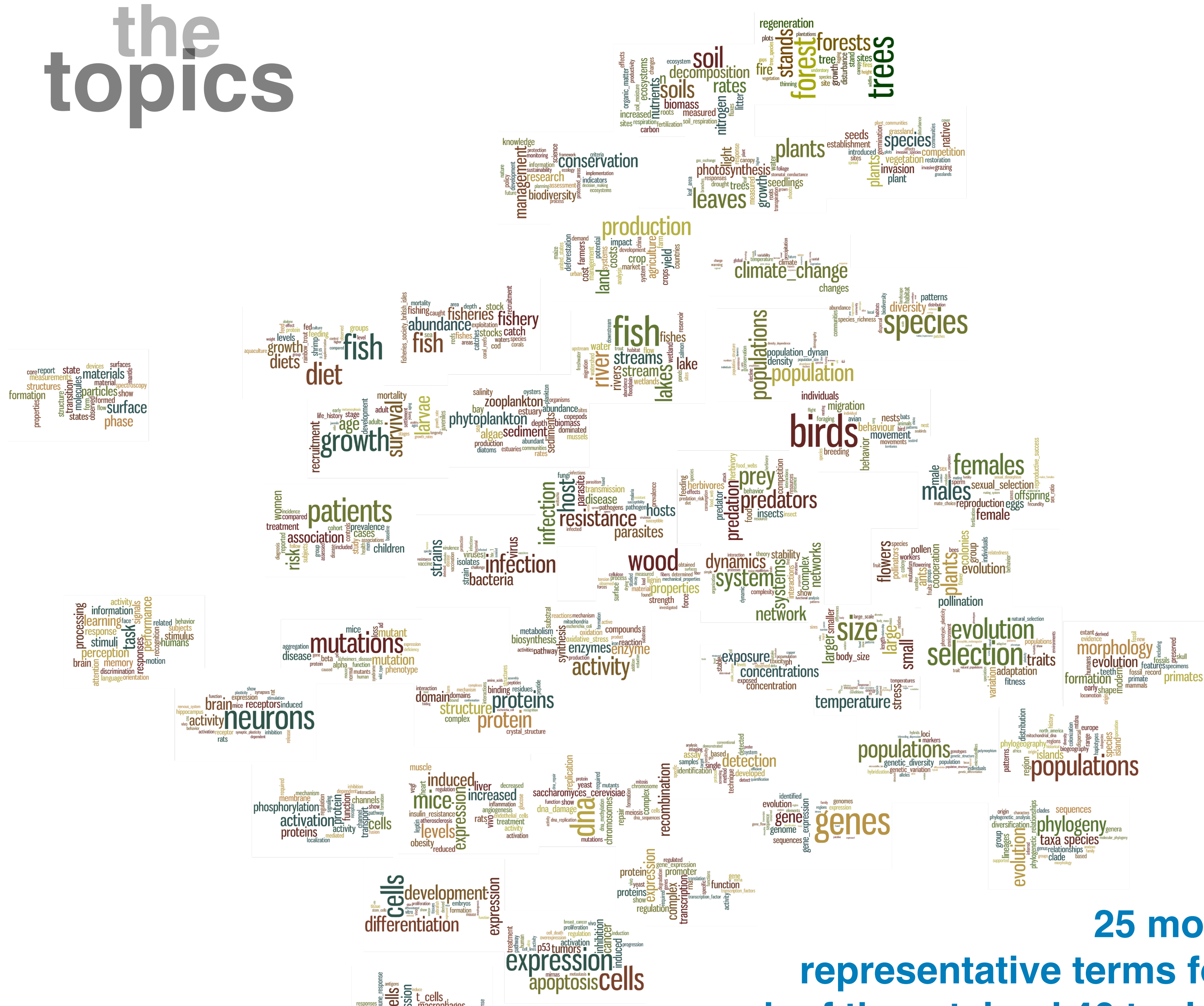

probabilistic generative process
statistical inference
Topic A
Topic B
bank
money
loan
river
stream
1.0
0.5
0.5
1.0
DOC1: money^A bank^A loan^A bank^A money^A money^A bank^A loan^A
DOC2: money^A bank^A bank^B river^B loan^A stream^B bank^A money^A
DOC3: river^B bank^B stream^B bank^B river^B river^B stream^B bank^B
?
DOC1: money^? bank^? loan^? bank^? money^? money^? bank^? loan^?
DOC2: money^? bank^? bank^? river^? loan^? stream^? bank^? money^?
DOC3: river^? bank^? stream^? bank^? river^? river^? stream^? bank^?

Figure recreated from Steyvers & Griffiths 2007, Fig. 2

Topic modeling, in this case Latent Dirichlet Allocation (LDA)[1], is an unsupervised probabilistic method for extracting a quantitative representation of semantic content for a document corpus based on observed patterns of term co-occurence. The LDA model assumes Dirichlet priors for terms in topics and and topics in documents, and discovers latent 'topics' by repeated sampling across the entire corpus. Once the latent topics (represented by groups of co-occurring terms) have been discovered, per-document topic 'mixtures' or combinations are inferred.

1. Steyvers, M. and T. Griffiths (2007) Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis and W. Kintsch (Eds.) *Handbook of Latent Sematic Analysis*, pp. 427-448. Earlbaum, Mahwah, NJ.
2. Kohonen, T., et al. (1996) SOM_PAK: the Self-Organizing Map program package. Helsinki University of Technology. http://www.cis.hut.fi/research/som_lvq_pak.shtml

the topics



topics removed

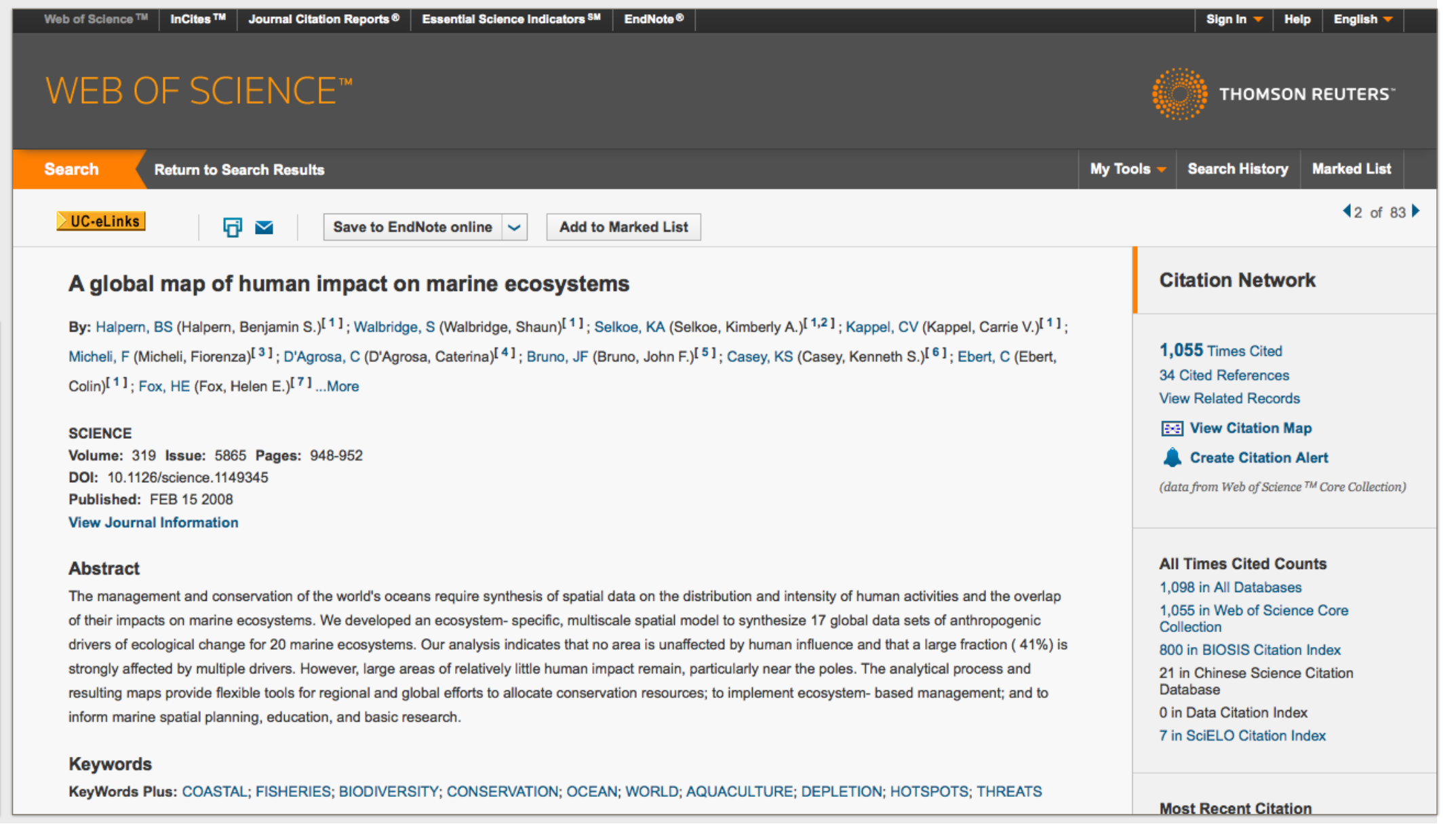
25 most representative terms for each of the retained 46 topics

The documents in the corpus were organized based on topic similarity using a self-organizing map algorithm[2] and then grouped into clusters based on their most dominant topics. Topic cluster positions were then manually adjusted slightly based on k-means clustering solutions. Topic clusters are organized so that similar topics are near each other. Within each topic cluster, the largest words are the most representative terms for that topic.

documents as topic mixtures

This table shows the inferred LDA topic mixture for one of the synthesis publications. The Web of Science screenshot on the right shows the title, abstract and keywords that were used by the topic model to make this inference. The first column of the table shows the topic ID for the topics associated with this document; the second shows the estimated weights for each of those topics in the document; and the third column shows the most representative terms for each of these topics.

TopicID	Weight	Representative Terms
topic 02	0.3529	conservation management research biodiversity knowledge scienc
topic 03	0.1569	species diversity patterns habitat species_richness communities ab
topic 18	0.1373	phytoplankton sediment zooplankton algae sediments bay abunda
topic 16	0.0392	fish fishery abundance fisheries catch stocks fishing stock fishes ca
topic 31	0.0392	dna recombination saccharomyces_cerevisiae chromosomes replic
topic 39	0.0392	mutations mutation mutant disease mice phenotype loss alpha be
topic 17	0.0196	populations islands species island region phylogeography distribut
topic 21	0.0196	climate_change changes climate temperature precipitation global
topic 26	0.0196	fish river lakes streams stream lake rivers fishes water wetlands w
topic 27	0.0196	selection evolution traits adaptation variation fitness populations



A global map of human impact on marine ecosystems

By: Halpern, BS (Halpern, Benjamin S.)[1]; Walbridge, S (Walbridge, Shaun)[1]; Selkoe, KA (Selkoe, Kimberly A.)[1,2]; Kappel, CV (Kappel, Carrie V.)[1]; Micheli, F (Micheli, Fiorenza)[3]; D'Agrosa, C (D'Agrosa, Caterina)[4]; Bruno, JF (Bruno, John F.)[5]; Casey, KS (Casey, Kenneth S.)[6]; Ebert, C (Ebert, Colin)[1]; Fox, HE (Fox, Helen E.)[7] ...More

SCIENCE
Volume: 319 **Issue:** 5865 **Pages:** 948-952
DOI: 10.1126/science.1149345
Published: FEB 15 2008
View Journal Information

Abstract

The management and conservation of the world's oceans require synthesis of spatial data on the distribution and intensity of human activities and the overlap of their impacts on marine ecosystems. We developed an ecosystem- specific, multiscale spatial model to synthesize 17 global data sets of anthropogenic drivers of ecological change for 20 marine ecosystems. Our analysis indicates that no area is unaffected by human influence and that a large fraction (41%) is strongly affected by multiple drivers. However, large areas of relatively little human impact remain, particularly near the poles. The analytical process and resulting maps provide flexible tools for regional and global efforts to allocate conservation resources; to implement ecosystem- based management; and to inform marine spatial planning, education, and basic research.

Keywords

KeyWords Plus: COASTAL; FISHERIES; BIODIVERSITY; CONSERVATION; OCEAN; WORLD; AQUACULTURE; DEPLETION; HOTSPOTS; THREATS

Citation Network

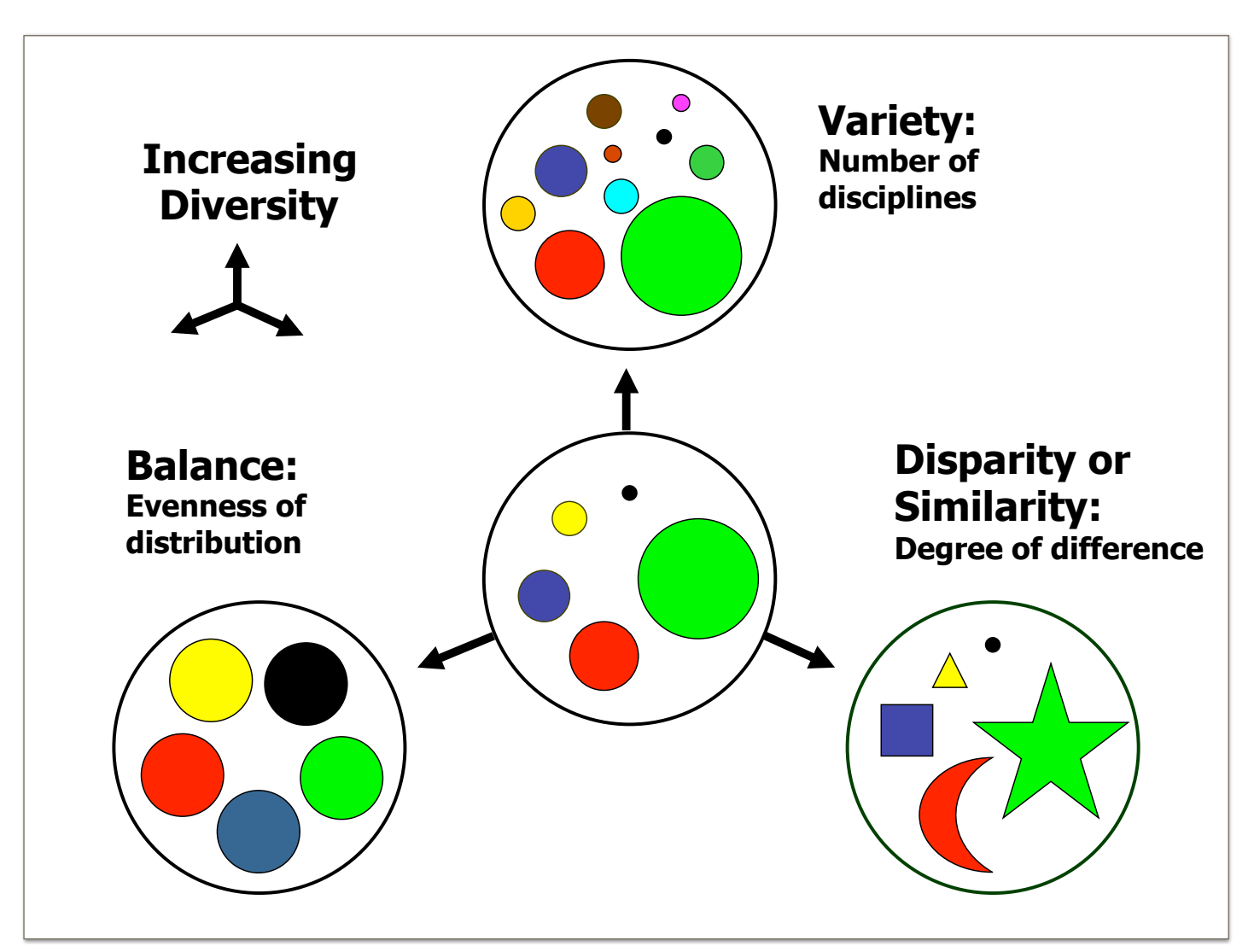
1,055 Times Cited
34 Cited References
View Related Records
View Citation Map
Create Citation Alert
(data from Web of Science ™ Core Collection)

All Times Cited Counts
1,098 in All Databases
1,055 in Web of Science Core Collection
800 in BIOSIS Citation Index
21 in Chinese Science Citation Database
0 in Data Citation Index
7 in SciELO Citation Index

Most Recent Citation

Rao-Stirling diversity: variety, balance, and disparity

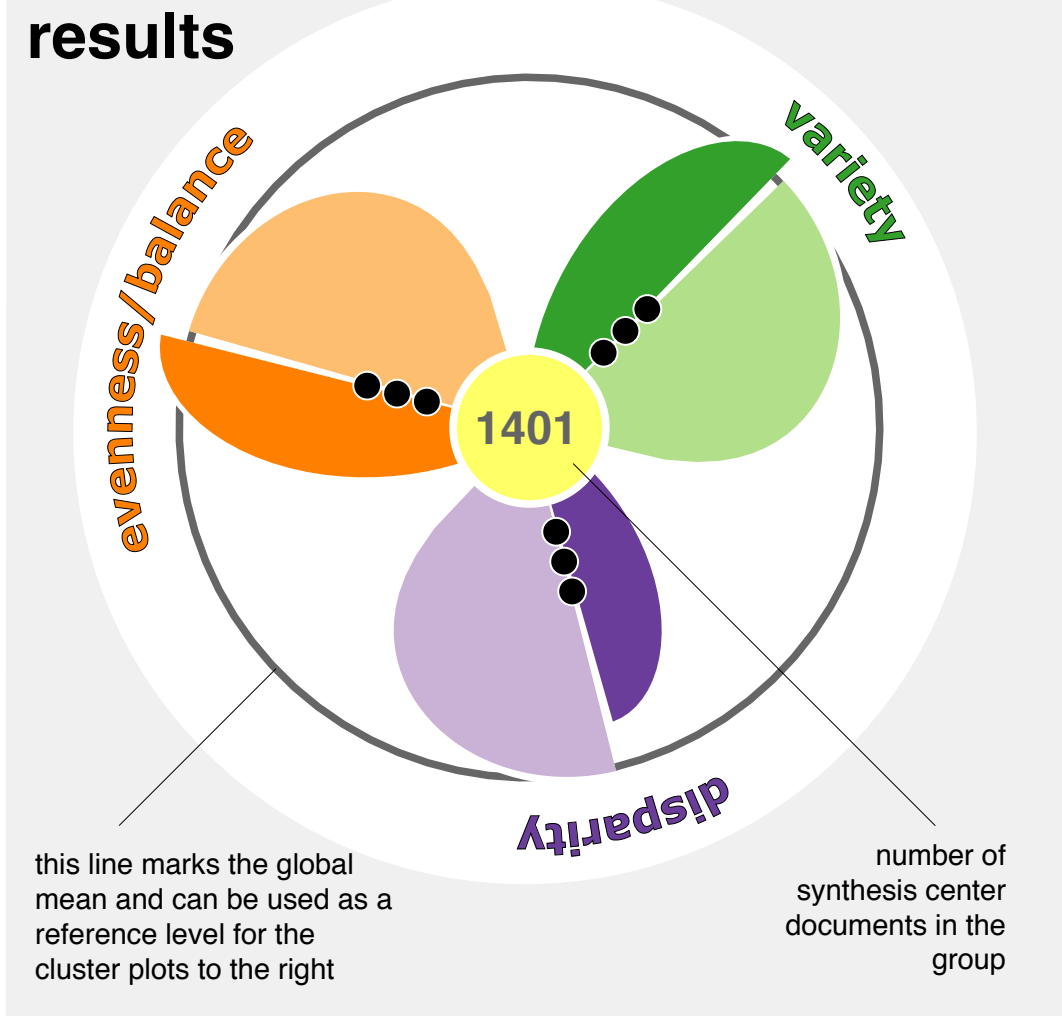
The concept of diversity has been applied in various ways in ecology and other natural sciences, information sciences, and social sciences. Rao (1982)[3] and (Stirling (2007)[4] have advanced analytical frameworks that distinguish different aspects or components of diversity: ***variety*** (the number of categories associated with an entity), ***balance*** or ***evenness*** (how evenly represented the categories are), and ***disparity*** (how dissimilar the given categories are). We apply these three components of diversity to analysis of the topic mixtures that we obtained through LDA topic modeling. To calculate variety, we counted the number of topics associated with each document; for evenness, we evaluated the shares or proportions of each of those topics in a document; and for disparity, we first measured similarity/dissimilarity between topics by evaluating how rare all pairwise combinations of two topics were (cosine similarity/distance) and then used those distances to derive disparity scores based on the combinations of topics observed in a single document.



From Rafols & Meyer, 2010[5]; based on Stirling, 1998[6]

3. Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. Theoretical Population Biology, 21, 24–43.
4. Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface, 4(15), 707–719.
5. Rafols, I. and M. Meyer (2010) Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. Scientometrics 82:263–287.
6. Stirling, A. (1998). On the economics and analysis of diversity. SPRU Electronic Working Paper. http://www.sussex.ac.uk/Units/spru/publications/imprint/sewps/sewp28/sewp28.pdf Accessed Feb 10, 2015.

global results



this line marks the global mean and can be used as a reference level for the cluster plots to the right

number of synthesis center documents in the group

HOW TO INTERPRET THE PETAL PLOTS

Petal color indicates the diversity measure: green = variety; orange = evenness/balance; and purple = disparity

Petal darkness (color value) distinguishes synthesis center documents from other documents. Synthesis center documents are represented by the darker petals, and all other documents by the lighter petals.

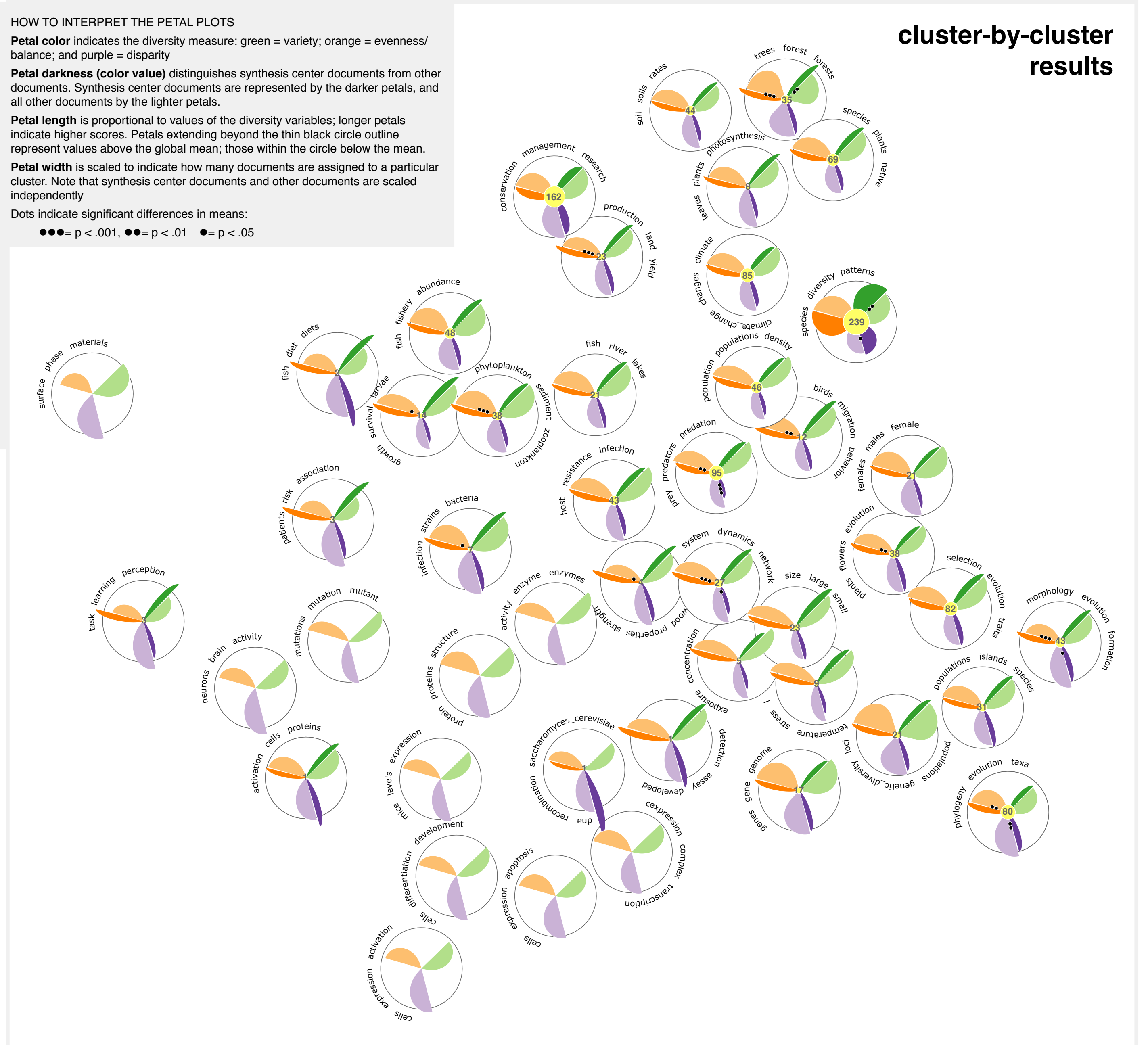
Petal length is proportional to values of the diversity variables; longer petals indicate higher scores. Petals extending beyond the thin black circle outline represent values above the global mean; those within the circle below the mean.

Petal width is scaled to indicate how many documents are assigned to a particular cluster. Note that synthesis center documents and other documents are scaled independently

Dots indicate significant differences in means:

●●● = p < .001, ●● = p < .01 ● = p < .05

cluster-by-cluster results



t-tests

We performed unequal variance t-tests to evaluate differences in means for each of the three diversity components. First, a global t-test was performed for each diversity component (variety, evenness/balance, and disparity) that evaluated all synthesis center documents vs. all other documents in the corpus. Results of these global t-tests are represented in the large flower plot above. Because averages for some measures appeared to vary with dominant topic, the same t-tests were also performed for synthesis center documents vs. all other documents in each topic cluster. Results of these per-topic tests are represented in the diagram to the right.

This figure shows the average variety, evenness/balance, and diversity scores for the documents in each topic cluster. Each topic cluster appears as a small flower; positions are the same as those used in the word cloud figure above. Here one can observe that synthesis center publications appear in over three-quarters of the topic clusters, but that they are concentrated in topics related to species diversity, conservation management, predator-prey relationships, climate change, evolution and phylogeny, and plant species. Synthesis center documents have significantly higher balance/evenness scores in many of the clusters, and significantly lower disparity scores in a smaller number. In one case (*species, diversity, patterns* topic), synthesis center publications have significantly higher disparity, but the disparity scores for that cluster overall are well below the global mean. Synthesis center publications also demonstrated higher variety scores in the species diversity and trees topic clusters, but further analysis suggests these differences are mainly attributable to differences in document length.

exploring predictors

regression analyses

p = positive correlation
n = negative correlation
ns = no significant correlation

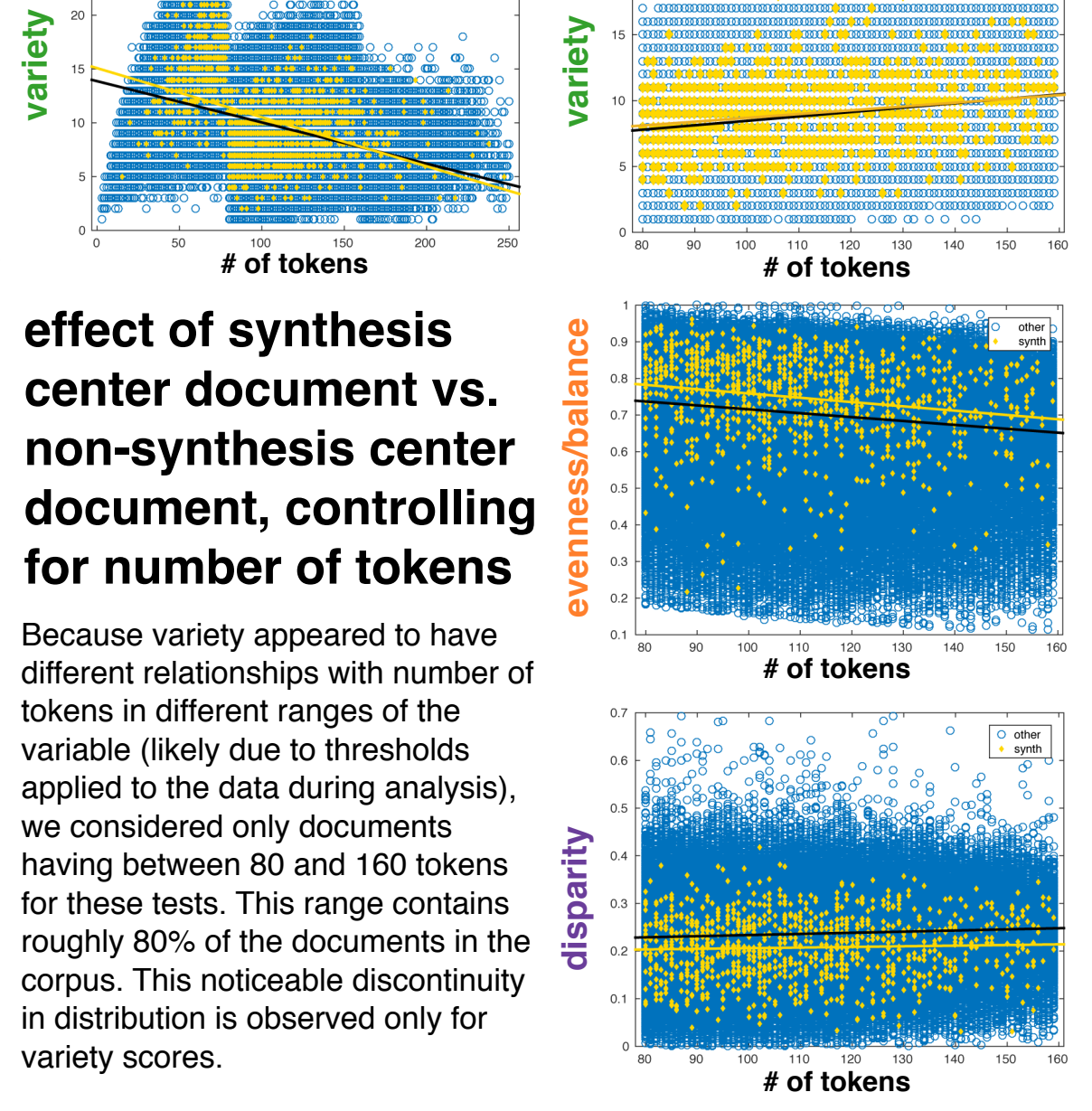
P-values for all models shown are less than .0001.

proportion of document tokens ignored due to 'junk' topic removal			number of tokens upon which inference is based			number of references cited			number of unique author addresses			number of authors			year of publication			topic cluster			R-squared (variety)	R-squared (evenness/balance)	R-squared (disparity)
v	e	d	v	e	d	v	e	d	v	e	d	v	e	d	v	e	d	v	e	d			
n	p	n	n	n	n	p	p	n	p	n	p	n	n	p	n	p	p				0.143	0.0696	0.0455
n	p	n	n	n	n	p	p	n	p	p	p	ns	n	ns	p	p	p	~	~	~	0.176	0.159	0.0993
			n	n	n	p	p	n	p	n	p	ns	n	p	n	p	p				0.108	0.031	0.0388
			n	n	n	p	p	n	p	p	p	ns	n	ns	ns	p	p	~	~	~	0.132	0.133	0.097
n	p	n				n	ns	n	p	n	p	n	n	p	n	n	p				0.0655	0.0272	0.0247
n	p	n				n	n	n	p	p	p	ns	n	ns	n	ns	ns	~	~	~	0.106	0.118	0.0783
						n	p	n	p	n	p	ns	n	p	n	n	p				0.008	0.000857	0.0129
						n	n	n	p	p	p	ns	n	ns	n	ns	n	~	~	~	0.0382	0.102	0.0727

A set of regression analyses show reasonable explanatory power for topic category across all three diversity components, but especially for evenness/balance and disparity. The number of tokens in the document and topic weight removed account for a majority of the accounted for variation in variety scores. However, the models that included all predictors accounted for only 17.6% of the variance in variety, 15.9% in evenness/balance, and 9.9% in disparity. Other potential predictors accounted for very little variance, but the signs of their coefficients yield interesting patterns. For example, while number of authors often has non-significant relationships with variety and disparity, number of unique addresses has predictably positive relationships with these variables.

Because number of tokens and topic weight removed were important predictors for variety in particular, we performed analysis of covariance to examine differences between synthesis center documents and all other documents when these variables were accounted for. Here we show ANCOVA results for number of tokens. No significant difference in slope or intercept was observed for variety when this variable was accounted for, but evenness/balance remained significantly higher for synthesis center documents, and disparity significantly lower. Because these differences remain when number of tokens and topic weight removed are accounted for, it is reasonable to interpret the cluster-by-cluster differences in evenness/balance and disparity shown in the flower plots at left as reflective of differences not accounted for by these variables.

ANCOVA



effect of synthesis center document vs. non-synthesis center document, controlling for number of tokens

Because variety appeared to have different relationships with number of tokens in different ranges of the variable (likely due to thresholds applied to the data during analysis), we considered only documents having between 80 and 160 tokens for these tests. This range contains roughly 80% of the documents in the corpus. This noticeable discontinuity in distribution is observed only for variety scores.

insights, limitations & next steps

Whereas previous studies had investigated whether interdisciplinary centers were more or less diverse without specifying aspects of diversity, the novelty of our approach is that is capable of **discerning the type of thematic diversity** that characterizes "synthesis centers": an even combination of topics (relatively high balance) that are clearly distinct but that are not extremely disparate in cognitive terms (relatively low disparity) appear to characterize the work produced by the two centers studied. We have demonstrated the fruitfulness of applying theoretical measures of diversity to thematic quantitative models of research publications, and expect that these measures, when combined with other quantitative and qualitative approaches, will significantly advance our understanding of synthesis science. However, while we have identified significant differences in diversity between synthesis center publications and other publications and cast doubt upon the idea that these differences might be accounted for by larger numbers of authors or collaborating institutions, our analyses did not suggest mechanistic explanations for how the observed differences are achieved. Furthermore, the strength of our analyses is limited by the types and quantities of research publication data that are readily available for computational analysis and the lack of information about the scientific processes through which research publications are generated. As we continue to explore the patterns in this data, we will also seek to expand the number of synthesis centers included in our analyses and combine this approach for identifying publication diversity with other methods for discerning synthetic practices and qualities at various stages of the research process.

acknowledgements

A portion of this work was supported by the National Science Foundation under grant number 1242749. NESCent and NCEAS have generously supported the meetings that initiated this research project and guided its development. Many thanks to Benjamin Adams, who assisted with collection of the Web of Science corpus, and to Debbie Donahue (NCEAS) and Laura Sheble and Casey Rawson (NESCent) for providing lists of synthesis center publications. Many of the data visualizations were created using the Text Visualization Toolbox — a MATLAB toolbox to visualize large corpus of documents, by Stacy Rebich Hespanha and João Hespanha; available at http://www.ece.ucsb.edu/~hespanha, Feb. 2010. Wordle (www.wordle.net) was used to create the word clouds, and Colorbrewer (www.colorbrewer2.org) was used for color selection.