Daume et al. Forest Ecosystems 2014, 1:11
http://www.forestecosyst.com/content/1/1/11

Forest Ecosystems
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling

Stefan Daume[1,2*], Matthias Albert[3] and Klaus von Gadow[1]

Abstract

Background: With mounting global environmental, social and economic pressures the resilience and stability of forests and thus the provisioning of vital ecosystem services is increasingly threatened. Intensified monitoring can help to detect ecological threats and changes earlier, but monitoring resources are limited. Participatory forest monitoring with the help of "citizen scientists" can provide additional resources for forest monitoring and at the same time help to communicate with stakeholders and the general public. Examples for citizen science projects in the forestry domain can be found but a solid, applicable larger framework to utilise public participation in the area of forest monitoring seems to be lacking. We propose that a better understanding of shared and related topics in citizen science and forest monitoring might be a first step towards such a framework.

Methods: We conduct a systematic meta-analysis of 1015 publication abstracts addressing "forest monitoring" and "citizen science" in order to explore the combined topical landscape of these subjects. We employ 'topic modelling', an unsupervised probabilistic machine learning method, to identify latent shared topics in the analysed publications.

Results: We find that large shared topics exist, but that these are primarily topics that would be expected in scientific publications in general. Common domain-specific topics are under-represented and indicate a topical separation of the two document sets on "forest monitoring" and "citizen science" and thus the represented domains. While topic modelling as a method proves to be a scalable and useful analytical tool, we propose that our approach could deliver even more useful data if a larger document set and full-text publications would be available for analysis.

Conclusions: We propose that these results, together with the observation of non-shared but related topics, point at under-utilised opportunities for public participation in forest monitoring. Citizen science could be applied as a versatile tool in forest ecosystems monitoring, complementing traditional forest monitoring programmes, assisting early threat recognition and helping to connect forest management with the general public. We conclude that our presented approach should be pursued further as it may aid the understanding and setup of citizen science efforts in the forest monitoring domain.

Keywords: Forest monitoring; Citizen science; Participatory forest monitoring; Probabilistic topic modelling; Text analysis

* Correspondence: stefan.daume@ecoveillance.org
[1]Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany
[2]Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden
Full list of author information is available at the end of the article

Background

The ability of ecosystems worldwide to provide essential products and services is being threatened by major environmental, social and economic changes (Millenium Ecosystem Assessment 2005), and there is a rising demand for intensive monitoring to detect threats and potentially catastrophic changes earlier (Biggs et al. 2009). Forests provide many vital ecosystem services, but with increasing ecological and economic pressures their resilience and stability are under threat. Forest managers and scientists are thus required to constantly re-evaluate and communicate strategies for intensified monitoring; this includes general environmental monitoring for emerging threats but also "traditional" forest monitoring using field plots and remote sensing for forest management purposes. In addition, there is an urgent need to inform and educate the general public on the value of forest ecosystems and the direct and indirect anthropogenic influences on forests (European Environment Agency 2011a; European Environment Agency 2011b).

Participatory forest monitoring – involving local communities and stakeholders in forest monitoring activities - plays an increasingly important role in delivering useful information, especially in areas where communities are relying heavily on forests for their livelihood and where a community's forest use can have massive impacts on the ecosystem (Evans and Guariguata 2008). Participatory monitoring is thus one avenue to provide additional resources to intensify forest monitoring.

In research generally, *"citizen science"* – the volunteer participation of members of the public in scientific projects - has emerged as a valuable tool in data collection, processing and dissemination, and offers effective channels for educating the general public on research (Bonney et al. 2009). Many citizen science projects cover subjects in the environmental domain (Silvertown 2009; Bonney et al. 2009), but citizen science extends over a broad set of application areas (such as astronomy, cancer research, etc.) utilising a wide range of skills, interests and motivations.

Citizen science biodiversity monitoring projects in general (Silvertown 2009) can potentially deliver information relevant to forest monitoring programmes. In fact, volunteers are already contributing to specific forest monitoring challenges. The *Living Ash Project* (http://livingashproject.org.uk/) for example aims to counter the effects of *Ash dieback* disease by calling for members of the public to tag and regularly monitor ash trees with the long-term objective to identify pest-resistant trees. Mobile and web technologies in particular help to facilitate these contributions: Ferster and Coops (2014) report that citizen scientists can use smartphone applications to collect data on forest fuel loading to identify wildfire hazards, and the *Forest Watchers* web application (http://forestwatchers.net) calls on volunteers to identify remote deforested areas in aerial images.

While these projects can make a potentially dramatic difference to existing monitoring efforts, they still represent singular and often localized efforts. A solid, generic and applicable framework or toolset for utilising the true potential of citizen science projects in the forestry domain still seems to be lacking. We propose that a better understanding of shared and related topics in citizen science and forest monitoring can be a first step towards opening up citizen science as an additional resource in the forest monitoring toolset.

Accordingly, this contribution explores the potential of citizen science initiatives in forest monitoring from a high-level perspective through an assessment of topical overlaps in the published literature on "citizen science" and "forest monitoring". Specifically, we are interested in a fine-grained analysis and the discovery of latent topics that may point to opportunities in employing citizen science for the benefit of forest science. Such a meta-analysis could be a first step in encouraging new developments and specific designs of citizen science initiatives in the forest monitoring domain.

Data and methods

For the proposed meta-analysis we employ an approach known as "topic modelling", an unsupervised probabilistic machine learning method for the automatic analysis of large text collections (Blei 2012). Topic modelling has seen a rising number of applications in recent years with an emphasis on applications in the digital humanities (Blevins 2010; Templeton et al. 2011; Yang et al. 2011), but also for bibliometric analysis of publications in the natural sciences (Griffiths and Steyvers 2004; Blei and Lafferty 2007). The technique has been employed both to discover topics in text collections and to structure document sets for advanced searching.

In this study, we apply probabilistic topic modelling to analyse a combined collection of scientific articles on the subjects of "forest monitoring" and "citizen science". We aim to provide a description of the combined topical landscape of these two broad thematic sets of publications, explore to what extent shared topics exist, which topics are clearly separated but potentially related and discuss the potential of this approach in providing new insights and opportunities for citizen science applications in the forest monitoring domain.

Data

We applied topic modelling to a set of documents obtained through a search in the literature database Scopus for documents published from 1994 to 2013, explicitly mentioning the terms "forest monitoring" or "citizen science" in the

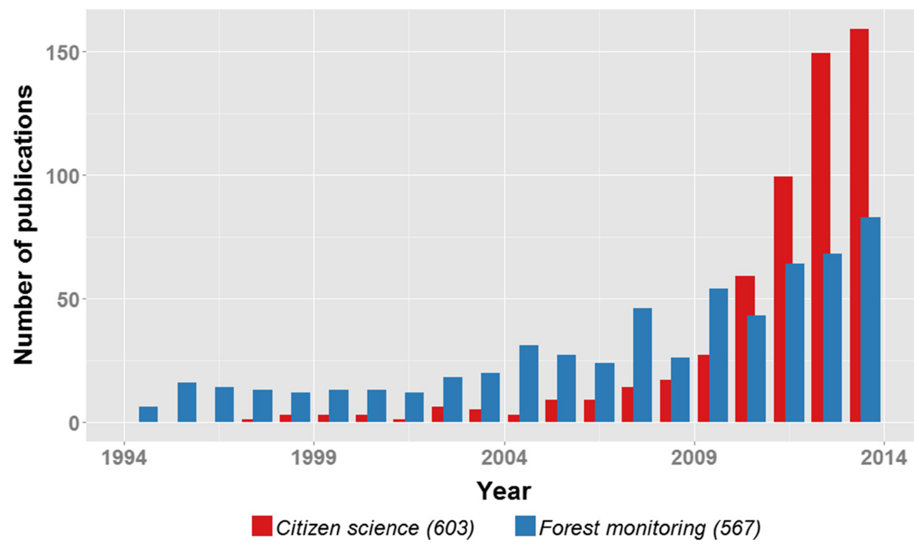

Figure 1 Distribution of "citizen science" and "forest monitoring" publications from 1994 to 2013 according to Scopus. The topic analysis included publications explicitly mentioning "citizen science" or "forest monitoring" in the title, abstract or keywords (based on a search in Scopus).

title, abstract or keywords. Figure 1 shows the development of the number of publications for the two document sets. The increase of the citizen science material since 2005 is rather dramatic. Only two articles (Roman et al. 2013; Butt et al. 2013), both published in 2013, contained both search terms; for our analysis we assigned Roman et al. (2013) to the "forest monitoring" and Butt et al. (2013) to the "citizen science" document set.

We obtained the abstracts for each matching publication for analysis, but excluded all documents not published in English as well as documents with abstracts of less than 100 words, which left 477 documents on "citizen science" and 538 documents on "forest monitoring". Many of the "forest monitoring" publications present a global coverage, though with an apparent bias towards studies focusing on Europe and North America. Our "citizen science" publications refer almost exclusively to projects in North America and Europe. This bias is also reflected by the geographical distribution of the corresponding authors of the two sets of publications.

Prior to running the topic modeller the text corpus is split into tokens and so-called stop-words (e.g. "the", "and", "if") are removed. The quality of the topic analysis can often be further improved by removing additional domain specific stop-words; we added "citizen", "science", "forest" and "monitoring" to the stop-word list since one of either combination would have occurred in every document which effectively turned them into stop-words. In addition, all words occurring only once were removed from the text corpus. This left us with a vocabulary of 6.181 unique terms, occurring a total of 100.274 times in the 1.015 abstracts.

Probabilistic topic modelling

Probabilistic topic models represent a suite of algorithms for analysing large document collections and identifying the distinct latent topics in these documents (Steyvers and Griffiths 2007; Blei 2012). Topic models are based on the assumption that documents are typically composed of multiple topics. Each topic in turn may be viewed as a distinct set of unique words that frequently occur together. All documents in a set share multiple topics, but individual documents will exhibit only a subset of all available topics to a certain degree. More formally let:

- W be the unique set of words in
- D a set of documents containing
- T topics, where
- each topic t is a discrete probability distribution Φ_t over all words w and
- each document d has a specific distribution Θ_d over all topics T.

Topic modelling is based on the assumption that each document d is the result of a generative process by which iteratively a topic t is first drawn from Θ_d and then a word w is drawn from Φ_t until the document is complete. Topic modelling algorithms reverse this assumed document creation process in order to infer topics and topic compositions that best explain a set of observed variables, here represented by the word occurrences in a given set of documents.

Figure 2 illustrates informally the intuition behind topic modelling: assuming that the topic composition of a document and the frequencies with which words appear

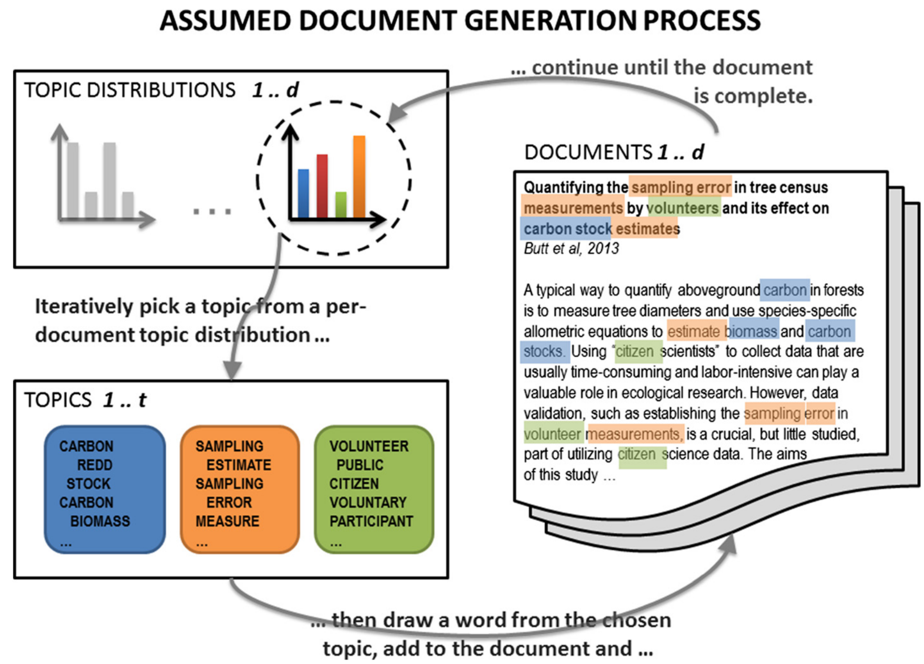

Figure 2 Informal explanation of the intuition behind topic modelling (adapted from (Blei 2012)). An excerpt of a sample abstract (Butt et al. 2013) is quoted for illustration purposes. A document is assumed to be the result of an iterative process that selects topics from a document-specific probabilistic topic distribution and words from a topic-specific probabilistic word distribution. Topic modelling algorithms reverse this process in order to find assignments for the two distributions that best explain a set of documents.

in a topic are known, a document can be generated by iteratively choosing words from a topic according to the frequencies of the topics. A topic modelling algorithm then reverses this process by, simply put, assigning the words in a given "observed" set of documents to topics, and topic distributions to documents, such that a set of documents generated on the basis of these distributions best fits the set of "observed" documents.

Topic models typically employ variational inference (Asuncion et al. 2009) to estimate the best topic-word and topic-document assignments. We use a topic model called *Latent Dirichlet Allocation (LDA)* first described by Blei et al. (Blei et al. 2003). In LDA the assumed prior distributions for Θ_d and Φ_t are Dirichlet distributions with concentration parameters α and β respectively. The choice of these so-called hyper-parameters determines the sparsity of the distributions and thus the variability in likelihood with which words will be assigned to topics and topics to documents. LDA has emerged as a reliable and popular topic modelling approach successfully applied in many different domains. Furthermore, it offers several freely available implementations. We use the MALLET machine learning package (McCallum 2002) which provides an open source implementation of LDA.

LDA configuration – choosing the number of topics

A key choice in running topic modelling algorithms is the number of distinct topics that are expected to be covered by the document corpus. The number of topics T and the priors α and β are the only required input parameter for LDA, but they have a significant impact on the resulting topic assignments. Choosing larger topic numbers may result in a fragmentation of topics which may not always be easy to interpret semantically. However, MALLET offers a feature called *hyper-parameter optimisation* which alleviates the impact of the chosen topic number (Wallach et al. 2009a) and allows to safely work with larger topic numbers.

It can be argued that the choice of T is ultimately an arbitrary one driven by the research questions and the intended use of the resulting topic model; small topic numbers will result in semantically broad topics, with increasing topic numbers, those broader topics will be split in semantically more refined topics. Several evaluation methods allow however a quantitative assessment of the optimal number of topics (Wallach et al. 2009b). We followed an approach chosen by Griffiths and Steyvers (2004) and compared the converged log-likelihood (LL) per token (returned by the LDA algorithm) as a measure of best model fit for topic numbers T ranging from 10 to 300. We repeated 10 topic analyses for each T in this range and measured the final LL/token which suggested 100 topics as a suitable topic number for our analysis (see Figure 3).

We thus ran MALLET's implementation of LDA with 100 topics. The algorithm was run for 2000 iterations

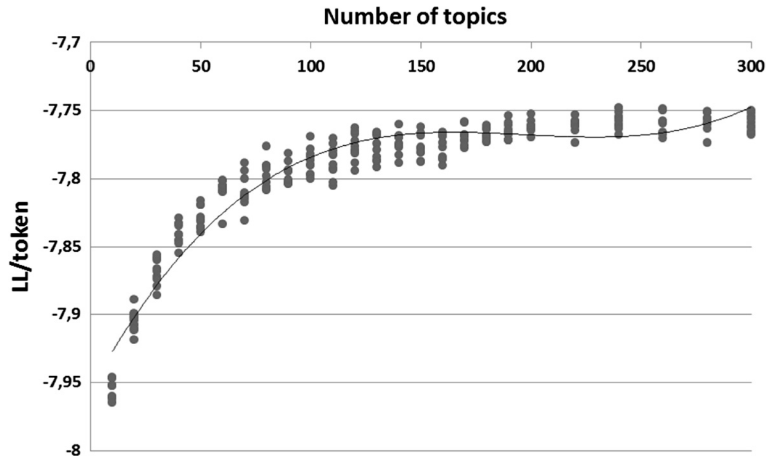

Figure 3 Evaluation of topic model fit with different topic numbers. The relationship between the number of topics and the log-likelihood (LL) per token as a measure of best model fit for topic numbers is shown for 10 sample MALLET LDA modelling runs for topic numbers in the range from 10 to 300.

with the hyper-parameter optimization feature enabled, producing a topic model that will be presented in the next section.

Results

The MALLET LDA topic modelling implementation produces two main outputs that will be referred to further analysis:

1. **Topic word sets for each topic:** the collection of terms with associated occurrence frequencies that characterise a topic.
2. **Topic composition for each analysed document:** the share of each topic in a given document.

Table 1 lists a sample selection of the 100 topics returned by the topic modelling algorithm for our document corpus. For each topic the 10 most frequent terms and relative word frequencies in the topic are provided. It is important to note that the topic modelling algorithm returns purely a distribution of topic terms that do not come with a semantic interpretation. Suitable labels can however often be inferred from the word frequencies. In the following we will either refer to a topic by its ID (0–99) assigned by the topic modelling algorithm or labels that we assigned on inspection of the most frequent terms in a topic.

A term is not necessarily exclusive to one topic. We find the term “results” for example as a top term in both topics 69 and 38 (see Table 1). In both cases it co-occurs with terms that are characteristic for scientific publications in general and would thus be expected in publications on citizen science as well as forest monitoring; both topics were accordingly labelled “science study”. The term “change” can be found in topic 38 (“science study”) and topic 32 (“climate change”). For generic terms like this the most frequent co-occurring terms as well as the term’s specificity to a topic can help to infer suitable labels. This may also clarify topic semantics in case of ambiguous term combinations. Topic 85 for example combines astronomy terms like “galaxy” and “supernovae” with “dna” and “genetic”. Figure 4 plots terms according to their frequency in and specificity to a topic for three sample topics. Considering these two dimensions suitable topic labels - here “galaxies”, “risk perceptions” and “birds” - can usually be suggested even for heterogeneous or ambiguous word combinations.

For each document in the analysed corpus the resulting topic model will include a topic composition distribution which specifies the shares of each topic in a given document. Figure 5 shows a sample topic composition for one (Butt et al. 2013) of the two publications that matched both the search term “citizen science” and “forest monitoring”. This example illustrates that only a small number of topics are active in this document. A comparison with the publication abstract confirms that the topic composition shown in Figure 5 appears to closely reflect the content of the analysed abstract.

For our analysis we were furthermore interested in the distribution of topics between the “citizen science” and “forest monitoring” document corpora. Figure 6 combines the topic composition of all analysed documents and shows the distribution of topics for the combined document corpora.

The cumulative topic distribution in Figure 6 includes only topic proportions greater than 0.02. The topic modelling algorithm attempts to assign a share of each of the chosen 100 topics for every document, but as the sample topic composition in Figure 5 illustrated, this will result in a large number of very small and negligible proportions.

Table 1 Most frequent words and relative word frequencies by topic for a sample set of topics

Topic 69 "science study"		Topic 38 "science study"		Topic 67 "information systems"		Topic 32 "climate change"	
Results	0.047	Study	0.029	Information	0.049	Climate	0.180
Methods	0.046	Results	0.027	Systems	0.036	Change	0.129
Method	0.044	Change	0.024	Development	0.035	Effects	0.037
Study	0.030	Time	0.022	Paper	0.033	Water	0.026
Accuracy	0.030	Analysis	0.019	Support	0.024	Ecosystems	0.024
Based	0.028	Studies	0.018	Developing	0.019	Response	0.022
Compared	0.023	Large	0.017	Process	0.019	Management	0.021
High	0.021	Significant	0.016	Framework	0.017	Integrated	0.016
Evaluated	0.014	Years	0.016	Key	0.016	Ground	0.016
Developed	0.014	Found	0.015	Based	0.013	Impacts	0.015
Topic 0 "volunteer surveys"		**Topic 30 "education"**		**Topic 24 "plant phenology"**		**Topic 85 "galaxies"**	
Volunteers	0.161	Students	0.112	Plant	0.082	Galaxy	0.065
Volunteer	0.118	Learning	0.056	Phenology	0.078	Galaxies	0.054
Collected	0.041	Education	0.056	Plants	0.066	Zoo	0.044
Scientists	0.030	Classroom	0.024	Species	0.059	Project	0.026
Groups	0.025	Teaching	0.020	Phenological	0.043	dna	0.026
Recording	0.021	School	0.019	Interactions	0.033	Morphological	0.026
Professional	0.020	Literacy	0.018	Network	0.032	Spiral	0.023
Surveying	0.020	Teachers	0.018	Networks	0.032	Supernovae	0.021
Environment	0.019	Educational	0.017	Observations	0.023	Genetic	0.021
Motivations	0.018	Experiences	0.013	Timing	0.023	Classifications	0.016
Topic 53 "forest growth"		**Topic 87 "SAR"**		**Topic 91 "ozone"**		**Topic 20 "carbon stocks"**	
Tree	0.126	Sar	0.068	Ozone	0.103	Carbon	0.091
Trees	0.082	Coherence	0.038	Concentrations	0.051	Redd	0.052
Growth	0.064	Radar	0.034	Sites	0.048	Countries	0.049
Species	0.032	Backscatter	0.032	Measured	0.036	National	0.044
Structure	0.032	l-band	0.023	Site	0.029	Change	0.032
Area	0.023	ers	0.022	Passive	0.027	Deforestation	0.031
Composition	0.021	Stands	0.022	Symptoms	0.023	Stocks	0.024
Plots	0.020	Biomass	0.022	Measurements	0.021	Climate	0.022
Diameter	0.020	Areas	0.020	Critical	0.019	Inventory	0.019
Conditions	0.017	Images	0.020	Sampling	0.019	Reporting	0.017

The table shows sample generic topics (top row), typical citizen science topics (middle row) and typical forest monitoring topics (bottom row). For each topic the topic ID and a representative label is provided.

The threshold of 0.02 was chosen, because the average number of words per analysed abstract after removal of stop-words was approximately 100 - a topic proportion of 0.02 thus corresponds to two words, which we propose is the absolute minimum for a semantic interpretation of a topic assignment. On average only 9 topic proportions per document are greater than 0.02, which however cumulatively explain approximately 90 % of the document.

Figure 6 illustrates that several topics have a large contribution from either corpus, thus occurring with high frequency and large proportions in both "citizen science" and "forest monitoring" publications - examples include topics 6, 38, 69 (see Table 1 for topic words). These topics combine keywords which are typical for scientific studies in general, thus topics which can be expected to be shared between the two corpora.

More specific large topics shared between the two corpora exist as well: topics 67 (labelled "information systems") and 88 ("large-scale analysis") are examples that seem to fit data intensive research fields. Given that these topics are characteristic for both domains the question arises whether shared topics in general point to synergies that could guide intensified citizen science contributions in forest monitoring. Similarly, examples

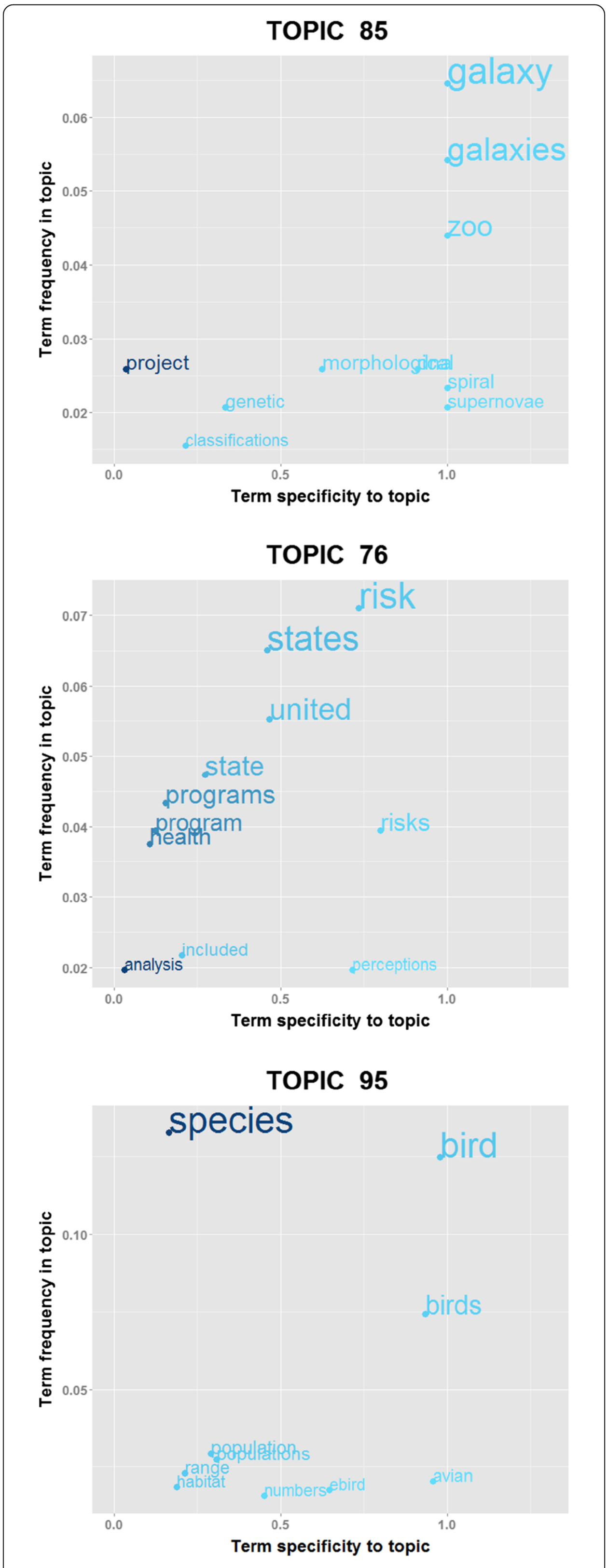

Figure 4 Term/topic frequency and specificity for three sample topics: "galaxies" (T85), "risk perceptions" (T76), "birds" (T95). A term that is exclusive to one topic has a specificity of 1. The relative size of the plotted words is proportional to their frequency in the topic. The colour gradient from light to dark blue indicates larger frequencies of a word in the complete document set.

of specific and shared but less frequent topics - for example 47 ("local/community-based"), 76 ("risk perceptions"), 96 ("urban environments") or 97 ("natural resource management") - have to be evaluated from the same angle and we will refer to those in more detail in the discussion section.

Typical or exclusive topics in either the "citizen science" or "forest monitoring" publications are also of particular interest. Here the question arises whether these are niche topics, truly un-related domains or potential examples of non-utilised citizen science opportunities in forest monitoring. Examples that we can discuss in "citizen science" are topics such as 0 ("volunteer surveys"), 24 ("phenology"), 27 ("citizen science"), 35 ("social media") and 85 ("galaxies"); dominant topics in the "forest monitoring" corpus include 1 ("crown studies"), 16 ("clearcuts"), 53 ("forest growth") and 87 ("SAR/remote sensing").

We conclude the result section with a network representation of the topical landscape of the analysed documents (Figure 7). Each document and topic is represented by a node in the network graph. An arc between a document and a topic represents a share of this topic in the connected document. The size of the topic nodes reflects their overall share in the analysed corpus. All topic proportions less than 0.02 in an individual document were excluded from this representation.

The network representation in Figure 7 provides a comprehensive visual summary of the topical structure of the combined document corpus, and confirms and extends the results in Figure 6. While the two document corpora have an intersection around major generic shared topics such as 6/38/69 ("science study"), 67 ("information systems") or 88 ("large-scale analysis"), they are visually clearly separated in the network layout. Corpus-specific topics such as 39 ("remote sensing") or 27 ("citizen science") are located in the centre of the respective document cloud, confirming that they are largely exclusive to these document corpora.

Discussion

Given the global pressures on forest ecosystems and the resulting challenges for forest managers and researchers, forest monitoring can benefit from additional and intensified efforts through citizen science projects. However, the topical landscape obtained through our analysis suggests that this opportunity is not yet pursued to a large extent. Although shared topics exist, the obtained topic

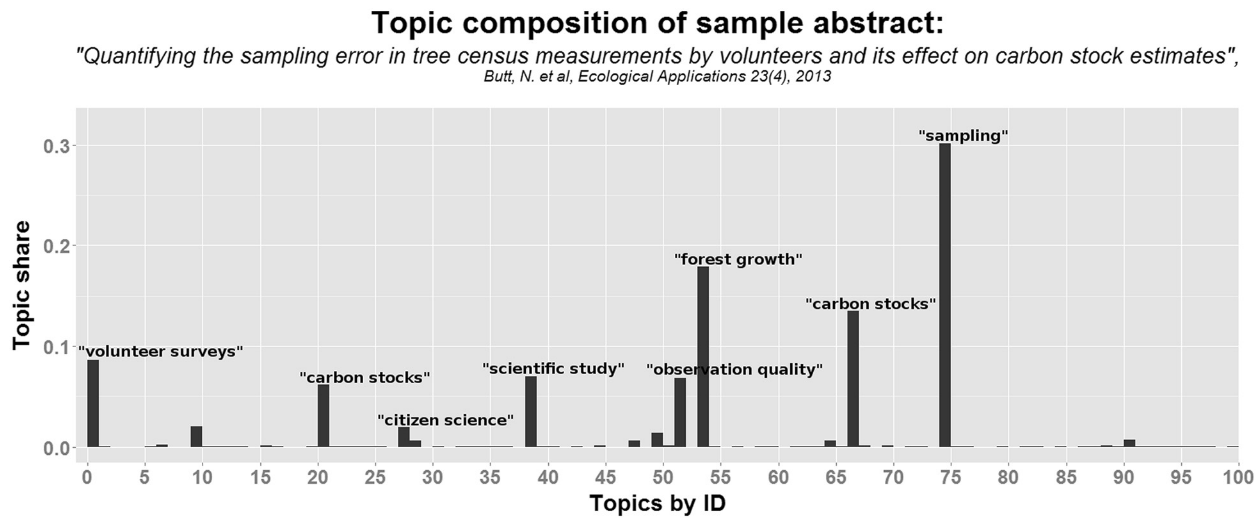

Figure 5 Topic composition of one publication abstract (Butt et al. 2013). The topic proportions reflect the output of a MALLET LDA topic analysis with 100 topics. Topic labels were added for the major topics.

model confirms the results hinted at by only two publications (Roman et al. 2013; Butt et al. 2013) in our document set that matched both the terms "citizen science" and "forest monitoring".

Obviously, the generalizability and conclusiveness of the results is limited by the size of the document collection and the analysed documents. Compared to similar studies – for example (Griffiths and Steyvers 2004), which used 28.154 abstracts, with more than 3 million words and a vocabulary of 20.551 words - our set of 1.015 abstract and a vocabulary of 6.181 terms occurring 100.274 times is significantly smaller. We were nevertheless able to identify many topics with consistent semantics - see for example topic 30 ("education") or topic 95 ("birds") – and the topic composition of sample articles (see Figure 5) seemed to reflect the content well. However, we also found topics like 85, which – while dominated by terms justifying the label "galaxies" - also included terms referring to genomics ("dna", "genetic"), pointing at a lack of granularity that can be attributed to the size of the document set and the vocabulary. A closer look at topic compositions of several sample documents in our set suggests that the quality of the topic assignments for a document correlates with the size of the text - longer abstracts display a more representative topic composition; taking into account other case studies in topic modelling we conclude that larger documents and thus vocabularies would probably deliver more representative topic structures for individual documents and topics with more refined and consistent semantics. A further improvement in the semantic interpretation and consistency of discovered topics might be achieved by exploring variations of topic models that consider word bigrams – the reoccurrence of e.g. the bigram "biodiversity loss" allows a more conclusive

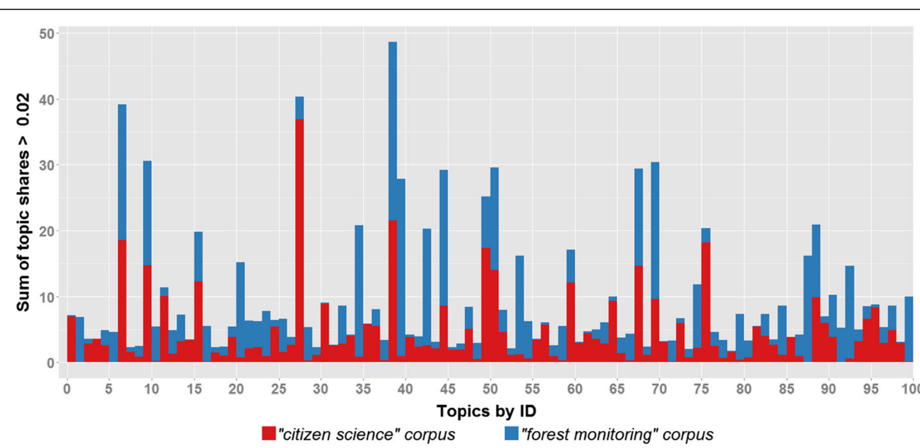

Figure 6 Contributions of topic shares by analysed document corpus. Cumulative shares per topic contributed by the "citizen science" (red) and "forest monitoring" (blue) corpus respectively. Result of a MALLET LDA topic analysis with 100 topics.

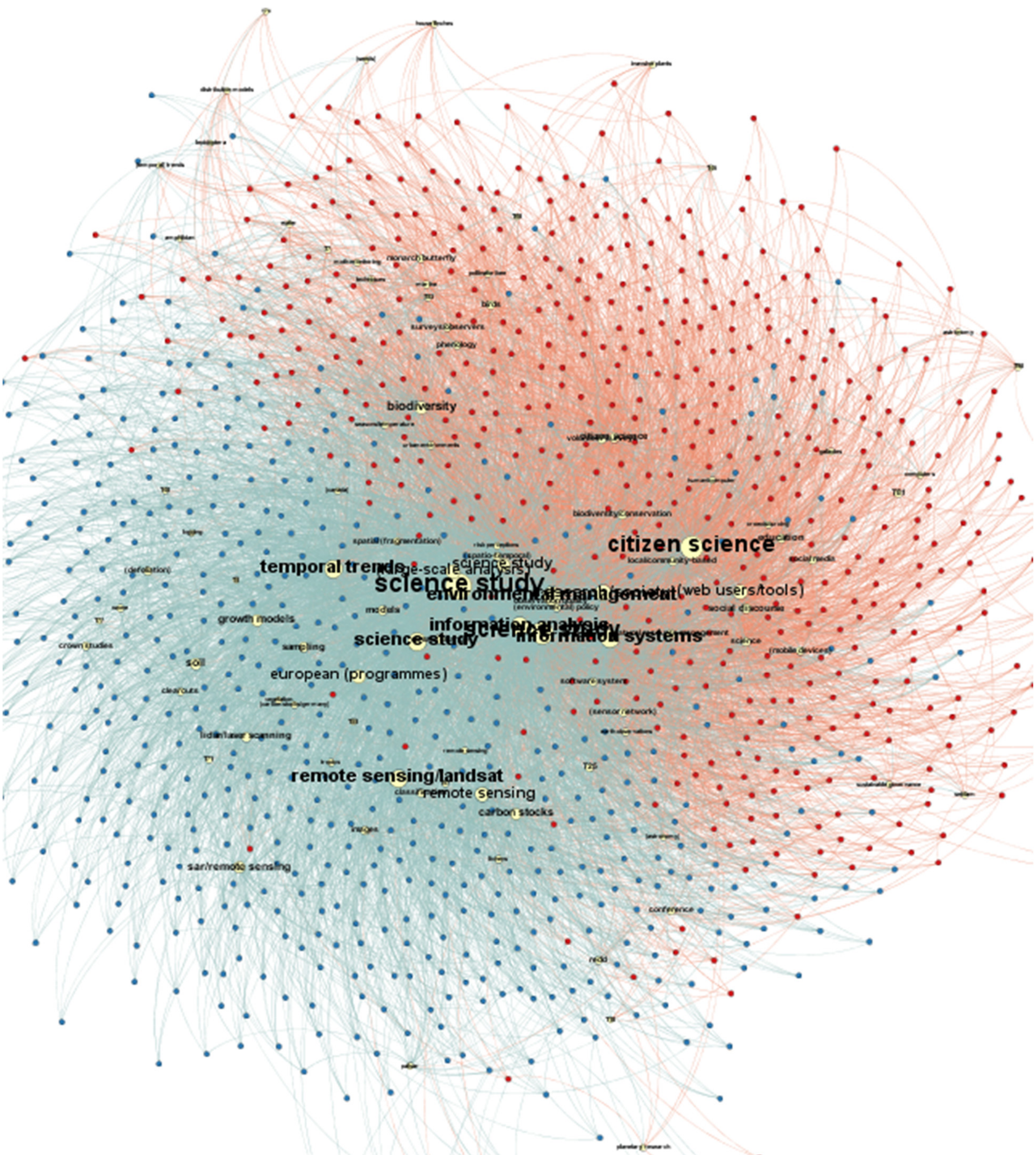

Figure 7 Network representation of the topics and documents in the analysed corpus. Blue nodes represent articles on "forest monitoring", red nodes articles on "citizen science" and yellow nodes the topics they are connected with; only topic shares greater than 0.02 were included as arcs in the network. The graph was generated with the Gephi (https://gephi.org) network visualization tool using a Fruchterman-Reingold network layout.

interpretation of a topic than the individual occurrence of the two words "biodiversity" and "loss".

Despite these potential methodical improvements, we nevertheless gained interesting initial insights in the combined topical landscape of "citizen science" and "forest monitoring" publications. Shared topics can be found and extend beyond the generic topics that would be expected in scientific publications in general. Shared topics such as "urban environments" (96) or "local/community-based" (47) indicate common themes, topics such as “spatio-temporal” (82) or “software development” (13) common tools and techniques, and topics like “risk perceptions” (76) or “climate change” (32) hint at shared research motivations. It can be argued that these results are “stating the obvious” and similar results could be obtained through traditional manual literature analysis. However, topic modelling is a scalable method that can be applied equally to very large document corpora, full-text analysis of publications and a much larger number of topics, and thus suggests topic modelling as a suitable method not only for a snapshot analysis but also for a continuous analysis of growing document sets. In addition, the consistency of our results with “the obvious” supports our other observations for the document set, particularly that major forest monitoring topics – e.g. “carbon stock” (20) estimates or “forest growth” (53) – are not shared between the document corpora.

In contrast, certain topics with large cumulative shares in the document corpus which are exclusive to or typical for either “citizen science” or “forest monitoring” publications point at interesting opportunities. Topics such as “galaxies” (85) or “astronomy” (43) indicate successful citizen science projects involving the analysis and classification of telescopic images by volunteers. Several articles in our corpus refer for example to the *Galaxy Zoo Supernovae* project (http://supernova.galaxyzoo.org) on the *Zooniverse* citizen science web platform, where volunteer participants were asked to compare changes between images of a specific region of the night sky taken at different times in order to identify supernovae (Smith et al. 2011). Participants were not required to have a background in astronomy, but still delivered classification results of “remarkable quality” (Smith et al. 2011).

In the forest monitoring corpus “remote sensing” (topics 42/87) emerged as a major topic (see Figure 7) and an area that will involve similar tasks and skills as the classification of telescopic images in the supernovae project. While rooted in different domains, both topics focus on image analysis and classification and thus have not only skill sets and techniques in common, but possibly also a citizen science community that could be mobilised for citizen science initiatives in the forest monitoring domain. Indeed, examples of remote sensing projects with volunteer participation can be found, for example in land cover monitoring (http://geo-wiki.org (Fritz et al. 2009), http://forestwatchers.net) or biomass estimates (Fritz et al. 2013), but are still an exception. A possible explanation is that “citizen science” as a research tool is still at an early stage of recognition in the forest monitoring domain, but also that there are concerns over the quality of citizen science data which will determine the applicability of inferred results (See et al. 2013). With reference to the example of remote sensing we propose that an understanding of topical landscapes across domains could contribute to citizen science projects delivering high quality data and results by learning from communities with similar tasks and techniques, finding participants with matching skills and utilising tested frameworks from other domains.

More topic examples largely exclusive to the forest monitoring document corpus which might benefit from intensified monitoring through citizen science are e.g. “carbon stock” estimates (20) or “ozone” effects (91) – citizen science has been explored in these areas (see for example (Sachs 2008)), but not as major research tool in forest monitoring. The topic “education” (30) on the other hand is almost exclusively found in the “citizen science” domain. In light of an increased need to communicate forest policies, threats and values to the general public, this observation points to citizen science as an important communication channel that should find more consideration in the forestry domain.

This exploratory study indicates that the two research areas represented by the document corpora on citizen science and forest monitoring exhibit shared topics, but that promising opportunities to utilise citizen science for key forest monitoring themes still lie dormant. Citizen science projects will be most successful, both in terms of research outcomes and the perceived value for participating volunteers, when projects are designed with a good understanding of the formal models of participation (Bonney et al. 2009; Shirk et al. 2012) and a clear alignment with key research process steps (Newman et al. 2012). We believe that the consideration of the combined topical landscape of citizen science and its (potential) application areas can contribute to the deliberate design of citizen science projects and the success of these projects. The discovery of shared latent topics could be of value when directing researchers and stakeholders in either field to matching resources (articles, studies, methods), connect communities and thus facilitate citizen science projects in the forest domain.

However, these first findings - while intriguing - are still too limited to permit general conclusions. We believe that our initial results confirm topic modelling as a valuable method, but that the conclusiveness of the results could be improved by broadening the thematic scope and the size and number of the analysed documents - for this exploratory analysis we chose to focus on publications explicitly mentioning the terms “citizen science” and “forest monitoring” and hence missed, by design, many citizen science projects in for example forest threat monitoring; furthermore, we analysed abstracts only.

Future research should therefore not only extend the topic analysis to full-text articles but should also pursue a broader thematic focus and include publications from other databases such as NGO project studies as well as publications that apply a different terminology to the subject area for example by using terms like “crowdsourcing”, "public participation in research", "forest inventory", "forest modelling" or "forest planning" instead of "citizen science" and "forest monitoring". When using a larger dataset and running the topic analysis with larger numbers of topics, more-fine-grained topics pointing to specific techniques, skill sets or communities might emerge that would allow to draw conclusions that are more generalizable and point to specific promising citizen science opportunities in the forest monitoring domain.

Conclusions

The application of probabilistic topic modelling for characterizing the shared topical landscape of publications on citizen science and forest monitoring confirmed that the method is useful as a scalable approach for a meta-analysis of large document collections in the chosen domain. While the conclusiveness of the findings is somewhat limited by the number of documents analysed, even this exploratory topic analysis indicates interesting shared motivations and skills, and under-utilised opportunities for citizen science projects in forest monitoring can be inferred from this study.

Citizen science projects in the area of forest monitoring have the potential to contribute to the earlier recognition of forest threats, supplement resources in traditional inventory programs, provide pointers for areas requiring intensified monitoring, indicate public demands on forests and connect forest practitioners and researchers with the general public. In the interest of utilising citizen science for intensified monitoring efforts, communication and public awareness, the presented topic modelling approach should be pursued further and may assist both citizen science and forest monitoring communities in connecting resources and stakeholders, thus possibly aiding in the future deliberate design of more numerous and ambitious citizen science initiatives in the forestry domain.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The first author (SD) conducted the data collection and topic analysis, compiled the results and wrote the first draft of the article. SD, MA and KG then jointly restructured and improved the presentation during several major revisions. All authors read and approved the final manuscript.

Acknowledgements

This study was inspired by many discussions on topic modelling the first author had with Emma Sundström and Ingo Fetzer of the Stockholm Resilience Centre. This input is gratefully acknowledged. Finally, the authors would like to thank the anonymous reviewers for their valuable feedback which helped to improve this publication.

Author details

[1]Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany. [2]Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden. [3]Northwest German Forest Research Institute, Grätzelstraße 2, 37079 Göttingen, Germany.

Received: 29 May 2014 Accepted: 10 July 2014
Published: 30 July 2014

References

Asuncion A, Welling M, Smyth P, Teh YW (2009) On Smoothing and Inference for Topic Models. UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, United States, pp 27–34

Biggs R, Carpenter SR, Brock WA (2009) Turning back from the brink: detecting an impending regime shift in time to avert it. Proc Natl Acad Sci U S A 106:826–831, doi:10.1073/pnas.0811729106

Blei D (2012) Probabilistic topic models. Commun ACM 55:77–84, doi:10.1109/MSP.2010.938079

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1:17–35

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Blevins C (2010) Topic Modeling Martha Ballard's Diary. In: Pers. Blog. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/. Accessed 2 May 2014

Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59:977–984, doi:10.1525/bio.2009.59.11.9

Butt N, Slade E, Thompson J, Malhi Y, Riutta T (2013) Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecol Appl 23:936–943, doi:10.1890/11-2059.1

European Environment Agency (2011a) Forests, health and climate change - Urban green spaces, forests for cooler cities and healthier people. http://www.eea.europa.eu/publications/forests-health-and-climate-change

European Environment Agency (2011b) Europe's forests at a glance — a breath of fresh air in a changing climate. http://www.eea.europa.eu/publications/europes-forests-at-a-glance

Evans K, Guariguata MR (2008) Participatory Monitoring in tropical forest management - a review of tools, concepts and lessons learned. Center for International Forestry Research (CIFOR), Bogor, Indonesia, p 50

Ferster CJ, Coops NC (2014) Assessing the quality of forest fuel loading data collected using public participation methods and smartphones. Int J Wildl Fire doi:10.1071/WF13173

Fritz S, McCallum I, Schill C, Perger C, Grillmayer R, Achard F, Kraxner F, Obersteiner M (2009) Geo-Wiki.Org: the use of crowdsourcing to improve global land cover. Remote Sens 1:345–354, doi:10.3390/rs1030345

Fritz S, See L, van der Velde M, Nalepa RA, Perger C, Schill C, McCallum I, Schepaschenko D, Kraxner F, Cai X, Zhang X, Ortner S, Hazarika R, Cipriani A, Di Bella C, Rabia AH, Garcia A, Vakolyuk M, Singha K, Beget ME, Erasmi S, Albrecht F, Shaw B, Obersteiner M (2013) Downgrading recent estimates of land available for biofuel production. Environ Sci Technol 47:1688–1694, doi:10.1021/es303141h

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci U S A 101(Suppl):5228–5235, doi:10.1073/pnas.0307752101

McCallum AK (2002) MALLET: A Machine Learning Language Toolkit. http://mallet.cs.umass.edu/

Millenium Ecosystem Assessment (2005) Ecosystems and Human Well-Being: Synthesis (Millennium Ecosystem Assessment Series). p 160

Newman G, Wiggins A, Crall A, Graham E, Newman S, Crowston K (2012) The future of citizen science: emerging technologies and shifting paradigms. Front Ecol Environ 10:298–304, doi:10.1890/110294

Roman LA, Mcpherson EG, Scharenbroch BC, Bartens J (2013) Identifying common practices and challenges for local urban tree monitoring programs across the United States. Arboric Urban For 39:292–299

Sachs S (2008) Using Students to Monitor the Effects of Ground-level Ozone on Plants. In: Weber, Samantha, and David Harmon (ed) Rethinking Protected Areas in a Changing World: Proceedings of the 2007 GWS Biennial Conference on Parks, Protected Areas, and Cultural Sites. The George Wright Society, Hancock, MI, pp 277–279

See L, Comber A, Salk C, Fritz S, van der Velde M, Perger C, Schill C, McCallum I, Kraxner F, Obersteiner M (2013) Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS One 8:e69958, doi:10.1371/journal.pone.0069958

Shirk JL, Ballard HL, Wilderman CC, Phillips T, Wiggins A, Jordan R, McCallie E, Minarchek M, Lewenstein BV, Krasny ME, Bonney R (2012) Public participation in scientific research: a framework for deliberate design. Ecol Soc 17:art29, doi:10.5751/ES-04705-170229

Silvertown J (2009) A new dawn for citizen science. Trends Ecol Evol 24:467–471, doi:10.1016/j.tree.2009.03.017

Smith AM, Lynn S, Sullivan M, Lintott CJ, Nugent PE, Botyanszki J, Kasliwal M, Quimby R, Bamford SP, Fortson LF, Schawinski K, Hook I, Blake S, Podsiadlowski P, Jönsson J, Gal-Yam A, Arcavi I, Howell DA, Bloom JS, Jacobsen J, Kulkarni SR, Law NM, Ofek EO, Walters R (2011) Galaxy zoo supernovae. Mon Not R Astron Soc 412:1309–1319, doi:10.1111/j.1365-2966.2010.17994.x

Steyvers M, Griffiths T (2007) Probabilistic topic models. In: Landauer T, McNamara D, Dennis S, Kintsch W (eds) Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum, pp 427–449

Templeton C, Brown T, Bhattacharyya S, Boyd-Graber J (2011) Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus. Chicago Colloquium on Digital Humanities and Computer Science 2011, Chicago, IL

Wallach H, Mimno D, McCallum A (2009a) Rethinking LDA: Why Priors Matter. In: Advances in Neural Information Processing Systems 22, NIPS 2009 Proceedings. 23rd Annual Conference on Neural Information Processing Systems 2009, Vancouver, British Columbia, Canada, Proceedings of a meeting held 7-10 December 2009

Wallach HM, Murray I, Salakhutdinov R, Mimno D (2009b) Evaluation Methods for Topic Models. Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML '09. ACM Press, New York, New York, USA, pp 1–8

Yang T-I, Torget AJ, Mihalcea R (2011) Topic modeling on historical newspapers. In: LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, Stroudsburg, PA, pp 96–104

doi:10.1186/s40663-014-0011-6
Cite this article as: Daume *et al.*: Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling. *Forest Ecosystems* 2014 **1**:11.