HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Exploring the Methods and the Need for Data Anonymization of Species Locational Data

Jason Huang[1], Linlin Li[2], Kevin Zekai Li[3], Sean Choi[4] and Kelly Chin[#]

[1]St. George's School, Canada
[2]St.John's School, BC, Canada
[3]Elgin Park Secondary School, BC, Canada
[4]Mulgrave School, Canada
[#]Advisor

ABSTRACT

Our research provides datasets on different national parks and the different species that were seen there before. The names of the national park, as well as the scientific and common name of each species studied, are provided at the top of each dataset..The datasets give better understanding on how we can sustain the biodiversity of specific areas. While creating our method of collecting data, we encountered problems involving data being taken by unauthorized users for unethical actions. Our solution to that involves using geospatial masking and AI-assisted anonymization to ensure that the sensitive information of species occurrences are not shared. Data security,legal and ethical concerns,and possible threats to the regions' biodiversity are the main factors we focused on while deciding with our method of creating the research paper and solutions to sustaining biodiversity.

Introduction

Increasing attention is being paid to extinction issues, and related topics such as climate change and poaching. Extinction can be defined as a species or a group of animals having no living members, or no longer in existence (Gonzalo et. al., 2022). Species extinction poses a multitude of threats to the Earth's ecosystem. An ecosystem can be considered as an interconnected web: every living organism in said ecosystem affects one another. When these organisms are removed one after another, it leaves holes in the interconnected web, ultimately causing a ripple effect that affects other species in the ecosystem. There are various ways extinction of species could happen; however, one of the most prominent contributors are the constant threat poachers pose on endangered species - around 12 million animals are poached each year. Poachers illegally target animals, killing them for financial motives. Often, poachers are able to use unprotected biodiversity data to locate certain endangered species, which contributes to the species extinction. Due to this urgent problem, data masking is of the utmost importance.

Data masking involves the process of modifying sensitive data in a way that it has no or little value to unauthorized users, while still being able to be used by authorized personnel. In order to mask data properly, the data set should be primarily investigated. The current dataset provides us with the different species within national parks. It provides us with the scientific name and the common name for the species (Kaggle, 2017). Additionally, a limitation of the dataset would be that it does not provide a concrete number regarding the population of a certain species, but rather provides a statement saying if the species had been seen at the park. Any of the current species on the national park's list of datasets are present, and were seen before - thus, as different national parks are located in various regions, species have a range of different sightings or occurrences.

Material and Method

Biodiversity Informatics in United States National Parks: A Data Overview

Dataset Synthesis: "park.csv" and "species.csv"

Dataset Synthesis Source and Download Information: This report examines two datasets: "park.csv" and "species.csv," which are comma-separated values text files (CSV) provide detailed information on the variety of species found in American national parks. These datasets are made available by the National Park Service and can be downloaded from the Kaggle website (Cooke et. al., 2017).

Data Collection and Verification: The National Park Service has compiled these datasets based on verified records—such as sightings, physical samples, and scientific studies—that confirm the presence of various species in the parks. The data is openly shared with the public, but information on certain sensitive species is withheld to protect them.

Contents of the Data: The datasets include up-to-date information about the plants and animals in each park, and they are constantly being updated. Because the classification of species can change, and because new information is always being collected, the lists are always improving. Each record includes details like both the common and scientific species' names, their status in the park, and whether they native to the area.

Parks Dataset ("parks.csv"):

- Park Code: Unique identifier for each park.
- Park Name: Official denomination of the national park.
- State: Geographical location by state.
- Acres: The park's expanse, measured in acres.
- Latitude and Longitude: The precise geographical coordinates.

Species Dataset ("species.csv"):

- Species ID: The taxonomic identifier of the species.
- Park Name: The park wherein the species is observed.
- Category: Taxonomic categorization (e.g., Mammal, Bird, etc.).
- Order and Family: Hierarchical taxonomic ranks.
- Scientific Name: The binomial nomenclature.
- Common Names: Vernacular appellations.
- Record Status: The current status of data entry.
- Occurrence: Verification of species presence.
- Nativeness: Endemic or exotic status.
- Abundance: Population visibility and density.
- Seasonality: Temporal pattern of presence.
- Conservation Status: Threat classification according to the US Fish & Wildlife Service.
- Unnamed: 13: This column potentially includes ancillary data.

Importance of the Data: These datasets are useful for scientists, conservationists, educators, and policymakers. They help us understand the diversity of life in our national parks and can be used to make decisions about how to protect and manage these natural resources.

Methods of Data Anonymization: Our proposed methods of anonymizing the location data of national parks containing critical species include both simple geospatial masking and AI-assisted anonymization, a method which involves machine learning clustering.

Geospatial Masking: In geospatial masking, the CSV file was then read using the "pd.read_csv" method from the Pandas framework to create the "parks_data" DataFrame, and the data for the coordinates under the labels "longitude" and "latitude" were designated as sensitive columns and subsequently slightly altered using the np.random.uniform(-5,5), which generates a random floating point value from a uniform distribution within the range -5 to 5 inclusive and adds it to the actual value of the coordinate. The range of -5 to 5 was used because it strikes a balance between obscuring sensitive geographical information while still maintaining relative spatial proximity to the actual locations. While larger ranges provide more data security, they also disrupt the spatial relationship of the parks relative to their actual locations. Geospatial masking was utilized for protecting the data of parks because such data is relatively low in sensitivity; geospatial masking is an efficient yet effective methodology.

AI-Assisted Anonymization: AI-assisted anonymization is a method that provides stronger data security by further obscuring the locational data of the parks. Our method of AI-assisted anonymization is centered around clustering the location data of the parks using the machine learning algorithm K-means based on their geographical region and re-assigning regions to these clusters. The preliminary step was normalization, which adjusts the variables and features of the data to a common scale without distorting differences in the range of values (Gittleman, 2023). Our algorithm utilizes the MinMaxScaler from scikit-learn to normalize the longitude and latitude data.

The clustering of the data involved using the K-means algorithm, a relatively computationally-efficient algorithm which takes measurements of Euclidean distance to assess the similarities between data points (Egnyte, 2024).

The K-means algorithm uses Expectation-Maximization (Egnyte, 2024). The algorithm takes a parameter K for the intended number of clusters, or regions in this context, which determines the number of initial centroids (central points in clusters of data points) that are placed randomly (Egnyte, 2024). The E-step consists of calculating the Euclidean distance between each park and each centroid and assigning the datapoint to the closest cluster; the M-step consists of re-calculating the centroid of each cluster (Egnyte, 2024).

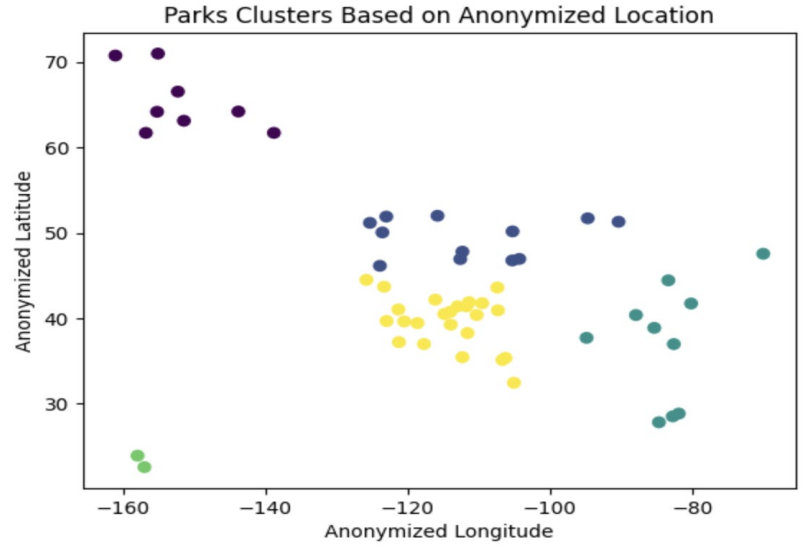

Figure 1. This figure is a scatter plot demonstrating the location of parks in the "parks_data" DataFrame grouped by color representing their cluster association.

It is important to note that the chosen K affects the outcome and using suboptimal K values will yield suboptimal clusters. We have chosen a K value of five because it is representative of the major regions of the United States. Less than five centroids would risk over-generalizing the regions of the US, whereas too many centroids would cause the clusters to be sensitive to minor fluctuations in the data distribution and the initialization of centroids.

The anonymization involves altering the region that appears in the "Region" column of the "parks_data" DataFrame. A dictionary, "cluster_to_region" is created where cluster zero corresponds to the Northern USA, cluster 1 to the Southwest USA, cluster 2 to the Eastern USA, cluster 3 to Alaska, and cluster 4 to Hawaii. The regions that each cluster corresponds to differs from its actual region. The "parks_data" DataFrame is then updated with a "Cluster" column, which indicates which cluster the park belongs to.

Cluster to region	
0	*Northern USA*
1	*Southwest USA*
2	*Eastern USA*
3	*Alaska*
4	*Hawaii*

Figure 2. This table shows the cluster_to_region, which assigns a different region to each cluster.

For each park in the "parks_data", the value in the "Region" column is then mapped with the corresponding region associated with the cluster as defined in the "cluster_to_region" dictionary as seen in Figure 4.

Results

The primary goal of our code is the anonymization of sensitive location data. Our results consist of both visualizations of the anonymized data along with copies of the anonymized data itself. Through randomly adjusting the coordinates of parks and through using clustering and mapping, we are able to safeguard the confidentiality of sensitive information while still preserving details about general geographical location. The methods we employ provide sufficient security and are relatively scalable.

We created a Pandas DataFrame "anonymized_parks_data" that is a copy of "parks_data" with the coordinates altered by adding random values between -5 and 5. One of the greatest threats to biodiversity is poaching, which is responsible for the endangerment of many critical species. This altered DataFrame protects sensitive species by making it harder for unauthorized users to pinpoint the exact locations these species are located in. Without this geospatial masking, unauthorized users might gain unwanted direct access to coordinates in which the habitats of critical species are located.

Our results also consist of two maps created using the Folium library in Python: "anonymized_parks_map.html", which containing the locations of the geospatially masked parks, and "clustered_parks_map.html", containing the locations of the parks grouped by region following anonymization.

Figure 3. This figure shows the "anonymized_parks_map.html" map.

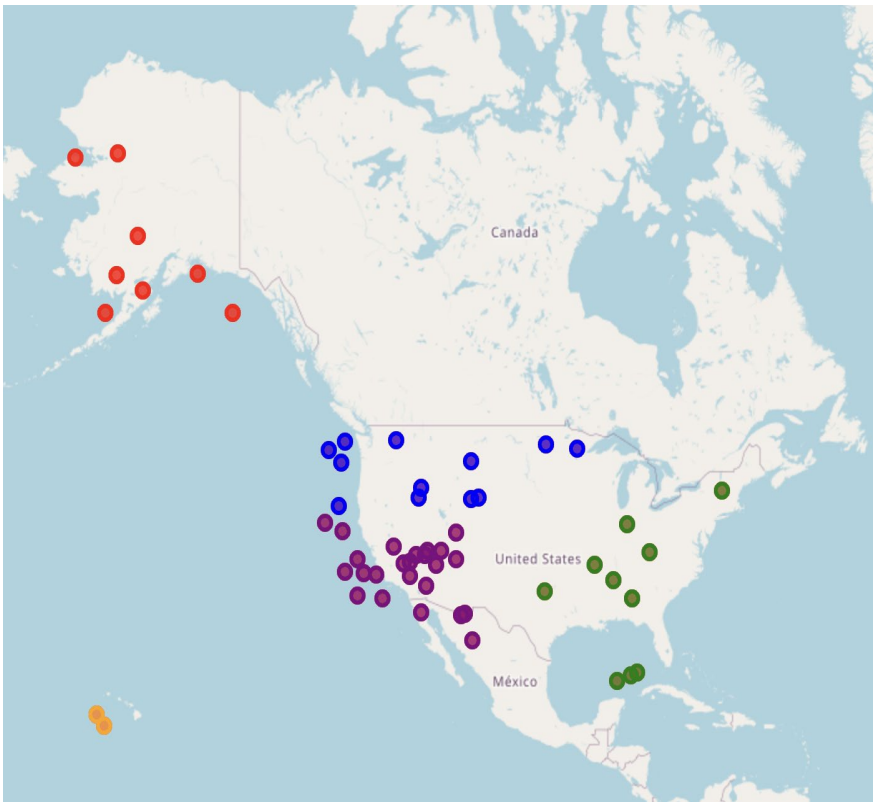

Figure 4. This figure shows the "clustered_parks_map.html" map.

Lastly, we created a scatter plot using MatPlotLib (see Figure 1) which plots the locations of the parks where each point is color-coded based on the cluster they are associated with. The scatter plot provides a visualization of clusters that allows for statistical understanding of cluster relationships along with algorithmic evaluation of the K-means algorithm; contrastingly, the Folium maps provide more geospatial context.

Our solution focuses on a solution that focuses on maximizing protection through encryption techniques. Such techniques ensure that even if access is gained by unauthorized users, they cannot decipher the data, thus protecting data confidentiality. Methods like geospatial masking machine learning clustering prevent unauthorized users who gain access into biodiversity databases from identifying the locations of critical species, effectively protecting endangered species from poachers, habitat destruction, or other forms of exploitation. Without these anonymization measures, such users would have easy access to sensitive details that enable them to locate the habitats of critical species. Unwanted access to this data can lead to targeted exploitation which can potentially disrupt fragile local ecosystems and pose permanent existential threats to endangered species.

Discussion

Threats to Biodiversity Data

The misuse of data from the surveillance of animal movements is the main cause behind risks to biodiversity such as poaching, habitat destruction, and illegal trade. As animals are often tracked using radio, acoustic or satellite transmitters that are physically attached to them, information can be obtained by receivers that detect signals given off by the equipment (IBM, 2024). Thus, the specific locational data can be intercepted with appropriate receivers listening for tagged animals in an area (Dabbura, 2022). In particular, radio and acoustic transmitters experience a higher vulnerability to misuse, as oftentimes the receivers are inexpensive and easily obtainable by potential poachers. In addition, the signals given off by these electronic tagging technologies are often unencrypted, leaving it open to unauthorized access and allowing them to exploit the system. Anonymization and masking techniques can be incorporated into the signals from transmitters so that it would provide an inaccurate or non-specific location to those who are unauthorized. Another application could be to also alter the information so only those with access can decode or interpret the signals correctly (Dabbura, 2022).

Currently, most of the technology in protecting sensitive species from poaching involves using AI to predict poaching trends (Mindy, 2022). However, a new source of information for poachers results from the rise and accessibility of the internet (Norton, 2024), where the interception of data allows unauthorized users and poachers to obtain information on specific locations of animals. This causes animals to experience an increased vulnerability to human exploitation, which promotes the disturbance of ecosystems. Information is often obtained by accessing databases, maps, or public outreach websites. Additionally, due to current efforts to fight biodiversity loss, researchers are using large amounts of data to generate trends in ecosystems(Thompson), such as tracking wildlife patterns or making predictions about population changes of endangered species (AIWS, 2024).This puts data at risk, where reports showing positional data of rare animal distributions can be used by poachers to target and illegally harvest the animal (Dabbura, 2022). Geospatial masking and AI-assisted anonymization can also be used when publishing results so that aggregated data or generalized information on animal movements is provided instead of exact locational data. Researchers will be able to incorporate data as normal at a lesser degree of risk towards the endangered animals.

Data Sharing vs. Data Security

Data sharing grants individuals access to information and insights of the organization's works (Oyekunle, 2022). It is a fundamental part of publishing data research and including many types of sharing. Internal data sharing is to individuals within the organization (Ravikiran, 2023). External data sharing entails providing data to people outside of the party. Data sharing efficiently collaborates groups between each other by offering people with products and services and removes barriers that prevent people from accessing knowledge and data from specific topics (Lennox and Harcourt, 2020). Although data sharing has become an essential step to data sharing and other variations of publishing knowledge, it still contains risks and mitigations. Vital information within the data can potentially be accessed by unauthorized users from cyberattacks and ransomware . That is why data security plays an important role in protecting data and preventing threats such as ransomware, theft and corruption (U.S. Department of the Interior, 2024). Encryption transforms normal text into scrambles that only authorized users have learned to decode and read; Data erasure overwrites data stored in a device and makes it unrecoverable. Data masking hides the original data with modified work added by the users. Data resiliency allows an organization to recover data that is lost and minimizes impact when there is a cybercriminal activity. With data sharing providing information to others for collaboration and data security preventing cyber risks, datasets and research papers are able to present thorough analysis while also keeping threats from alternating or abusing the data.

Technological Solutions

Due to advances in technology, there are various solutions that can be investigated and utilized to secure biodiversity data. One such method is the blockchain technology (Machine Learning, 2024), a framework that is used to store public transactional records, or blocks. This type of storage is commonly known as a "digital ledger." The digital signature of the owner authorizes each transaction in this ledger, ensuring its authenticity and preventing any form of tampering (Machine Learning, 2024). Because of this, the additional layer of security assures the integrity of biodiversity data. Advanced encryption, such as end-to-end encryption, is an additional measure that protects data privacy and secrecy both during transmission and storage. With end-to-end encryption, data is protected from unwanted interception in transit by undergoing encryption during transmission and is only decrypted by the intended recipient. This method prevents sensitive information concerning animal movements, habitats, and conservation activities from being compromised during communication between researchers, conservationists, and appropriate authorities, which makes it extremely important when dealing with biodiversity data. As mentioned prior, a large risk of poaching is due to the interception of signals from animal monitoring equipment. End-to-end encryption is especially suitable in adding security to the transmitted signals and preventing unauthorized access by unauthorized users. These methods of data security do not only apply for the conservation of biodiversity data. When used in conjunction with our AI-assisted anonymization and masking technology, its applications can be extended beyond. AI-assisted anonymization can be trained with algorithms to adapt based on evolving threats. Hence, there are many industries that can use this new technology. Health care could benefit by more effectively masking patient data in sensitive research without compromising privacy. In finance, this can be used to enhance security protocols regarding the protection of customer data in financial transactions and analyses. With advancements in technology also comes greater security risks for online databases. As such, the importance of continued innovation must be emphasized to mitigate threats.

Legal and Ethical Considerations

There are multiple considerations one must take into account regarding the usage and collection of biodiversity data, the most prominent being Informed Consent. To gain a better understanding of Earth's diverse ecosystems

and species, many data collectors trespass into territories without being mindful of who the land may belong to. It is always crucial to obtain informed consent from individuals or communities when using or collecting biodiversity data, as this ensures that their land is treated with respect (Databricks, 2024). Alongside informed consent, the data that is being analyzed must be confidential to a certain extent. Owners of datasets that involve any personal or endangered species information must be aware of the implications that sharing the data can cause. By protecting the identities and location of the endangered species, the ecosystem is able to thrive without the risk of poachers decimating the populus further. Additionally, owners must consider the long term impacts of sharing biodiversity data. One must comply with environmental impact assessment regulations and conduct through assessments before undertaking biodiversity research projects (Bosworth, 2011) When biodiversity data is used to investigate species that reside on indigenous properties, it is vital that one acknowledges and respects Indigenous knowledge and practices related to biodiversity, which ensures that the collection and use of knowledge are conducted in a culturally appropriate and sensitive manner. Another crucial point revolving around the ethicality of data sharing is the concept of benefit sharing. When biodiversity data leads to commercial products or services, ensuring fair benefit sharing with local communities, indigenous peoples, or countries is essential. This can be done by implementing benefit-sharing agreements, helping ensure that the benefits from biodiversity resources are shared fairly, as well as contributing to conservation efforts. Legally, regulations and international conventions exist in regards to governing biodiversity data collection, sharing, and use. This includes compliance with national and international frameworks such as the Convention of Biological Diversity (CBD), the Nagoya Protocol on Access and Benefit Sharing, and regional or local regulations.

Conclusion

We utilized the datasets of different national parks, each with different areas covered and different amounts of species occurrence, to identify the biodiversity of that specific region. Our method of anonymizing the data for data includes geospatial masking and AI-assisted anonymization. These two effectively prevent unauthorized users and other unauthorized users from identifying locations of species and information about them. Without it, people are able to exploit these data for illegal and unethical action such as poaching endangered species or destroying natural habitats. We provided two maps using the Folium library in Python, one being an anonymized parks map, and the other being a clustered parks map. We created a scatter plot using MatPlotLib and color coded the points to show the relationships between the regions and its biodiversity with others. With the method and material used for the research on biodiversity, we were able to anonymize the data, while also presenting clear graphs and plots on the biodiversity of each region. We discussed threats to biodiversity data, data sharing and data security, technological solutions, and legal and ethical considerations. We took these factors into account during our research, because when these factors are not taken into account for the research, threats such as poaching or cyberattacks can emerge. For every research on biodiversity, members need to protect the region they analyze by applying these factors to their method of collecting data.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

Reference

1. AI World School. (2024, 2 March). This is why AI in wildlife conservation is so glorious! AI World School, https://aiworldschool.com/research/this-is-why-ai-in-wildlife-conservation-is-so-glorious.

2. Bosworth, Andrew, et al. (2011). *Ethics and biodiversity.* UNESCO Digital Library, http://unesdoc.unesco.org/ark:/48223/pf0000218270?posInSet=1&queryId=db9e18aa-bcc2-4826-a752-e5275086dd05.

3. Cooke, Steven J., et al. (2017). *Troubling issues at the frontier of animal tracking for conservation and management.* The Society for Conservation Biology, https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.12895

4. Dabbura, Imad. (2022). *K-Means Clustering: Algorithm, applications, evaluation methods, and drawbacks.* Medium, Towards Data Science, https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

5. Databricks, (2024). *What is data sharing?* Databricks, https://www.databricks.com/glossary/data-sharing .

6. Egnyte, (2024, January 5). *What Is Data Sharing?* Egnyte https://www.egnyte.com/guides/life-sciences/data-sharing .
7. Elsevier,Cortés-Capano, Gonzalo, et al. (2022). *Ethics in biodiversity conservation: The meaning and importance of pluralism.* Biological Conservation, https://www.sciencedirect.com/science/article/pii/S0006320722003123?via%3Dihub

8. Gittlemen, John L. (2023). Extinction. In Encyclopædia Britannica, Encyclopedia Britannica, https://www.britannica.com/science/extinction-biology

9. IBM (2024, January 5). What is data security? IBM, https://www.ibm.com/topics/data-security .
10. Kaggle, (2017, January 5). B*iodiversity in national parks*. https://www.kaggle.com/datasets/nationalparkservice/park-biodiversity

11. Lennox, Robert J., and Robert Harcourt. (2020). *A novel framework to protect animal data in a world of ecosurveillance.* Bioscience, vol. 70, no. 6, p. 10.

12. Machine Learning, (2024, January 5). *Normalization.* Google, https://developers.google.com/machine-learning/data-prep/transform/normalization

13. Mindy Support. (2022, January 25) Stopping animal poaching with AI and data annotation. *Mindy Support*, https://mindy-support.com/news-post/stopping-animal-poaching-with-ai-and-data-annotation/.

14. Norton, Kara. (2020, 30 October) The 21st century threat to wildlife is "cyberpoaching." *PBS*, https://www.pbs.org/wgbh/nova/article/21st-century-threat-wildlife-cyberpoaching/

15. Oyekunle, Isa. (2022). *Data sharing: Definition, types, benefits, and examples*. Security Gladiators , https://securitygladiators.com/data-sharing/

16. S, Ravikiran A. (2023, October 18). *What is blockchain technology? How does blockchain work?* Simplilearn, https://www.simplilearn.com/tutorials/blockchain-tutorial/blockchain-technology

17. Thompson, Tosin. (2023, 6 January). How AI can help to save endangered species. *Nature*, https://www.nature.com/articles/d41586-023-03328-4. Accessed 2 March 2024.

18. U.S. Department of the Interior. (2024, January 5)"*National Parks Service"*, U.S. Department of the Interior, https://irma.nps.gov/NPSpecies/Search/SpeciesList/ACAD