AGILE: GIScience Series, 5, 33, 2024. https://doi.org/10.5194/agile-giss-5-33-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.
© Author(s) 2024. This work is distributed under the Creative Commons Attribution 4.0 License.

Locality and transferability: examining pre-built lexicons to elicit landscape values from natural language

Inhye Kong [1], and Ross S. Purves [1]

[1] Department of Geography, University of Zurich, Zurich, Switzerland

Correspondence: Inhye Kong (inhye.kong@geo.uzh.ch)

Abstract. Landscape values are perceived through complex interactions between people and their surroundings, and understanding such values is essential for policy. Previous research to explore landscape values through text analysis often relies on developing lexicons, which serve as value classification rules. However, it is unclear how transferrable such lexicons are between locations of differing environmental and cultural conditions.

In this work, we examine the transferability of lexicons from previous studies: one on based on Geograph data to contain natural language in the UK and another on based on TripAdvisor in the context of US national parks. Both lexicons have typologies for 1) attractiveness/aesthetics of landscape and 2) natural elements/mammal species/biological values of landscape. We apply these lexicons to a text corpus built with the Guardian's Country Diary.

Our initial findings were spatial distributions of lexical matches, along with match ratios and keywords at county level in the UK. Then we zoomed in to compare the lexical performance to learn that larger lexicons do not guarantee better results when the context mismatches. Indeed, there is a room for transferability of lexicons; nonetheless, it is crucial to acknowledge that lexicons are sensitive to locality, urging to consider site-specific biophysical and sociocultural difference in applying pre-built lexicons.

Keywords. landscape values, landscape character, natural language processing, text analysis, newspaper

1 Introduction

Landscape values are perceived through complex interactions between people and their surroundings. When people are in the landscape, they can have a variety of experiences that lead to the formulation of relational and instrumental values (IPBES, 2022). Understanding the public perception of landscape serves as a foundation for successful policymaking and planning. However, capturing a full picture of public perception towards landscape and its values has been a difficult task, majorly due to the lack of relevant data or time-consuming data collection processes. The complexity in which landscapes are valued pose another challenge, manifested in the subjectivity of landscape perception, diversity of interests and value systems, and ambiguity in natural language.

With digital transformation and significant advances in natural language processing, recent studies have been taking advantage of large-scale computational analysis to extract diverse landscape values from text data. One popular approach in contemporary landscape semantic research is to work with the sources of natural language content, such as social media and unstructured text archives, coupled with computational text analysis. A number of case studies have extracted landscape values by developing rule-based classification schemes, namely lexicons, to match texts to landscape values. For instance, Hale, Cook, & Beltrán (2019) annotated Flickr tags into 11 types of cultural ecosystem services in riverain landscapes of Idaho, U.S., and Chen et al. (2020) took Instagram posts to annotate a subset of texts into seven cultural ecosystem services in Nova Scotia, Canada. Ebner, Schirpke, & Tappeiner (2022) combined Flickr tags, online surveys, and stakeholder interviews to

identify and compare cultural ecosystem services in mountain lakes of South Tyrol.

Still, little effort has been made to accumulate such knowledge in a structured manner, nor examine to what extent the lexicons are transferable and applicable to other locations. Such as discussion is relevant to methodological reproducibility and transferability in science and big data community, to assist soft landing of following research by offering an off-the-shelf package. Indeed, the lexicons are attuned to different domains of interest (e.g., tourism) and local context, such as biophysical (e.g., flora and fauna), sociocultural (e.g., accessibility), and institutional settings (e.g., protected areas). Nonetheless, they may share common ground in which values people perceive (i.e., beauty, recreation).

In this study, we examine the locality and transferability of two sets of lexicons from previous studies. They both dealt with two aspects of landscape values (aesthetics/attractiveness, elements/species/biological), but were built on different data and contexts. For our analysis, we prepared a freshly acquired text corpus from the Guardian newspaper column, Country Diary. As such, we compare spatial distributions of popular locations for two aspects of landscape values, reveal keywords corresponding to them, and examine the extent of lexicon transferability.

2 Methods

2.1 Data collection

Country Diary is a daily newspaper column published in the British newspaper, The Guardian, featuring essays about "the countryside and nature". Country Diary fits our analysis since the content describe nature-oriented experiences. We acquired articles from the Guardian API ('*guardianapi'*, Odell, 2019) in R environment, using a tag "environment/series/country-diary". The temporal range for the query was set from 1990-01- 01 to 2023-11-15, but the oldest article available on the digital archive was 1999.

We retrieved 7,410 articles which were then screened to exclude historic archives (i.e., texts containing a phrase "originally published in") as well as erroneous retrievals (e.g., "obituary") using a string match and manual screening (n = 7,308). To assess the integrity of data, we extracted year and month of publications from 'web_publication_date' and plotted the number of monthly publications.

2.2 Locating the article

Country Diary generally follow a form that presents main place names at the head of the article, but specific locations vary over time or among authors. To define the main place names of articles, we chose five entries from the API search results (i.e., 'web_title', 'headline', 'trail_text', 'standfirst', first ten words in 'body_text'), and processed these using with the natural language processing (NLP) tool spaCy with a pre-trained language model for English ("en_core_web_trf").

Here, we considered two approaches to extract place names: 1) we applied named entity recognition and filtered the entity results for 'GPE' (geopolitical entities), 'LOC' (locations), and 'FAC' (facilities); 2) since we found that NER often failed to capture some entities we extracted consecutive proper nouns (i.e., PROPN) and tokens following the pattern, PROPN+ADP(of)+PROPN (e.g., 'Isle of Man'); in this procedure, we cleaned the author names from the metadata 'byline', since persons names are also proper nouns.

Candidate place names were then sent to Google Geocoding API, namely '*ggmap'* in R (Kahle & Wickham, 2013), to fetch geolocations (i.e., latitude, longitude). From the results, 'continent' and 'administrative area level 1' in the variable "type" were removed given their coarse geographic granularity. Likewise, 'Britain' and 'UK' in place name entries were removed. At this point, the number of candidate place names was 9700+ whereas the number of articles was 7,308.

As a last step to geolocate the articles, we took spatial boundaries of Counties and Unitary Authorities Boundary published in December 2022 (hereafter, CTYAU22) from Open Geography Portal, Office for National Statistics (geoportal.statistics.gov.uk). Of course, '*ggmap'* returned precise geolocations; however, we found that taking a county-level boundary fits to the spatial coverage of our text corpora, which can be of a small-scale location or a broader region. Of all levels of administrative boundaries, we chose CTYAU22 to offer optimal spatial granularity for visualization. Candidate place names for each article were assigned to the boundary using the R package '*sf*' (Pebesma & Bivand, 2023). Of 9700+ entries, 8200+ entries corresponded to valid CTYAU22 attributes (84.8%), and most of the remaining entries were valid locations abroad (i.e., Country Diary sometimes cover overseas places) or falsefully retrieved locations abroad (i.e., "Norfolk", which is semantically meant to be a county in the U.K. according to our text corpus, returned a city in Virginia, U.S. from '*ggmap'*). Then we aggregated the CTYAU22 codes for each article, and kept articles with a single CTYAU22 code, leaving 5,737 articles or 78.5% of the original set.

2.3 Identifying landscape values

To compare the locality and transferability of lexicons, it is important to find comparable lexicons that cover similar typologies in landscape values. In this study, we took lexicons from two publications: a) one was built from Geograph data to analyse attractiveness (word pairs), and natural elements and mammal species (single words) (Koblet & Purves, 2020); b) another was built upon TripAdvisor to analyse phrasal expressions for eight cultural ecosystem services, including aesthetic and biological values (Kong et al., 2023). We grouped the typologies into 1) attractiveness (Geograph 1, namely G1) /aesthetic values (TripAdvisor, namely T1), and 2) natural elements(G2)/mammal species(G3)/biological values (T2). The Geograph lexicon was derived in the context of the UK, as same as our test text corpora, but the total number of entries was small (G1: 175, G2: 36, G3: 45); on the other hand, the TripAdvisor lexicon was built on different biophysical and sociocultural context (i.e., national parks in the United States), but comprised of much larger numbers of terms (T1: 1200+, T2: 800+).

To apply the lexicons, we took the body text of the corpus (i.e., 'body_text') and treated with basic NLP using the same model as earlier ("en_core_web_trf"). Since the model returns the lemmatized tokens in American English, we added a version in British English using R package '*uk2us'* (Davies, 2021). Also, a selective set of stop words was deleted. In the end, pattern matching was applied to detect lexical items in the NLP-treated text corpus.

The results were plotted on maps to show the ratio of articles containing the lexical entries out of the total number of articles locating at the CTYAU22 level (i.e., in colour spectrum). Here, we used ratios than total numbers of articles, since the number of articles were strongly biased to certain areas (see Figure 2 below). On top of that, the most frequent lexical items to the corresponding CTYAU22 were overlaid to the map. Text sizes corresponded to the normalized values of the occurrences.

2.4 Data and Software Availability (DASA)

Abovementioned procedures were all conducted in R version 4.3.2. The code lines can be found in GitHub link (http://github.com/ihKong/lexicon_landscapevalues).

3 Results

3.1 Corpus overview

The monthly count of Country Diary corpus is plotted as Figure 1. As the article is published daily-basis, the monthly count should be approximately 30. In the early days of digital archiving, such as 1999 and 2000, the numbers fell short below 30, but they have stabilized since 2001 (with the exception of January and February 2002 and February and March 2010).

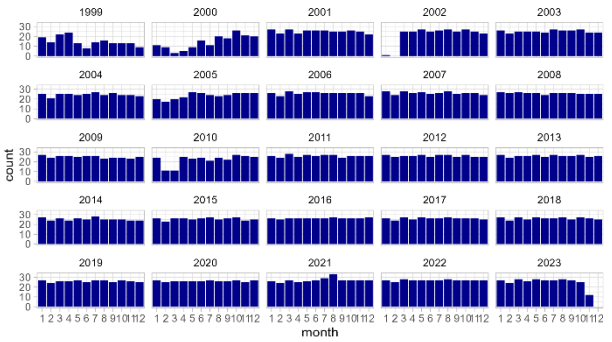

Figure 1. Number of monthly publications from 1999 to 2023

Then we aggregated the number of valid articles for CTYUA22 boundary. The result indicated the media bias to certain areas (Table 1, Figure 2): Shropshire ("Wenlock Edge (, Shropshire)" (508+182)) was covered the most, followed by Cumbria ("(The) Lake District" (191 + 124)) and Northumberland ("Northumberland" (269), "Allendale, Northumberland" (108)).

Table 1. Article counts per CTYUA22 boundary (table)

	CTYUA22CD	CTYUA22NM	count
1	E06000051	Shropshire	700
2	E10000006	Cumbria	440
3	E06000057	Northumberland	411
4	S12000017	Highland	283
5	E10000020	Norfolk	282
6	E10000027	Somerset	271
7	E10000014	Hampshire	259
8	E06000056	Central Bedfordshire	257
9	E06000047	County Durham	228
10	E10000007	Derbyshire	204

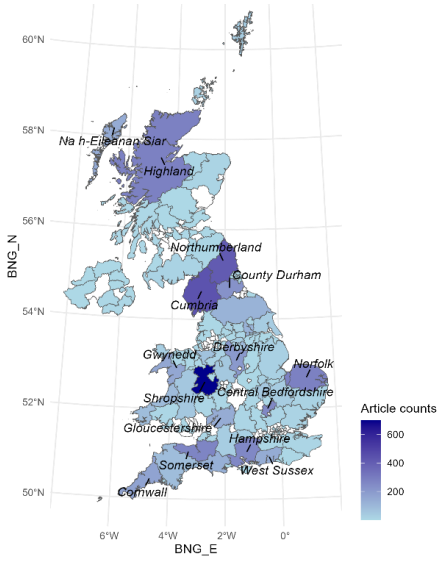

Figure 2. Article counts per CTYUA22 boundary (map)

3.2 Attractiveness/Aesthetics value

From the comparison between the attractiveness (G1) and aesthetic values (T1) lexicons, the results returned comparable performance on the map (Figure 3 and 4) and scatter plot (Figure 8, top), although the number of lexical entries in attractiveness (G1) was much smaller than the counterpart. The keywords representing the regions, however, were different: while aesthetic lexicon (T1) returned 'blue sky' to be the most popular word across most counties, attractiveness lexicon (G1) displayed more variety of keywords, such as 'coastal path' in Cornwall and southern Wales, and 'brown trout' in Highland.

3.3 Natural elements, Mammals/Biological values

The results from the lexicons of natural elements (G2), mammals (G3), and biological values (T2) are summarized in Figure 5-7. Lexical items for natural elements (G2) were found in almost every article corresponding to the area (> 75%, darkest green colour). The lexicons for mammals (G3) and biological values (T2) returned comparable results, despite some differences shown in spatial distribution as well as scatter plot (Figure 8, bottom). Moving on to the keywords, the most frequent words for natural elements (G2) was 'tree' in most of England, while 'rock' and 'water' stood out in Cumbria and Highland, respectively. Keywords for mammals (G3) featured interesting perception of species, such as 'sheep' in Cumbria and Somerset, 'dog' in Shropshire and Northumberland, and 'deer' in Highland. For the lexicon of biological values (T2), 'natural reserve' was most common, followed by 'small bird' which was

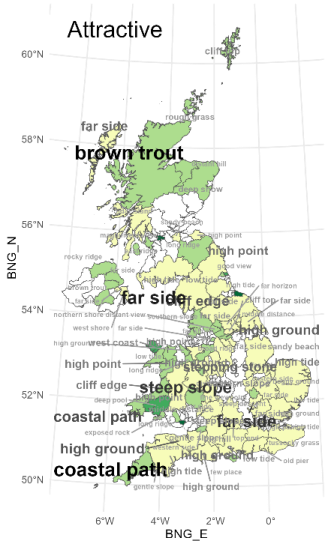

Figure 3. Distribution of lexicon matches for Attractiveness (G1)

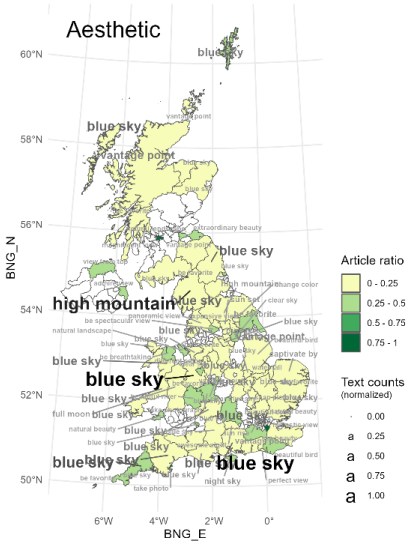

Figure 4. Distribution of lexicon matches for Aesthetic value (T1)

Figure 5. Distribution of lexicon matches for Natural elements (G2)

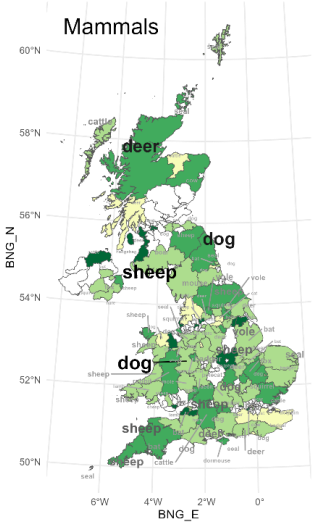

Figure 6. Distribution of lexicon matches for Mammals (G3)

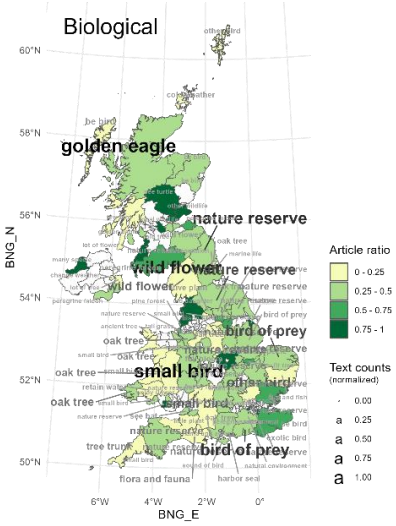

Figure 7. Distribution of lexicon matches for Biological values (T2)

captured in Shropshire the most. Highland was mentioned the most with ‘golden eagle’, whereas County Durham was with ‘wild flower’.

4 Discussion

The aim of this work was to examine the extent of transferability of landscape value lexicons from earlier works with a new corpus. The relevance of this work is to explore the possibility of creating a common ground for accumulating and sharing knowledge, at a time when concerns about scientific transferability are growing. Particularly for landscape value research, transferring lexicons can benefit researchers to cut time and effort for generic and universal values as a rapid off-the-shelf analysis package, while allowing researchers to focus on unravelling site-specific values.

The findings indicated that there is a room for transferability, since the lexicons showed decent returns for a new text corpus. The important messages here, however, are twofold: 1) the composition of lexical entries largely poses strong influence on the results, as we witnessed different keywords for the same regions, and 2) larger lexicons do not guarantee better results when the context mismatches. Indeed, lexicon building is subject to the locality, which often involves different biophysical and sociocultural context. In our case, the lexicons from TripAdvisor comprised of much larger lexical entries, but their context was of protected areas (i.e., national parks), whereas Geograph lexicon was not bound to certain types of landscapes. Moreover, different flora and fauna, landscape characters, activities people engage in, must have played a role in structuring lexical entries. For example, ‘(blue) sky’ was popular in the lexicons from TripAdvisor, whereas it was missing in Geograph; ‘gentle slope’ was popular in Geograph, whereas it was missing in TripAdvisor. Another deliverable in this study was to urge to consider different domains of value schemes that can be captured from different text corpora. For instance, the most common words from Country Diary (i.e., ‘bird’, ‘flower’, ‘garden’, ‘hedge’) were missing in Geograph-based lexicons, although they both were pertinent to the context of landscape in the U.K.

Technical challenges also remain. The most critical one lies on applying the lexicon with text matching, which exclusively applies to exact matches. It is rather straightforward when it comes to a word-to-word matching, but it gets complicated for phrasal matching unable to detect tokens in flipped orders (e.g., ‘garden of flowers’ will not be found with a lexical entry ‘flower garden’). Word-level text matching contains challenges as well, when identical semantics have different expressions, not just spelling (i.e., ‘color’ vs ‘colour’) but word usages, such as ‘crag’ (≈‘cliff’), ‘loch’ (≈ ‘lake’). This could be improved by using skip-gram matching or lexicon enrichment through word embedding.

The bias within text corpus cannot be overlooked. We found geographic preference of Country Diary towards certain locations, such as Wenlock Edge and Lake District, which may imply the narrative of representable countryside for newspaper publication. At least in our analysis, such a bias resulted in overstating the semantics of a few places, which had to be subdued with calculating ratios. Geolocating may have also influenced the results, such as Google Geolocation API to return NA some entity names, such as “Lake District”. Lake District was the only case we intervened to specify ‘Lake District National Park’ to fetch the geolocation (due to the popularity of the name in our corpus), but there were more entities that failed to get geolocations, especially the cases of colloquial names or names more popular outside of the U.K. (e.g., Tamar Valley is found both in Dartmoor, U.K. and Tasmania, Australia). To improve the accuracy, toponym archives, such as Geonames, can be supplemented to consider spatial hierarchies as well as alternate names.

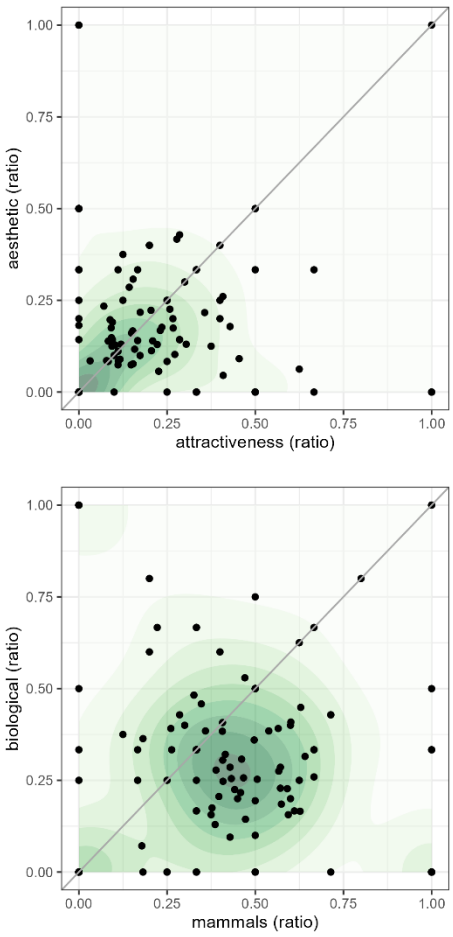

Figure 8. Scatter plots for comparing attractiveness-aesthetic value lexicons (top), and mammals-biological value lexicon (bottom)

5 Conclusion

Rich archives of natural language (i.e., unstructured text corpora) offer a great potential to elicit diverse dimensions of landscape values. Previous studies have accumulated valuable knowledge from a variety of data to distinguish landscape values, and such outcomes are encapsulated in lexicons. In this study, we examined the transferability of earlier lexicons to a new text corpus from news media, especially the one curated for landscape-oriented writings.

Overall, we confirmed a room for transferability to extrapolate the lexicon to a new text corpus, but it is important not to overlook the locality of landscape, which is not fully captured with a simple lexicon transfer. In the end, landscape semantic research needs to take advantage of existing lexicons with has a strong potential for transferability, but also requires comprehensive effort to fill the semantic gaps in landscape values from different text sources as well as local context. Essentially, values are subjective and context-dependent.

References

Chen, Y., Caesemaecker, C., Rahman, H. T., & Sherren, K. (2020). Comparing cultural ecosystem service delivery in dykelands and marshes using Instagram: A case of the Cornwallis (Jijuktu'kwejk) River, Nova Scotia, Canada. Ocean and Coastal Management, 193(November 2019), 105254. https://doi.org/10.1016/j.ocecoaman.2020.105254

Davies B (2021). _uk2us: Convert Words Between UK and US English_. R package version 0.1.0, <https://CRAN.R-project.org/package=uk2us>.

Ebner, M., Schirpke, U., & Tappeiner, U. (2022). Combining multiple socio-cultural approaches – Deeper insights into cultural ecosystem services of mountain lakes? Landscape and Urban Planning, 228(April), 104549. https://doi.org/10.1016/j.landurbplan.2022.104549

Hale, R. L., Cook, E. M., & Beltrán, B. J. (2019). Cultural ecosystem services provided by rivers across diverse social-ecological landscapes: A social media analysis. Ecological Indicators, 107(July), 105580. https://doi.org/10.1016/j.ecolind.2019.105580

IPBES (2022). Summary for Policymakers of the Methodological Assessment Report on the Diverse Values and Valuation of Nature of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Pascual, U., Balvanera, P., Christie, M., Baptiste, B., González-Jiménez, D., Anderson, C.B., Athayde, S., Barton, D.N., Chaplin-Kramer, R., Jacobs, S., Kelemen, E., Kumar, R., Lazos, E., Martin, A., Mwampamba, T.H., Nakangu, B., O'Farrell, P., Raymond, C.M., Subramanian, S.M., Termansen, M., Van Noordwijk, M., and Vatn, A. (eds.). IPBES secretariat, Bonn, Germany. https://doi.org/10.5281/zenodo.6522392

Kahle, D. and H. Wickham, (2013). ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

Koblet, O., & Purves, R. S. (2020). From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally. Landscape and Urban Planning, 197(February), 103757. https://doi.org/10.1016/j.landurbplan.2020.103757

Kong, I., Sarmiento, F. O., & Mu, L. (2023). Crowdsourced text analysis to characterize the U.S. National Parks based on cultural ecosystem services. Landscape and Urban Planning, 233(June 2022), 104692. https://doi.org/10.1016/j.landurbplan.2023.104692

Odell, E. (2019). _guardianapi: Access the 'Guardian' newspaper open data API_.

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446.