Identifying Cultural Resource Hotspots via Crowdsourcing and Expert Perspectives

Madeline Brown[1], Changjie Chen[2], Luwei Wang[2], Timothy Murtha[2]

[1]University of Maryland, Maryland/USA · mtbrown@umd.edu
[2]University of Florida, Florida/USA

Abstract: Prioritizing hotspots for cultural resource conservation remains a critical challenge for conservation design and planning on a large-landscape scale. Here, we propose novel methodologies for integrating expert evaluations of cultural resource distribution, diversity, and priorities with an emerging digital geospatial dataset tracing the movement of people to and from sites of interest (SafeGraph). Methods for data processing and analysis are explained in detail and code is shared to promote iterative research combining big data with small-n data. Here, we demonstrate the scope and potential of these methodologies by evaluating the correspondence and dynamics between expert and crowdsourced datasets via hotspot analysis and temporal visitation patterns. These emergent crowdsourced metrics of valuation (e. g. number of visitors and time spent in particular cultural landscapes/sites) can be contextualized and evaluated in combination with expert assessments of resource prioritization to better understand landscape values across a city-wide spatial extent.

Keywords: Cultural resources, spatial-temporal analysis, crowdsourcing, hotspot analysis, Washington DC

1 Introduction

Assemblages of human and nonhuman elements, processes, and actors collectively generate particular "types" of places or ecotopes (TRUSLER & JOHNSON 2008), which may be variably understood as cultural or hybrid landscapes. Some of these places are actively designed, while others are the emergent results of uncoordinated, but patterned activities. As these hybrid spaces are co-constructed − both physically and symbolically − by those who use them, their importance, form and meaning remains productively dynamic. Managing, evaluating and planning for the future of important cultural resource sites and landscapes requires a hybrid approach: incorporating histori c information, contemporary use patterns, and spatial analysis. Here, we propose a novel methodology for integrating expert evaluations of cultural resource distribution, diversity and priorities with emerging digital geospatial datasets tracing the movement of people to and from cultural sites of interest.

Prioritizing cultural resource management and preservation remains a challenge for land managers and planners, who may be working with limited budgets and personnel to navigate regulatory and scientific imperatives, as well as meet the expectations of the general public. Increasingly, landscape architecture researchers and planners are turning to crowdsourced social media and other digital tools to understand public preferences for and visitation rates to cultural resource sites, points of interest, and green spaces (ROBERTS 2017, TENKANEN et al. 2017, GOLDBERG et al. 2018, HAMSTEAD et al. 2018). One novel tool with potential to inform landscape-scale cultural resource management is SafeGraph, which offers an open-source platform to maintain a comprehensive dataset of point of interest (POI) for places globally that is collected using GPS signals from smartphone applications.

Journal of Digital Landscape Architecture, 7-2022, pp. 155-163. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-724-3, ISSN 2367-4253, e-ISSN 2511-624X, doi:10.14627/537724016.

Across a regional landscape scale, cultural and natural resources exhibit a patchy, patterned distribution. Increasingly, conservation scientists and practitioners are recognizing the need to address environmental challenges at multiple scales rather than within bounded protected areas or sites. This approach is not limited to natural resource conservation, as archaeological and cultural resource sites also benefit from a landscape-scale perspective (DOELLE et al. 2016). Focus on these multiple scales is sometimes correlated with the topic of urban growth, where the surrounding areas of cultural resources are converted into urban lands. Urbanization has become the critical context in understanding cultural and natural resources as part of coupled human-natural systems (MCDONALD et al. 2008). This perspective shift can be summarized as a switch from fortress conservation models to landscape conservation design or landscape mosaic models, wherein different patches of human and nonhuman habitats embody a spectrum of conservation values (PERFECTO et al. 2009, LEONARD et al. 2018). Thinking in terms of assemblages, hybrid landscapes, and anthropogenic mosaics also has potential to transform how we assess cultural resource landscapes as part of broader social-ecological systems, yet the science of how to effectively identify priority hotspots and functional diversity remains underdeveloped. Here, we argue that big data and crowdsourced data will be critical tools for advancing the science of cultural resource hotspot identification and the resulting establishment of management priorities.

One particularly useful data source for assessing site visitation is the Patterns dataset from SafeGraph[1], a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places. SafeGraph offers high-quality Points of Interest (POI) data via three primary datasets: *Core Places*, *Geometry*, and *Patterns* (SAFEGRAPH 2021a). Specifically, the Patterns dataset contains "daily visit count" for over 4.5 million POIs based on GPS signals from over 45 million mobile devices, among which each observation is associated with a unique but anonymized user ID (SQUIRE 2019, YABE et al. 2020). To enhance privacy, SafeGraph excludes census block group information if fewer than two devices visited an establishment in a month from a given census block group. Although SafeGraph only aggregates data from approximately 10% of devices in the United States, it maintains a representative sample across census demographics (e. g., education, ethnicity, gender, income) at the county-level (O'DONOGHUE et al. 2021).

The SafeGraph pattern datasets also provide information about daily count of visitors of each individual POIs. This dataset is distinctive as a tool for mapping visitation, as unlike intercept surveys which may only capture visitor demographics at particular times or places, these data allow for simultaneous analysis of visitors at multiple timescales and any location within the site as a whole, making it less vulnerable to sampling biases. Despite the potential utility of these data, aggregated visitation data from sources like SafeGraph also exhibit limitations in terms of a clear data processing and analysis pipeline to link them with the nuances of cultural resource management in particular social and geographic contexts. Effectively integrating big data with qualitative assessments of cultural resource priorities and values from land managers and practitioners has the potential to improve landscape conservation design and planning and reduce costs of assessing site importance or functions. Moreover, such approaches may identify gaps in understanding or values between stakeholders, including public visitors, civic groups, and government agencies.

[1] https://www.safegraph.com/

2 Methods

In this paper, we combine and assess two primary datasets: 1) point-of-interest data from SafeGraph and 2) cultural resource prioritization data from an online survey of cultural resource practitioners and experts. The latter dataset was collected using a survey designed with cultural domain analysis methods from cognitive anthropology (BORGATTI 1998, QUINLAN 2005, BERNARD 2006), distributed to individuals working in cultural resource management in Washington DC, USA and the surrounding region. Cultural domain analysis supports the identification of cognitive or cultural *domains*, that is, shared concepts, knowledge or beliefs within a particular cultural group. This methodology shows promise for landscape conservation design and planning due to the ability to differentiate boundaries between domains as well as core and peripheral ideas/entities within a single domain.

Survey respondents were asked to freelist up to ten categories of cultural resources and up to ten specific tangible and intangible cultural resources (including sites, practices, events, etc.) that fit within each category. Based on these responses, we assembled a regional dataset of cultural domain expert perceptions of the distribution and diversity of cultural resources. Therefore, our definition of cultural resources is emergent based on compiling numerous specialist definitions, but includes scenic vistas, museums, historic sites and buildings, roadways, landscapes, parks and natural areas, among other sites. Data analysis was conducted with R (packages include: AnthroTools and tidyverse (R CORE TEAM 2021, PURZYCKI & JAMIESON-LANE 2016, WICKHAM et al. 2019), ArcGIS, and Python.

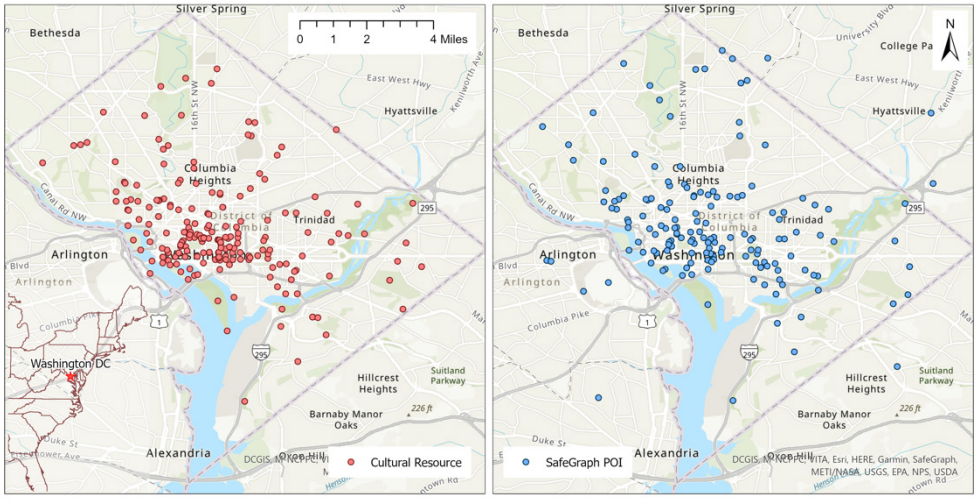

Fig. 1: Left: Cultural resources in Washington DC identified by specialists. Right: Matching records found in SafeGraph core places.

2.1 Methods for Matching SG (SafeGraph) and CRP (Cultural Resource Prioritization) Datasets

Matching the two datasets was a crucial part of this study for the purpose of comparing expert opinions and public visitation patterns. This multi-step process is further documented in a publicly available Github code repository (CHEN 2022). First, the coordinates, i. e., latitude/longitude, of the CRP dataset were geolocated using R. Second, the coordinates and names of each location were queried against the Core Places dataset using the SafeGraph *Places API* (SAFEGRAPH 2021b). If a record was found, such query returns a placekey, a unique identifier for SafeGraph POIs, as well as other pre-selected information including location name, category, coordinates, and brand association. Note that the Places API conducts an ad hoc query and returns the "most likely" candidate in the SG database, which does not guarantee an exact match. As shown in Figure 1, automatic detection is not 100% accurate due to the discrepancies in names and coordinates between the two datasets. In other words, a CRP location used in a query may yield a completely unrelated record in SG.

To further check whether two matched records indeed refer to the same physical location, we first calculated the *Levenshtein distance* (Eq. 1), also known as edit distance, between the CRP names and the SG names (LEVENSHTEIN 1966). The Levenshtein distance measures the minimum edits (insertion, deletion, and replacement of characters) that one needs to make to change one word to the other.

$$\boldsymbol{D}_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if } min(i,j) = 0 \\ min\begin{cases} \boldsymbol{D}_{a,b}(i-1,j)+1 \\ \boldsymbol{D}_{a,b}(i,j-1)+1 \\ \boldsymbol{D}_{a,b}(i-1,j-1)+\mathbf{1}_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases} \quad \text{(Eq. 1)}$$

Given two arbitrary words denoted as a and b with length I and J, Eq. 1 shows how the distance is calculated. The algorithm starts by initializing the first row and column of a distance matrix, $\boldsymbol{D}_{a,b}$. Then, the $(i,j)^{\text{th}}$ element of the matrix equals to the minimum of three cases, each one of which corresponds to a case of insertion, deletion, and a replacement, respectively. Note that the $\mathbf{1}_{(a_i \neq b_j)}$ is an indicator function, whose value equals 1 if only $a_i \neq b_j$, that is the leading i characters of a is not equal to the leading j characters of b. The Levenshtein distance divided by $max(i,j)$ returns a value between 0 and 1, which is a similarity ratio of the two measured words. In this study, we used a similarity ratio of 0.6 as a threshold to determine whether a CRP name and a SG name are matched.

Next, we measured the physical proximity between a CRP record and its corresponding record from SG based on the query method mentioned before. Since the coordinates of both datasets are in latitude/longitude, we applied the Haversine formula (Eq. 2), or great-circle distance, to calculate distances (SINNOTT 1984). In general, a smaller distance increases the credibility of a match between records from the two datasets. In this study, we considered a match inaccurate if two records have a Haversine distance greater than 500 meters.

$$d = 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right)+\cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \quad \text{(Eq. 2)}$$

3 Results

In total, 50 experts named 1,775 specific cultural resources that they perceived to exist within Washington DC and the surrounding region, 1,216 of which have a location that could be geocoded. These site-based cultural resources are the focus here. Resources named by experts include a range of tangible and intangible, historic and contemporary, monumental and natural, as well as other diverse resource dimensions. Subsetting to the DC area, there are 699 references to cultural resource sites, with 205 unique places based on their spatial coordinates (some sites are multi-component). This is referred to as the CRP dataset. These places were aggregated and analyzed for diversity and spatial hotspots.

3.1 Dataset Matching (and Mismatching)

Available SafeGraph data includes real-time visitation information for POIs globally. Subsetting the spatial extent of these data to the Washington DC region yielded 17,649 POIs. Of the 205 unique locations in the CRP dataset, the SafeGraph API detected 187 pairs of matches. After applying the criteria of Levenshtein distance (similarity ratio greater than 0.6) and Haversine formula (distance less than 500 meters), 104 out of the 187 pairs were confirmed as "exact" matches. By mapping their coordinates in GIS, the remaining 83 pairs were manually checked. Ultimately, 46 out of the 83 manual checks showed mismatches, with the remaining 37 points matched to particular locations in both datasets. In total, we found matches between 141 out of the 205 potential CRP places (66.78%).

Upon further examination, we found that mismatches and uncertainties primarily related to either: 1) multi-location "sites" or 2) districts or regions. For example, "Washington D.C. boundary stones" are noted in the survey dataset as cultural resources. Although these stones have clear physical locations, there are multiple stones located around the boundary of the city, making this a multi-sited cultural resource "site". Moreover, because these stones follow the boundary of this region, including them in the analysis might add additional noise rather than clarity. The second issue we identified during the matching process relates to historic districts, neighborhoods, or other larger regions of cultural significance. Because SafeGraph is framed in terms of *points* of interest, these data are not readily translatable into a polygonal or regional spatial analysis. Moreover, it is not always feasible or meaningful to identify a singular point to represent an entire district or region.

Evaluating the correspondence between SG and CRP datasets indicated a relatively high level of overlap between locations of interest. Therefore, replicating this matching methodology has potential to support rapid comparison between small-n social science surveys and studies and digital crowdsourced datasets when site names and/or spatial coordinates are known.

3.2 Spatio-Temporal Analysis of Cultural Landscapes

Given the temporal dimension of the SafeGraph data, it is possible to assess how visitation rates vary across cultural resources seasonally. To test this ability, we extracted data of a full calendar month, September 2021, from the Safegraph *patterns* dataset. We then aggregated daily visit counts by days of the week. As shown in Figures 2 and 3, visitation rates vary drastically across POIs. In other words, these data reveal discrepancies in the popularity of

different cultural places, as well as changes in visitation rates across days of the week. For example, Monday has the least number of total POI visits, whereas Thursday has the most.

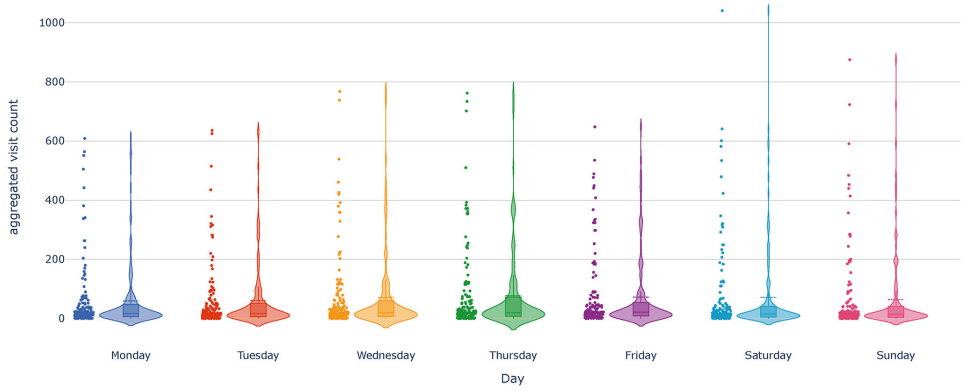

Fig. 2: Cultural resource site visitation counts and generalized trends by day of the week

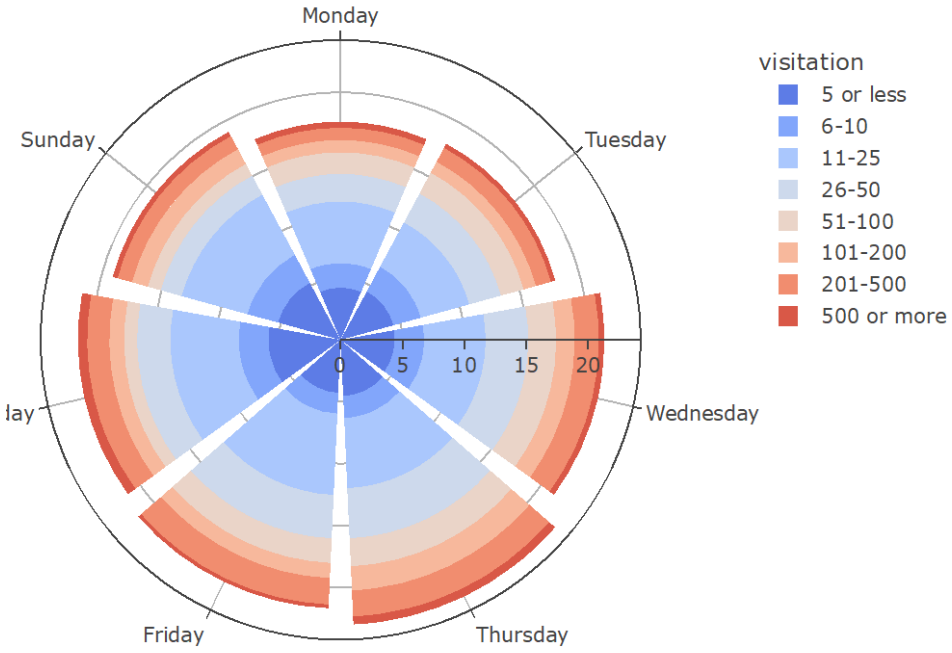

Fig. 3: Relative site visitation counts by day of the week. In this plot, visitation is binned into categories representing different levels of visitation intensity across POIs. The lengths of the seven "spokes" are stretched to reflect total number of visits by each day of week.

To analyze spatial trends, we adopted hotspot analysis via heatmap, which visualizes the intensified trend based on the number of visits to each site and the number of visitors (Figure 4). This analysis of hotspots compares cultural resource sites identified and valued by professional specialists and POI. Comparing these figures reveals differences in how popular or visited each site is versus the number of unique individuals visiting each site. Visitation is concentrated near the National Mall area of D.C. and its vicinity. The two figures align, though the intensity of hotspots on the left is greater, corresponding to expectations about repeat visitors to cultural resource sites increasing the overall visit counts. Maps included in this paper are intended to demonstrate the potential of matching diverse datasets, their implications for management and prioritization will be explored elsewhere.

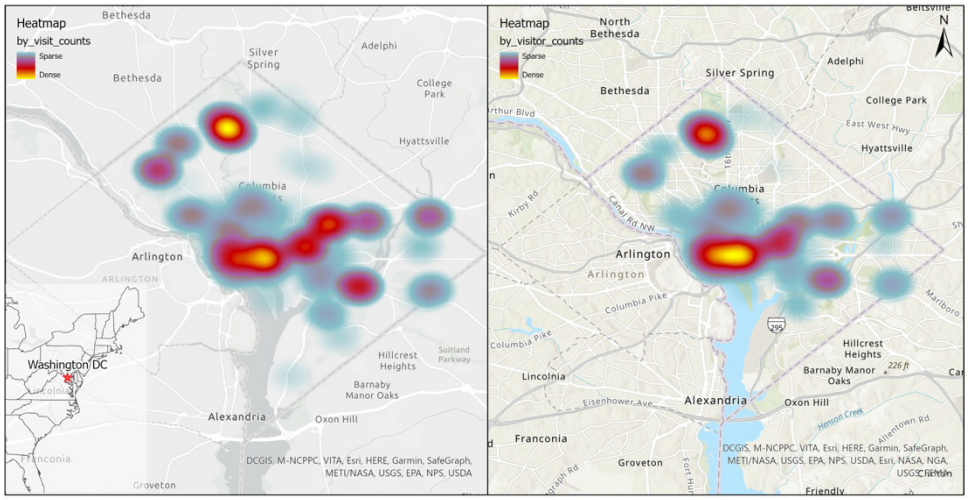

Fig. 4: Left: Hotspot analysis on heatmap by visit counts. Dense areas shown with yellow, revealing the most frequent count of places visited. Right: Hotspot analysis on heatmap by visitor counts. Indicated trends in visitor numbers in each unique location. Dark-colored areas indicate highest number of unique visitors.

4 Discussion and Conclusion

We find that matching crowdsourced data with expert perceptions of cultural resource sites is possible with a reproducible data processing and analysis pipeline. Based on automatic matching and manual checks, around two-thirds of cultural resource sites were matched to specific geographic points. After matching, these point data may be used in additional hotspot, management prioritization, and temporal analyses to further understand the use and importance of these sites of cultural significance. Careful attention to how matching is conducted will be critical for ensuring the reliability and relevance of aggregated crowdsourced data for local cultural resource management decision-making. We identify two avenues for further exploration in matching datasets: 1) handling multi-sited resources (e. g. boundary stones or a set of themed resources such as a particular type of statue or building); and 2)

regional or polygonal analysis (e. g. historic districts or neighborhoods). Because POI data from sources such as SafeGraph also include commercial establishments (e. g. restaurants, and shops), further methodologies require development to determine how to best incorporate these into analyses of cultural landscapes. Particularly when evaluating how "sense of place" might differ between a designated historic district and a neighboring district, incorporating the whole suite of diverse POIs might be informative for planning purposes.

Setting cultural and natural resource priorities across a large landscape may productively rely on identifying hotspots where variables are particularly dense or diverse. In the case of cultural resource management in the United States, hotspot analysis combining multiple dimensions (e. g. expert assessments, social media trends, and visitation data) may spark new forms of evaluating landscape-scale preservation priorities. Using novel big data sources, it will be possible to compare cultural resource specialists' values and management plans for important cultural resource sites with user visitation spatial and temporal patterns. Moreover, combining dynamic spatiotemporal data of visitation patterns to cultural sites with formal or informal management priorities and values has potential to support more rapid and low-cost decision-making compared to on-the-ground visitor studies. Further enhancing the efficiency of such analyses will be the increased adoption of reproducible methodologies and open-access sharing of code repositories. Finally, comparing metrics of site value may also reveal gaps between site use, equitable access, historic value, and management resource allocations (HAMSTEAD et al. 2018), when spatial visitation data are combined with demographic socio-economic data or other datasets. Here we present a reproducible methodology for the first step in collating multi-layered digital landscape-scale cultural resource datasets in order to develop more holistic and nuanced understandings of how cultural landscapes are valued and experienced across multiple populations.

References

BERNARD, H. R. (2006), Research Methods in Anthropology: Qualitative and Quantitative Approaches. 4th Ed. Altamira Press, Lanham, MD.

BORGATTI, S. P. (1998), Elicitation Techniques for Cultural Domain Analysis. In: SCHENSUL, J. & LECOMPTE, M. (Eds.), The Ethnographer's Toolkit, 3, 1-26. AltaMira Press, Walnut Creek, CA.

CHEN, C. (2022), Match SafeGraph POI data with Cultural Resource Places in Washington DC. Code repository. https://github.com/chjch/DLA2022_MatchSafeGraph.

COHN, J. P. (1988), Culture and conservation: a greater sensitivity to local culture could increase the success of both conservation and development projects. BioScience, 38 (7), 450-453.

DOELLE, W. H., BARKER, P., CUSHMAN, D., HEILEN, M., HERHAHN, C. & RIETH, C. (2016), Incorporating Archaeological Resources in Landscape-Level Planning and Management. Advances in Archaeological Practice, 4 (02), 118-131. doi:10.7183/2326-3768.4.2.118.

GOLDBERG, L. K., MURTHA, T. M. & ORLAND, B. (2018), The Use of Crowdsourced and Georeferenced Photography to Aid in Visual Resource Planning and Conservation: A Pennsylvania Case Study. In: GOBSTER, P. H. & SMARDON, R. C. (Eds.), Visual resource stewardship conference proceedings: landscape and seascape management in a time of change. Gen. Tech. Rep. NRS-P-183. U.S. Department of Agriculture, Forest Service, Northern Research Station, Newtown Square, PA, 116-126.

HAMSTEAD, Z. A., FISHER, D., ILIEVA, R. T., WOOD, S. A., MCPHEARSON, T. & KREMER, P. (2018), Geolocated Social Media as a Rapid Indicator of Park Visitation and Equitable Park Access. Computers, Environment and Urban Systems, 72 (November), 38-50. doi:10.1016/j.compenvurbsys.2018.01.007.

LEONARD, P. B., BALDWIN, R. F. & HANKS, D. (2017), Landscape-Scale Conservation Design across Biotic Realms: Sequential Integration of Aquatic and Terrestrial Landscapes. Scientific Reports, 7 (1). 14556. doi:10.1038/s41598-017-15304-w.

LEVENSHTEIN, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10 (8), 707-710.

MCDONALD, R. I., KAREIVA, P. & FORMAN, R. T. (2008), The implications of current and future urbanization for global protected areas and biodiversity conservation. Biological conservation, 141 (6), 1695-1703.

MULDER, M. B. & COPPOLILLO, P. (2005), Conservation: linking ecology, economics, and culture. Princeton University Press.

O'DONOGHUE, A., DECHEN, T., PAVLOVA, W., BOALS, M., MOUSSA, G., MADAN, M., THAKKAR, A., DEFALCO, F. J. & STEVENS, J. P. (2021), Reopening businesses and risk of COVID-19 transmission. Npj Digital Medicine, 4 (1), 1-5. https://doi.org/10.1038/s41746-021-00420-9.

PERFECTO, I., VANDERMEER, J. & WRIGHT, A. (2009), Nature's Matrix: Linking Agriculture, Conservation and Food Sovereignty. Earthscan, London.

PURZYCKI, B. G. & JAMIESON-LANE, A. (2016), AnthroTools: A Package in R. https://anthrotools.files.wordpress.com/2016/05/anthrotools_guide.pdf.

QUINLAN, M. (2005), Considerations for Collecting Freelists in the Field: Examples from Ethobotany. Field Methods, 17 (3), 219-234. doi:10.1177/1525822X05277460.

R CORE TEAM (2021), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

ROBERTS, H. V. (2017), Using Twitter Data in Urban Green Space Research: A Case Study and Critical Evaluation. Applied Geography, 81 (April), 13-20.

SAFEGRAPH (2021a), About SafeGraph. https://docs.safegraph.com/docs/about-safegraph.

SAFEGRAPH (2021b), Places API overview. https://docs.safegraph.com/reference/places-api-overview-new.

SINNOTT, R. W. (1984), Virtues of the haversine. Sky and Telescope, 68 (2), 158.

SQUIRE, R. F. (2019), What about bias in the SafeGraph dataset? https://www.safegraph.com/blog/what-about-bias-in-the-safegraph-dataset.

TENKANEN, H., DI MININ, E., HEIKINHEIMO, V., HAUSMANN, A., HERBST, M., KAJALA, L. & TOIVONEN, T. (2017), Instagram, Flickr, or Twitter: Assessing the Usability of Social Media Data for Visitor Monitoring in Protected Areas. Scientific Reports, 7 (1). doi:10.1038/s41598-017-18007-4.

TRUSLER, S. & JOHNSON, L.M. (2008), 'Berry Patch' as a Kind of Place − the Ethnoecology of Black Huckleberry in Northwestern Canada. Human Ecology, 36, 553-568.

WICKHAM, H., AVERICK, M., BRYAN, J. et al. (2019), Welcome to the Tidyverse. Journal of Open Source Software, 4 (43), 1686. https://doi.org/10.21105/joss.01686.

YABE, T., ZHANG, Y. & UKKUSURI, S. V. (2020), Quantifying the economic impact of disasters on businesses using human mobility data: a Bayesian causal inference approach. EPJ Data Science, 9 (1), 1-20. https://doi.org/10.1140/EPJDS/S13688-020-00255-6.