### **Class I. Data set descriptors**

1. **Data set identity**: Artificial Hotspot Occurrence Inventory (AHOI)
2. **Data set description**
	1. **Originator(s)**:
		1. Daniel S. Park. Department of Biological Sciences | Purdue Center for Plant Biology, Purdue University, West Lafayette IN 47907
		2. Yinying Xie. Department of Biological Sciences, Purdue University, West Lafayette IN 47907
		3. Hanna T. Thammavong. Department of Biological Sciences, Purdue University, West Lafayette IN 47907
		4. Rima Tulaiha. Department of Biological Sciences, Purdue University, West Lafayette IN 47907
		5. Xiao Feng. Department of Geography, Florida State University, Tallahassee, FL, 32306
	2. **Abstract**:
	Aim: Species occurrence records are essential to understanding Earth’s biodiversity and addressing global environmental issues. However, these data do not always reflect actual locations of occurrence. Certain geographic coordinates are assigned repeatedly to thousands of observation/collection records. This may result from imperfect data management and georeferencing practices, and can greatly bias the inferred distribution of biodiversity and associated environmental conditions. Nonetheless, these ‘biodiverse’ coordinates are often overlooked in taxon-centric studies, as they are identifiable only in aggregate across taxa and datasets, and it is difficult to determine their true circumstance without in-depth, focused investigation. Here we assess highly recurring coordinates in biodiversity data to determine artificial hotspots of occurrences.

Location: Global

Taxon: Land plants, birds, mammals, insects

Methods: We identified highly recurring coordinates across plant, bird, insect, and mammal records in the Global Biodiversity Information Facility, the largest aggregator of biodiversity data. We determined which are likely artificial hotspots by examining metadata from over 37 million records; assessing spatial distributions of associated datasets; contacting data managers; and reviewing literature. These results were compiled into the Artificial Hotspot Occurrence Inventory (AHOI)

Results: Artificial biodiversity hotspots generally comprised geopolitical and grid centroids. The associated uncertainty ranged from several square kilometers to millions. Such artificial biodiversity hotspots were most prevalent in plant records. For instance, over 100,000 plant occurrence records were assigned the centroid coordinates of Brazil, and points that have at least 1,000 associated occurrences comprised over 9 million records. In contrast, highly recurring coordinates in animal data more often reflected actual sites of observation.

Main Conclusions: AHOI can be used to i) improve accuracy of biodiversity assessments; ii) estimate uncertainty associated with records from artificial hotspots and make informed decisions on whether to include them in scientific studies; and iii) identify problems in biodiversity informatics workflows and priorities for improvement.

**Key words**: biodiversity, centroid, coordinates, duplication, georeferencing, grid

### **Class II. Research origin descriptors**

1. **Overall project description**:
	1. **Identity**: Artificial Hotspot Occurrence Inventory (AHOI)
	2. **Period of study**: February 25, 2021 – November 12, 2021
	3. **Objectives**: Our objectives for compiling highly-recurrent geographic coordinates in primary biodiversity data were to i) assess the presence of inaccurate, artificial hotspots of biodiversity and associated records; ii) identify common data sources and practices associated with artificial biodiversity hotspots; and iii) promote the use of accurate occurrence data for studies of ecology, evolution, and biodiversity. Our dataset represents an in-depth, focused investigation of the validity of primary biodiversity data which can be leveraged to infer the locations and environments species inhabit.
	4. **Experimental or sampling design**
		1. **Design characteristics**: Primary biodiversity records with coordinate data were downloaded from the Global Biodiversity Information Facility. We evaluated the frequency of each coordinate in plant, bird, insect, and mammal occurrence data and determined the circumstances of the top 100 most recurrent points for plants and the top 50 for each animal group by i) examining individual record metadata; ii) assessing the spatial distribution of associated datasets; iii) contacting relevant data managers; and iv) literature review (see Figure 1 for example).
		2. **Data collection period, frequency, etc**.: Primary biodiversity records were queried from the Global Biodiversity Information Facility on January 30 and May 10, 2021 for plants (Plantae; https://doi.org/10.15468/dl.th5tn8; https://doi.org/10.15468/dl.76jc24), July 29, 2021 for birds (Aves; https://doi.org/10.15468/dl.gtxhe3), and August 23, 2021 for insects (Insecta; https://doi.org/10.15468/dl.4q2972), and mammals (Mammalia; https://doi.org/10.15468/dl.cujmgz).
2. **Project personnel**: Principal and associated investigator(s), technicians, supervisors, students
	1. Daniel S. Park. Department of Biological Sciences | Purdue Center for Plant Biology, Purdue University, West Lafayette IN 47907
	2. Yinying Xie. Department of Biological Sciences, Purdue University, West Lafayette IN 47907
	3. Hanna T. Thammavong. Department of Biological Sciences, Purdue University, West Lafayette IN 47907
	4. Rima Tulaiha. Department of Biological Sciences, Purdue University, West Lafayette IN 47907
	5. Xiao Feng. Department of Geography, Florida State University, Tallahassee, FL, 32306

###  **Class III. Data set status and accessibility**

1. **Status**
	1. Latest update: Oct 10, 2022
	2. Latest archive date: Oct 10, 2022
	3. Metadata status: Oct 10, 2022
	4. Data verification: Oct 10, 2022
2. **Accessibility**
	1. **Storage location and medium**: Data will be archived and shared publicly on DRYAD following acceptance at Journal of Biogeography.
	2. **Contact person(s)**: Daniel S. Park (danielpark@purdue.edu), Xiao Feng (fengxiao.sci@gmail.com)
	3. **Proprietary restrictions**:
		1. **Release date**: Released upon acceptance at Journal of Biogeography.
		2. **Citation**: Provided upon acceptance at Journal of Biogeography.
		3. **Disclaimer(s)**: The Global Biodiversity Information Facility constantly updates data and metadata, and the number of records associated with artificial biodiversity hotspots are subject to change.

### **Class IV. Data structural descriptors**

1. **Data set file**
	1. **Identity**: AHOI.zip
	2. **Size**: evaluations of 306 unique coordinates comprising 37,343,922 occurrence records across 4 taxonomic groups (plants\_20220216.csv, birds\_20220216.csv, insects\_20220216.csv, mammals\_20220216.csv)
	3. **Format and storage mode**: 4 .csv files and 1 word file (metadata) compressed in a single .zip file.
	4. **Header information**: See Table 1 and Table 2
2. **Variable information**

See Tables 1 and 2

**Table 1.** Header information.

|  |  |  |
| --- | --- | --- |
| **Header name** | **Storage type** | **Definition** |
| ID | Character | unique identification codes for each set of coordinates |
| frequency | Integer | number of occurrence records associated with the coordinates |
| decimal\_latitude | Numeric | latitudinal coordinate values in decimal degrees |
| decimal\_longitude | Numeric | longitudinal coordinate value in decimal degrees |
| example\_description | Character | example of relevant locality/dataset information extracted from metadata |
| reasoning | Character | short reasoning behind coordinate category determination |
| category | Character | the category of coordinates (see Table 2 for details) |
| determination | Character | whether coordinates represent the location of actual observations/collections (TRUE) or artificial aggregates of records (FALSE) |
| notes | Character | relevant specific and/or sub-categorical information regarding the coordinates |
| doi | Character | Digital Object Identifier of queried GBIF data |

**Table 2**. Definitions of different coordinate categories.

|  |  |
| --- | --- |
| **Category** | **Definition** |
| georeferenced\_location | Coordinates are georeferenced to a landmark or an area (e.g., nature preserves, parks)  |
| geopolitical\_centroid | Coordinates represent the center of a geopolitical entity (e.g., countries, counties, cities) |
| grid\_centroid | Coordinates represent the centroid of a cell on a survey grid |
| collection\_site | Coordinates represent location of actual specimen collection (e.g., fossil dig sites, insect traps) |
| observation\_site | Coordinates represent location of actual observations (e.g., camera traps, bird watching) |

###

**Figure 1.** Example of AHOI point determination workflow. Steps 2 – 4 can be undertaken in different combinations/orders depending on the evaluated coordinates and available information.

### **Class V. Supplemental descriptors**

1. **Data acquisition:** See Section II B
2. **Quality assurance/quality control procedures**: In determining the grid systems, coordinates were mapped in GIS software (e.g., QGIS) to identify the spatial patterns and grid system resolutions. Distances between coordinates and known geopolitical centroid coordinates were examined. All entries were reviewed to ensure the consistency, quality, and accuracy of locality descriptions, reasoning and coordinate categories.
3. **Related materials**: Coordinates from the Getty Thesaurus of Geographic Names (TGN, <https://www.getty.edu/research/tools/vocabularies/tgn/>) were used for the plant data records from the Field Museum of Natural History as references, as associated metadata indicated that TGN was used for georeferencing.
4. **Computer programs and data-processing algorithms**: QGIS
5. **Archiving**
	1. **Archival procedures**: Data will be archived and shared publicly on DRYAD following acceptance at Journal of Biogeography.
6. History of data set usage
	1. **Data request history**: N/A
	2. **Data set update history**: last update was on Oct 10, 2022
	3. **Review history**: last review was on Oct 10, 2022
	4. **Questions and comments from secondary users**: N/A