AGILE: GIScience Series, 5, 18, 2024. https://doi.org/10.5194/agile-giss-5-18-2024
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.

Use of iNaturalist Biodiversity Contribution Data for Modelling Travel Distances to Parks Across the United States

Jiping Cao[1], Hartwig H. Hochmair[1]

[1] Fort Lauderdale Research and Education Center, Geomatics Sciences, University of Florida, Florida, USA

Correspondence: Jiping Cao (jipingcao@ufl.edu)

Abstract. Crowdsourcing platforms have become an important data source for modelling and observing human behavioural and social activities, such as mobility, social interactions and urban dynamics. This study uses observation data from iNaturalist, an online social network of voluntary users sharing biodiversity information, which was collected from 20,434 parks in the United States. It explores the relationship between park characteristics and the mean travel distance of users to parks. The latter is based on the average of distances between an iNaturalist user's typical main area of iNaturalist contributions and the locations of the user's observations falling inside a park of interest. The DBSCAN clustering algorithm is used to determine each user's main contribution area. An Eigenvector Spatial Filtering (ESF) model shows that the log of the average distance travelled to parks is positively associated with certain park management types (e.g. National Parks, State Parks) and biodiversity, but negatively associated with the population around a park. The results provide insights into the nature of iNaturalist user visitation patterns to parks which can be used for targeted outreach campaigns and a more user-centric approach to promote park attractions and biodiversity conservation.

Keywords. Crowdsourcing, travel distance, park, iNaturalist

1 Introduction

1.1 iNaturalist

In the era of digital engagement, environmental consciousness, and citizen science, iNaturalist stands as a beacon for biodiversity enthusiasts, researchers, and casual nature observers alike. Launched in 2008, with over 190 million observations and 7.4 million registered users in February 2024[1], this crowdsourcing application has revolutionized the way its users perceive and interact with the natural environment. At its core, iNaturalist facilitates the recording and sharing of biodiversity observations, i.e., the uploading of photographs of flora and fauna, which are then identified by the community (Unger et al., 2021). This collaborative platform not only contributes to the vast database of biodiversity records but also fosters a global community passionate about nature and conservation (Mesaglio and Callaghan, 2021).

While the primary focus of iNaturalist data analyses has been on assessing biodiversity (Chandler et al., 2017), sampling bias (Callaghan et al., 2021), and observation data quality (Hochmair et al., 2020), the application harbours a wealth of data that extends beyond species identification with the potential to analyse spatial and temporal patterns of iNaturalist user participation. An example is the distinction between residents and short term visitors (Dimson and Gillespie, 2023). Unlike traditional biodiversity databases, iNaturalist's user-generated content offers a unique lens through which one can explore how individuals interact with the natural environment, including their movement patterns, the distance they travel to make observations, and their choice of contribution locations.

1.2 User travel behaviour extraction

Crowdsourced data, such as tweets (Jurdak et al., 2015), photographs (Ma et al., 2020), or travel reviews (Owuor et al., 2023) are commonly used to analyze travel behaviour.

[1] https://www.inaturalist.org/stats

However, the analysis of travel behaviour based on iNaturalist observations remains underexplored although the unique nature of its data, which is rooted in the pursuit of biodiversity observation, combined with spatio-temporal information, offers a novel perspective on spatial behavioural patterns. Specifically, it allows for the investigation of travel behaviour within the framework of recreational and educational activities related to biodiversity (Di Cecco et al., 2021). This presents an opportunity to understand not just where people go, but why they choose certain destinations, particularly natural parks and reserves, over others (Tu et al., 2020).

To narrow this research gap, this study aims to use iNaturalist observation data to explore the relationship between park attributes (such as served population, and management type) and their attractiveness to users, measured by mean travel distance of users to parks. It therefore offers new insights into the role of crowdsourced biodiversity data for enhancing our understanding of human-nature interactions. To avoid the influence of spatial

2 Methodology

2.1 Data collection

iNaturalist offers a comprehensive dataset that includes species names, geolocations, timestamps, and the names of contributors and identifiers of observations. The latter are crucial for validating species identifications. Each observation that is provided with a photo, location, and observation date is automatically placed into the "Needs ID" category. An observation becomes research grade when the community agrees on species-level ID or lower, i.e., when more than 2/3 of identifiers agree on a taxon. Research grade observation data can be downloaded from the Global Biodiversity Information Facility (GBIF) Website. For this study, we collected iNaturalist research grade observations from across the United States covering the period from August 2022 to August 2023, which were subsequently loaded into a PostgreSQL database for further analysis. The download included observations from 224,123 users. Most of them (58.6%) contributed between 1 and 5 observations (Figure 1).

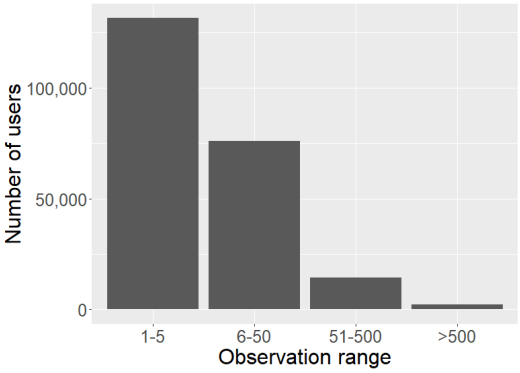

Figure 1. Distribution of users based on observation counts

Shapefiles for 143,689 U.S. parks were obtained from the Trust for Public Land[2]. The shapefiles include attributes such as park name, managing authority, address, and the demographics of the population living within a 10-minute walk radius. Due to the large number of management types provided in the shapefile (more than 20), and in order to simplify the analysis, management types were reclassified into seven key categories, i.e., national (for parks with names ending in "National Park"), federal, state, county, city, district, and regional. All other management types, such as "School", were aggregated under the "others" category. The total number of parks falling into the different park management categories (black bars), and those with at least one iNaturalist observation (grey bars), are shown in Figure 2. The subsequent analyses presented in this study utilize the latter category of parks.

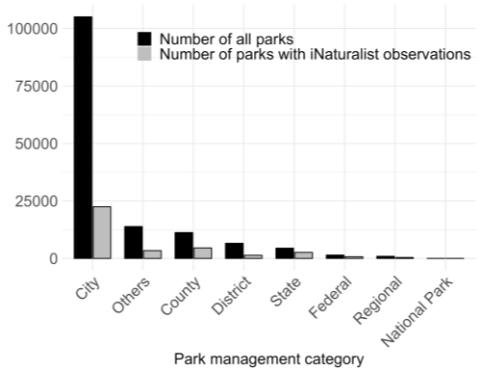

Figure 2. Number of all parks in the U.S. and subset of parks with iNaturalist observations

2.2 Primary activity location and travel distance

To assign each observation an estimated travel distance, the "home" region or, more specifically, center location of its observer's primary activity area, is needed. Since iNaturalist user profiles rarely provide a user self-reported location, this information has to be drawn from observations. Various approaches for this purpose have been developed for other crowd-sourced data platforms,

[2] https://www.tpl.org/park-data-downloads

such as OpenStreetMap (Zielstra et al., 2014) or Flickr (Bojic et al., 2016). In the first step, an observer's observation point locations were spatially clustered using DBSCAN, configured with a minimum cluster size of 3 points and a radius of 10 kilometres. Next, the cluster with the largest number of different contribution dates was selected as the observer's main activity region. Figure 3 demonstrates the process for a user's 465 contributions which were subdivided into 19 clusters. Using the points from the major cluster, the center point was subsequently computed and assumed as the primary activity location of an observer. This allowed to compute the distances between each observation location and a user's primary activity location.

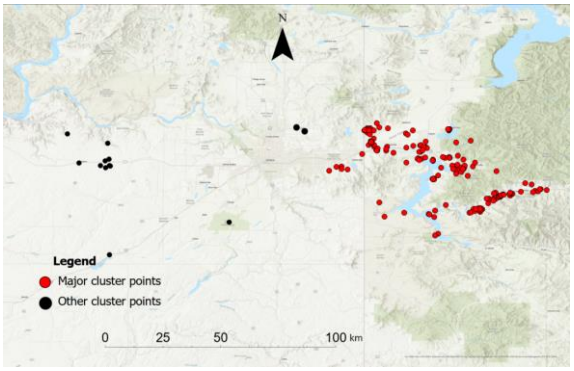

Figure 3. Primary activity cluster of one iNaturalist user

Next, the travel distance to each park was determined, where only parks with iNaturalist observations from at least two users were used. For this purpose, in a first step the average distance across the observations in a given park was computed for each observer. A second step computed the mean over all these average distances to obtain one distance associated with a park. The distribution of obtained travel distances to 20,434 parks for the 1,965,018 analyzed observations falling into parks, based on this approach, is shown in Figure 4.

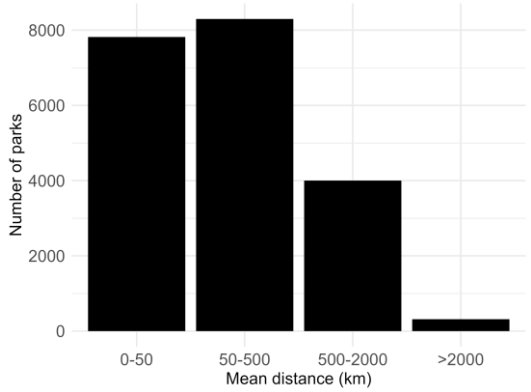

Figure 4. Distribution of mean travel distance to parks

2.3 Park attributes

Besides the average travel distance to a park, other contribution related attributes were computed for each park, including number of total iNaturalist observations (Figure 5) and number of distinct species identified in a park (Figure 6). The latter represents a proxy for biodiversity.

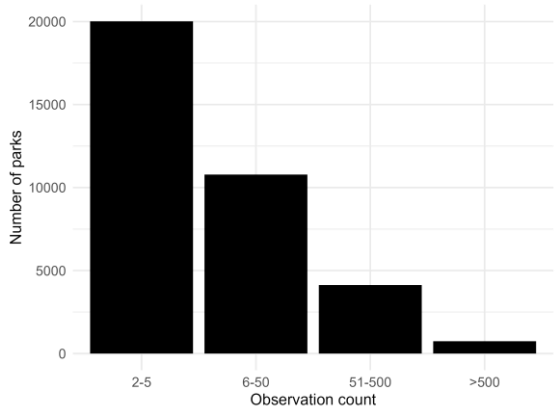

Figure 5. Distribution of observation counts in parks

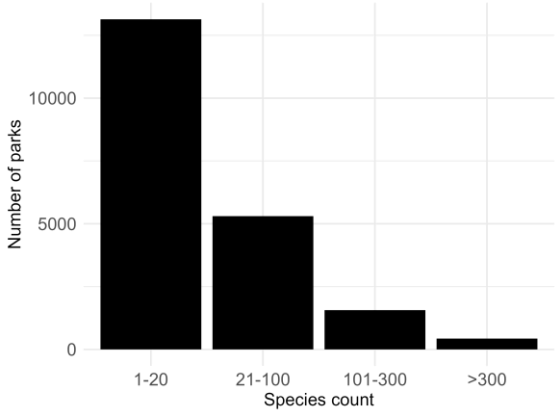

Figure 6. Distribution of species counts in parks

Descriptive statistics regarding observer-averaged travel distance and species count for the different park management categories for parks with at least two different users are presented in Table 1. It shows that mean travel distance and mean species count across analyzed parks is highest for national parks.

Table 1. Descriptive statistics for different park management types

Park Type	Distance (km)			Species Count		
	Max	Mean	Mdn	Max	Mean	Mdn
National Park	1,195	697	795	1,183	422	304
Federal	4,370	543	433	3,208	207	61
State	7,441	469	285	1,140	69	25
County	6,253	319	133	1,791	41	14
City	6,227	289	89	1,186	26	8
Others	5,181	287	109	913	37	11
District	2,265	221	92	1,018	45	12
Regional	3,855	202	58	627	46	13

2.4 Statistical analysis

When modelling travel distance to parks with the ordinary least squares (OLS) method, a strong positive autocorrelation was found in the model's residuals with a Moran's I of 0.35 ($p < 0.001$). To mitigate the influence of spatial autocorrelation on model results, an Eigenvector Spatial Filtering (ESF) model was constructed utilizing the "besf" function from the "spmoran" R package (Murakami and Griffith, 2019). Park management was used as a categorical predictor, consisting of eight levels. Other predictor variables initially considered included park size, biodiversity of a park (measured by the number of different species of iNaturalist contributions), population within a 10-minute walk from the park, expressed in 1,000s, and number of nearby parks within 10 km. Pairwise correlation analysis between predictor variables revealed a high significant correlation between biodiversity and iNaturalist contribution count (Pearson $r = 0.87$, $p < 0.001$), which led to the exclusion of the contribution count from the final ESF model. Also, park management type was found to be correlated with park size (e.g. city parks had the smallest mean area of parks) which led to the exclusion of the area variable from the final ESF model. The ESF model was manually built in a stepwise approach to improve model fit while controlling for multicollinearity. Besides the ESF regression, a Kruskal-Wallis test with Dunn's post hoc test was applied to determine the statistical significance between distances associated with different park management types.

3 Analysis Results

The ESF model ($r^2 = 0.307$) had a higher adjusted R-squared value than the OLS model ($r^2 = 0.022$). Analysis of 166 eigenvectors, filtered through the ESF model identified 66 eigenvectors as significant. The Moran's I of 0.049 ($p < 0.001$) among residuals of the ESF model demonstrates a substantial reduction in spatial autocorrelation compared to the OLS model.

Table 2 shows the results of the final ESF regression model where only significant predictors (except for park management types) were included. The Variance Inflation Factor was < 3 for all variables, indicating that no multicollinearity was present. Model results indicate that both biodiversity and nearby park number are positively associated with the average travel distance to parks. This shows that higher biodiversity and areas with a higher concentration of parks encourage travellers from farther away to visit a park after controlling for park management type. As opposed to this, parks in more populated areas offer close access to its local population, hence reducing necessary travel distances to visit a park. National, federal, and state parks are associated with longer visitor travel distances than parks with the "other" parks management category. However, there was no significant association with an increase of decrease of travel distances to smaller park types, such as county or city parks compared to parks with the "other" parks management category. These findings highlight the nuanced ways in which park management, biodiversity, and park surroundings shape visitor behaviours and preferences.

Table 2. ESF Regression results

	Coefficient
Intercept	291.2***
Biodiversity count	6.90***
Population	-9.09*
City Park	-2.40
County Park	-24.23
State Park	57.59***
Federal Park	166.7***
National Park	243.9***
Regional Park	-10.81
District Park	-35.20
Number of parks within 10km	0.065**
N	20,434
Residuals Moran's I	0.049
Adjusted R^2	0.307

Note: "Others" is the default category for park management

A Kruskal-Wallis Test indicated a significant difference in travel distance across park management categories, $H(7, n = 20{,}434) = 941.9$, $p < 0.001$. A Dunn's post hoc test with a Bonferroni adjusted alpha level of 0.0018 (0.05/28) showed that differences in distances between all park category pairs but Regional and District, and City and District were statistically significant.

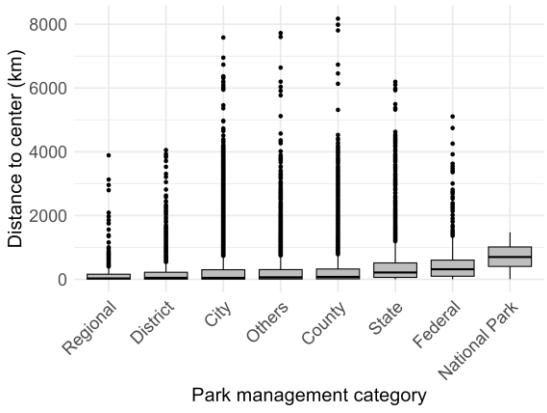

Figure 7. Boxplot of mean distances for park management category

4 Discussion

This study highlights the crucial influence of park management and biodiversity on visitor attraction, particularly travel distance to parks, showing that federal and national parks as well as parks in the vicinity of other parks draw visitors from further distances. In contrast, parks in more densely populated areas serve more localized populations. These insights advocate for management practices that balance ecological preservation with enhancing public access to green spaces. Future work aims to expand this research to the analysis of worldwide parks to and cross-reference visitation patterns with other crowd-sourced or ground-truth data.

References

Bojic, I., Belyi, A., Ratti, C., and Sobolevsky, S.: Scaling of foreign attractiveness for countries and states, Applied Geography, 73, 47-52, https://doi.org/10.1016/j.apgeog.2016.06.006, 2016.

Callaghan, C. T., Poore, A. G. B., Hofmann, M., Roberts, C. J., and Pereira, H. M.: Large-bodied birds are over-represented in unstructured citizen science data, Scientific Reports, 11, 19073, 10.1038/s41598-021-98584-7, 2021.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., and Turak, E.: Contribution of citizen science towards international biodiversity monitoring, Biological Conservation, 213, 280-294, https://doi.org/10.1016/j.biocon.2016.09.004, 2017.

Di Cecco, G. J., Barve, V., Belitz, M. W., Stucky, B. J., Guralnick, R. P., and Hurlbert, A. H.: Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science, BioScience, 71, 1179-1188, 10.1093/biosci/biab093, 2021.

Dimson, M. and Gillespie, T. W.: Who, where, when: Observer behavior influences spatial and temporal patterns of iNaturalist participation, Applied Geography, 153, 102916, https://doi.org/10.1016/j.apgeog.2023.102916, 2023.

Hochmair, H. H., Scheffrahn, R. H., Basille, M., and Boone, M.: Evaluating the data quality of iNaturalist termite records, PLOS ONE, 15, e0226534, 10.1371/journal.pone.0226534, 2020.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D.: Understanding Human Mobility from Twitter, PLOS ONE, 10, e0131469, 10.1371/journal.pone.0131469, 2015.

Ma, S., Kirilenko, A. P., and Stepchenkova, S.: Special interest tourism is not so special after all: Big data evidence from the 2017 Great American Solar Eclipse, Tourism Management, 77, 104021, https://doi.org/10.1016/j.tourman.2019.104021, 2020.

Mesaglio, T. and Callaghan, C. T.: An overview of the history, current contributions and future outlook of iNaturalist in Australia %J Wildlife Research, 48, 289-303, https://doi.org/10.1071/WR20154, 2021.

Murakami, D. and Griffith, D. A.: Eigenvector Spatial Filtering for Large Data Sets: Fixed and Random Effects Approaches, Geographical Analysis, 51, 23-49, https://doi.org/10.1111/gean.12156, 2019.

Owuor, I., Hochmair, H. H., and Paulus, G.: Use of social media data, online reviews and wikipedia page views to measure visitation patterns of outdoor attractions, Journal of Outdoor Recreation and Tourism, 44, 100681, https://doi.org/10.1016/j.jort.2023.100681, 2023.

Tu, X., Huang, G., Wu, J., and Guo, X.: How do travel distance and park size influence urban park visits?, Urban Forestry & Urban Greening, 52, 126689, https://doi.org/10.1016/j.ufug.2020.126689, 2020.

Unger, S., Rollins, M., Tietz, A., and Dumais, H.: iNaturalist as an engaging tool for identifying organisms in outdoor activities, Journal of Biological Education, 55, 537-547, 10.1080/00219266.2020.1739114, 2021.

Zielstra, D., Hochmair, H. H., Neis, P., and Tonini, F.: Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap, 10.3390/ijgi3041211, 2014.