Received: 3 September 2025 | Accepted: 18 September 2025

DOI: 10.1111/2041-210X.70166

Methods in Ecology and Evolution

APPLICATION

Release the HOGS: An unsupervised marker extraction, classification and georeferencing approach for biodiversity data

Harri Ravenscroft[1] | John Gardner[2] | Utku Perktas[3] | Erin Saupe[1]

[1]Department of Earth Sciences, University of Oxford, Oxford, UK

[2]Department of Chemistry, University of Oxford, Oxford, UK

[3]Department of Biology, Hacettepe University, Ankara, Turkey

Correspondence
Harri Ravenscroft
Email: harri.ravenscroft@stx.ox.ac.uk

Funding information
Natural Environment Research Council, Grant/Award Number: NE/S0007474/1; University of Oxford

Handling Editor: Aaron Ellison

Abstract

1. Distributional data are essential for understanding species and community responses to environmental and anthropogenic change. Large biodiversity databases provide key information on distributional patterns, but their temporal coverage can be limited.
2. Facilitating access to untapped occurrence data is a pressing need. We present the Historical Occurrence Georeferencing System (HOGS), a Python protocol that isolates occurrence markers from distribution maps, assigns them to taxonomic groups and georeferences each image relative to a baseline coordinate system, producing latitude–longitude coordinates for all mapped marker occurrences.
3. We tested our georeferencing protocol on two historical atlases of over 1000 sub-Saharan bird species, georeferencing 153,052 markers from 751 maps.
4. Our new georeferencing protocol enables rapid generation of occurrence data from historical maps, providing reference or enhanced distribution data for species and inputs for ecological niche modelling in biogeographical, evolutionary and conservation studies.

KEYWORDS
biogeography, conservation, georeferencing, historical data, modelling community ecology, spatial analysis

1 | INTRODUCTION

Anthropogenic activities, such as habitat degradation, have led to species extirpation and range contraction (Kemp, 2023). Human impacts on biodiversity are expected to continue into the future and instigate significant changes in the distribution of organisms on Earth (Pimm et al., 2014; Szabo et al., 2012). To better contextualise these changes, we need baselines prior to large-scale anthropogenic impacts (Bush, 2023; Kemp, 2023). Distributional baselines allow conservation practitioners to better assess the degree of change and to establish effective strategies to reconstitute those baseline conditions (Kiessling et al., 2019), manage resources, ensure land preservation and restoration, investigate the effects of climate change, among others (Huang & Frimpong, 2015; Seltmann et al., 2018).

Much of the biodiversity data existing in current online repositories (GBIF, 2024; Sullivan et al., 2009; Tobias et al., 2022), however, do not provide long-term (multi-decadal) perspectives because they are either too recent or are biased geographically due to uneven sampling, data storage and mobilisation efforts (Beck et al., 2014; Parker et al., 2024). This temporal mismatch fosters shifting baseline

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2025 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

syndrome (Bonebrake et al., 2010), wherein standards for environmental health are continually lowered. Undigitised specimen records (Nowak et al., 2021) can address this gap, offering untapped insight into past biodiversity patterns (Marcer et al., 2022; Nelson & Ellis, 2018). To use these records in global change studies, precise georeferencing—the process of interpreting a locality description or points on imagery into a spatially mappable representation—is essential (Marcer et al., 2022).

Georeferencing is the process of either verifying the associated geographic coordinates or assigning new ones to locality strings or occurrences on maps (Chapman & Wieczorek, 2020; Marcer et al., 2021; Thorn, 2022; Venceslau & Lopes, 2019). Most georeferencing tools for map-based occurrence data are essentially manual, requiring the user to generate a point array over the inputted map image, which necessitates considerable data entry, panning and on-screen data manipulation (Burt et al., 2020) using software packages such as ArcGIS (ArcGIS Pro, 2024) or QGIS (QGIS, 2024). Although the approach can be highly accurate, it incurs a significant time cost and can be prone to user error. Previous attempts to automate georeferencing of maps have focussed on the acquisition of basemap reference information (Heitzler et al., 2018; Zatelli et al., 2022). Although effective, these methods focus on linear features or character strings (e.g. contour lines, roads and words), and therefore do not have the flexibility to detect point-level occurrences depicted in distributional atlases, which is often desired for biodiversity analyses.

Here, we develop an unsupervised Historical Occurrence Georeferencing System (HOGS) to georeference point locations from high-resolution images of species distribution maps. Implemented using the Python programming language, HOGS extracts point occurrence markers from single digitised maps or a collection of maps and assigns latitude/longitude coordinates to the markers. This approach significantly enhances the efficiency of extracting georeferenced occurrence data from maps of any spatial resolution, geographic scale and time period. Additionally, HOGS is able to automate the extraction of occurrence data from maps when there are overlapping occurrences and when multiple marker types represent different taxon identities. HOGS georeferencing requires the selection of only a few known marker points on a map, which can be sites of interest such as graticule intersections or distinctive geographic features (e.g. mountain peaks and peninsula tips). We test HOGS using two historical distribution atlases for sub-Saharan Passerine and Non-Passerine birds: 'An Atlas of Speciation in African Passerine Birds' (Hall & Moreau, 1970) and 'An Atlas of Speciation in African Non-Passerine Birds' (Snow, 1978). These atlases synthesise distributional, ecological and evolutionary information, including over 1000 birds known to breed south of the Sahara (20° N) in Africa, for a total of 153,052 markers from 751 maps. They present an ideal case study due to the inclusion of varied marker shapes at different degrees of overlap across a wide area.

2 | METHODS

2.1 | Overview of HOGS

2.1.1 | Initial processing

Map scans

HOGS is a georeferencing protocol that takes occurrence marker data from maps and automates the assignment of latitude and longitude coordinates. A single digitised image of each of the maps is required as input. Following this, HOGS consists of a series of steps that first isolates markers (i.e. taxon occurrences) of interest, then categorises these markers using an unsupervised clustering approach to identify different populations, species, sexes or other taxon identities, and finally generates georeferenced occurrences (marker positions), given in decimal degree coordinates (Figure 1).

Marker separation

The first step in HOGS is to digitise the physical map and separate it into two layers: (i) the marker layer with occurrence points (Figure 2B), and (ii) the background layer with all other features. This is achieved via colour thresholding, which requires markers and background elements to differ in RGB values. Marker pixels are identified by calculating the mean RGB value for each pixel and flagging those below a user-defined threshold, which will be dependent upon the colours in the digital map copy. Common map colours, such as blue (RGB: 0, 0, 255) for aquatic features and green (RGB: 0, 255, 0) for vegetation, can be split easily from common black (RGB: 0, 0, 0) markers in this way.

HOGS retains a background layer in order to later reference map alignment. Background layer creation requires that all the pixels identified as markers are replaced with white (RGB: 255, 255, 255), effectively removing the markers from the image, using our function 'seperate_markers_and_map'. The remaining background, from which the occurrence markers have been removed, is retained.

2.1.2 | Initial marker processing

Marker extraction

Following isolation of the occurrence marker layer and conversion to a black-and-white binary colour scale, the shape and location of each isolated marker must be extracted. This can be difficult when the markers overlap, which can occur in areas of high marker density. Occurrence data may also be represented by different markers or shapes for different populations, species or breeding status, which need to be distinguished. The findContours algorithm from the Python OpenCV catalogue v4.12.0.88 (Bradski, 2000) is therefore used to distinguish the outer contours of each occurrence marker.

Our function 'get_shapes_and_centers' defines the contours (boundaries) of all the connected components (shapes) in the binary image marker as a list of points representing each shape's boundary.

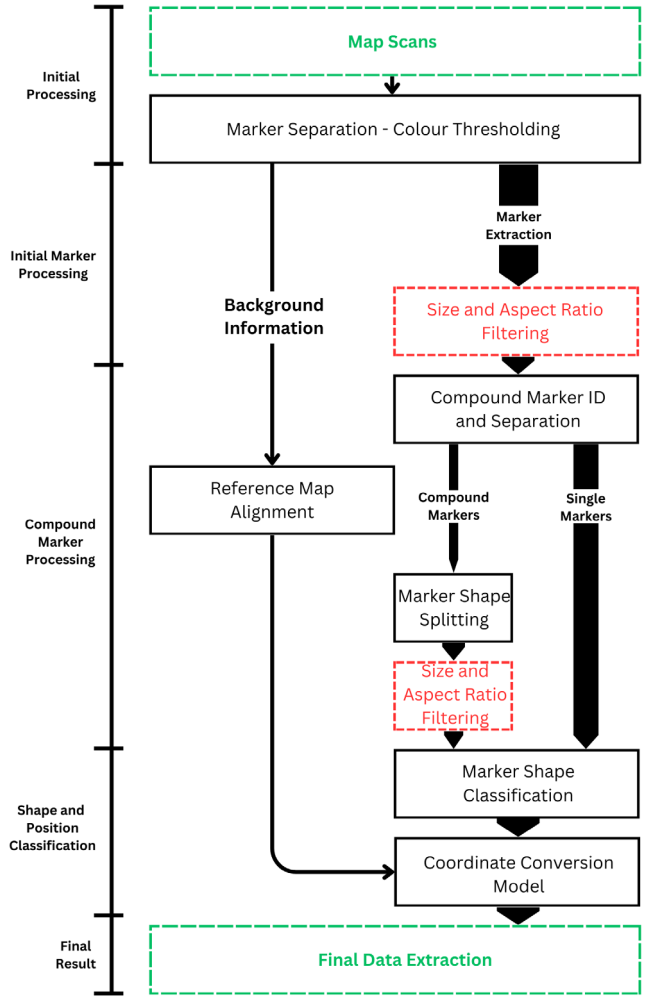

FIGURE 1 Georeferencing protocol steps in HOGS. Green dashed boxes denote inputs and outputs, red dashed boxes indicate data filtering steps, and black lined boxes represent HOGS processing steps. Inputs are high-resolution (typically ≥300 PPI: Adobe, 2025) digitised historical maps (Map Scans). Occurrence markers are separated from background features (e.g. boundaries, gridlines) and isolated (Marker Separation). Size and aspect ratio thresholds distinguish single markers from complex, overlapping compound markers (Compound Marker ID and Separation). Compound markers are split by identifying the outer contour, locating inflection points, selecting tie-lines and dividing along these lines (Marker Shape Splitting). Single and separated markers are pooled for PCA, then clustered by morphology using HDBSCAN (Marker Shape Classification), enabling species assignment from legend information. Marker centroid pixel coordinates are recorded relative to each map. Reference Map Alignment occurs in parallel, aligning all maps to a reference via automated outline/feature detection and Bayesian-optimised transformation. After manual selection of known geographic features, pixel coordinates are converted to decimal degrees (Coordinate Conversion Model). Final outputs are CSV files containing geographic coordinates (latitude and longitude), shape type and map ID for each marker (Final Data Extraction).

A contour is the curve that joins all the continuous points along a boundary with the same color or intensity. Only external contours are retrieved, while nested contours are ignored, since any occurrence marker with internal complexity such as white space could generate additional, incorrect occurrence markers.

For each contour found, the bounding box coordinates (i.e. pixel values) are retrieved. The centroid position for the bounding box of each shape is also calculated as the average of the minimum and maximum *x* and *y* pixel coordinates. The bounding box coordinates and centroid locations for each shape are stored within an array.

Markers are then filtered by size and aspect ratio. The default aspect ratio threshold (<3) removes line markers, dust and other artefacts. The minimum marker size is set to 20 pixels; at the high-resolution baseline of 300 pixels per inch (PPI) (Adobe, 2025), occurrence markers are rarely smaller than 30–50 pixels. This threshold may be adjusted for lower resolution maps or unusually small markers.

At this stage, the occurrence markers have not been grouped by shape, and overlapping/compound markers have not been identified.

2.1.3 | Compound marker processing

Compound marker identification and separation

Occurrence points from the marker layer must be identified as either single or overlapping (i.e. or compound) markers. Compound markers are identified and filtered from single marker occurrences based upon size thresholding:

$$\text{Compound shape} = \text{Shape extent} > (\text{Median extent} \times \text{Outlier cutoff}) \quad (1)$$

How a compound shape is defined based upon size thresholding.

In HOGS, the default outlier size cut-off multiplier is 1.5. This value allows for variation from scanning errors and printing flaws, while excluding overlapped markers. The threshold can be adjusted in the 'filter_outliers' function based on marker size in the input maps.

Marker shape splitting

The compound markers pose an issue, as they represent a conglomeration of distributional data markers that cannot be easily discriminated (Figure 3A). Removal of these shapes would reduce the information retrievable from a given data source and produce misleading classifications of distributional patterns, often in areas where data are most dense. Therefore, the compound markers must be isolated into smaller, single occurrence components to return their corresponding shapes and centres. To address this, HOGS automatically splits compound markers into their individual components, identifying their shapes and centres so they can be treated as single occurrences. This is achieved using contour extraction and shape-splitting methods (Figure 3), with full algorithmic details provided in the Supporting Information.

The final output is a set of single occurrence markers with centroid pixel coordinates relative to the original map. Processing time is typically under 10 s per test map.

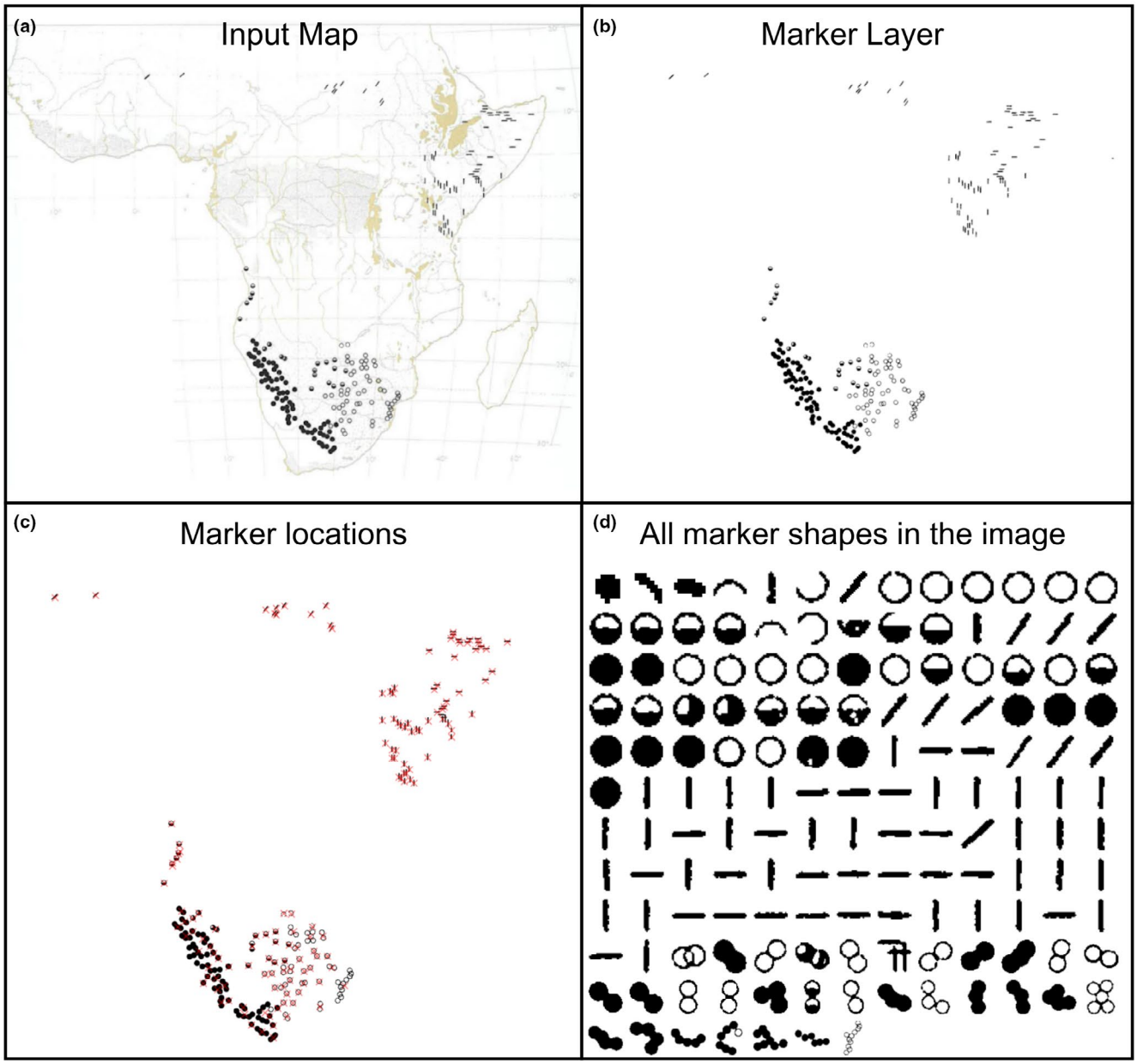

FIGURE 2 (a) An example of a historical occurrence map prior to any alteration, showing seven different species marker types for sub-Saharan African birds from Hall and Moreau (1970). Markers display members of the *Calendulauda sabota* species group (previously *Mirafra sabota*). (b) The occurrence marker layer has been isolated from the original map using the 'seperate_markers_and_map' function in the 'Marker Separation' step of HOGS. (c) Marked positions (red crosses) of all initial occurrence marker shapes in an example from Hall and Moreau (1970). (d) Different shapes initially found by the 'get_shapes_and_centres' function of HOGS. Both single and compound markers were identified.

2.1.4 | Shape and position classification

Marker shape classification

After deriving markers from the input map, we use clustering to assign shape labels to groups of occurrence markers, accommodating multiple marker types used to represent different taxonomic units, sexes or other taxon identities.

The approach used here groups markers into clusters in a three-stage process. First, all markers are reshaped to an n-by-n grid of the same size using our 'resize_shape' function. The default grid size value in our HOGS protocol is 40-by-40 pixels. This value can be modified based on the marker sizes for individual user maps. Resizing ensures that each marker has the same number of pixels, which is necessary to flatten and process them with PCA. If markers are initially smaller, then zero-value pixels are padded on the edges to the desired dimensions; otherwise, markers are scaled down, again using the 'resize_shape' function.

Second, two-dimensional marker shapes are flattened to a one-dimensional array for use in PCA using the NumPY function v2.1.0 array() in Python (Harris et al., 2020). Third, a clustering algorithm

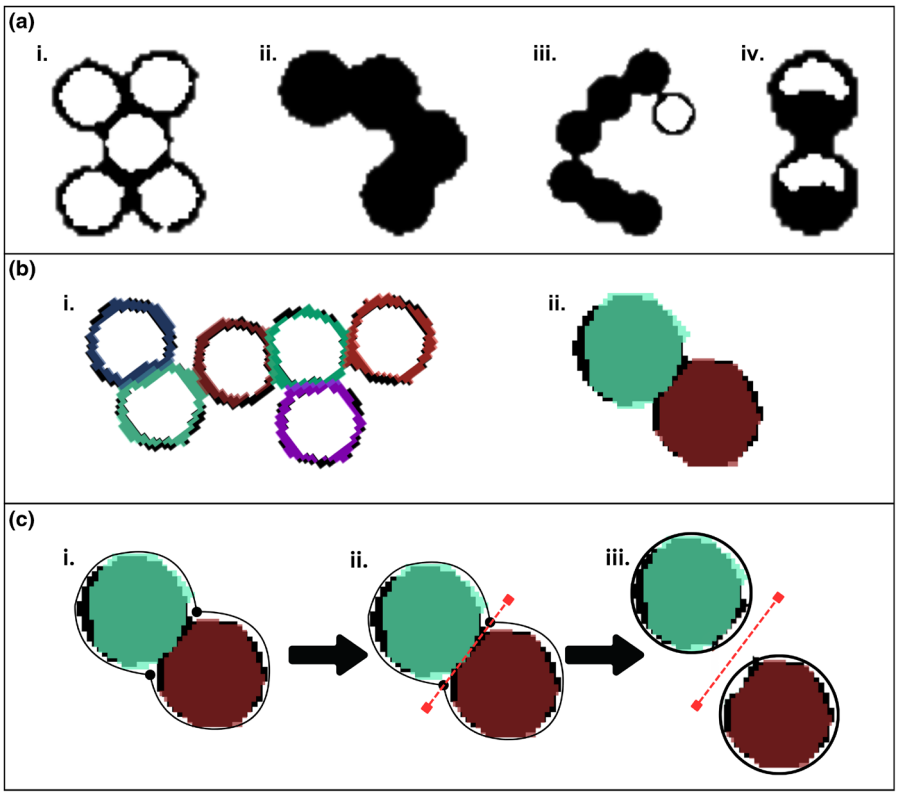

FIGURE 3 (a) Examples of compound markers derived from a single map input for three species from Hall & Moreau, 1970 (i) Five unfilled circles overlap only at edges, making shapes easy to distinguish. (ii) Four filled circles overlap more extensively, reducing inflection depth and making boundaries harder to define. (iii) Six filled circles plus one unfilled circle (different species) form a compound marker, illustrating potential cross-species data confusion. (iv) Two complex markers of the same shape overlap, showing how internal complexity and overlap compound identification challenges. (b) Compound markers, each component of this complex shape represents a single occurrence marker to be divided. (i) A compound marker composed of six overlapping unfilled circles. (ii) A compound marker composed of two overlapping filled circles. (c) Schematic demonstrating the compound marker splitting process in HOGS, which is used to define clusters of occurrence markers. These steps are defined within the function 'extract_single_shapes_and_centres'. (i) The outer shape of the compound marker is identified and the concave inflection points between the two components located. (ii) The cost function identifies the points to join across this pinch point. (iii) The two individual markers are divided along this tie-line.

(HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise, Campello et al., 2013) is used to cluster the markers based on their PCA embedding using the Python hdbscan v0.8.40 package (McInnes et al., 2017). This clustering identifies marker types that belong to a particular taxonomic unit (or other object) shown on a map. HDBSCAN does not force every marker into a cluster.Instead, some markers may be classified as noise (outliers), indicated by a label of −1, and are therefore of 'unknown' identity.

Reference map alignment

This step is relevant if multiple maps derive from the same source, with the same spatial extent. Each map from the same source is aligned to a single reference map from the dataset to automate georeferencing. The reference has markers removed via colour thresholding ('Marker Separation'). This approach assumes all maps in the set are identical, with only minor orientation or scale differences. Highly variable maps (e.g. different regions or projections) must be grouped separately. Alignment corrects small scanning and placement differences that, if uncorrected, could cause marker offsets of tens to hundreds of kilometres (Figure 4).

A tailored Bayesian optimisation function, implemented with the Optuna library v4.0.0 (Akiba et al., 2019), determines the optimal transformation parameters—translation (x, y), rotation (angle) and scaling—to align each processed map with the reference. The target function (Equation 2) minimises the difference between the transformed map and the reference. By default, the optimisation runs 300 trials, exploring translations of ±30 pixels, rotations of ±2°, and scaling from 0.99 to 1.01. These ranges and the number of trials can be modified within the 'get_transformation' function. Each trial tests a unique parameter combination, and the set that produces the closest alignment is selected.

$$1 - \frac{\Sigma(A \times B)}{\Sigma(A)} \quad (2)$$

Target function to optimise for alignment of the reference map (A) to the target map (B) in Figure 4. In this equation, A and B represent the 2D arrays of values that define the input figures. The target function

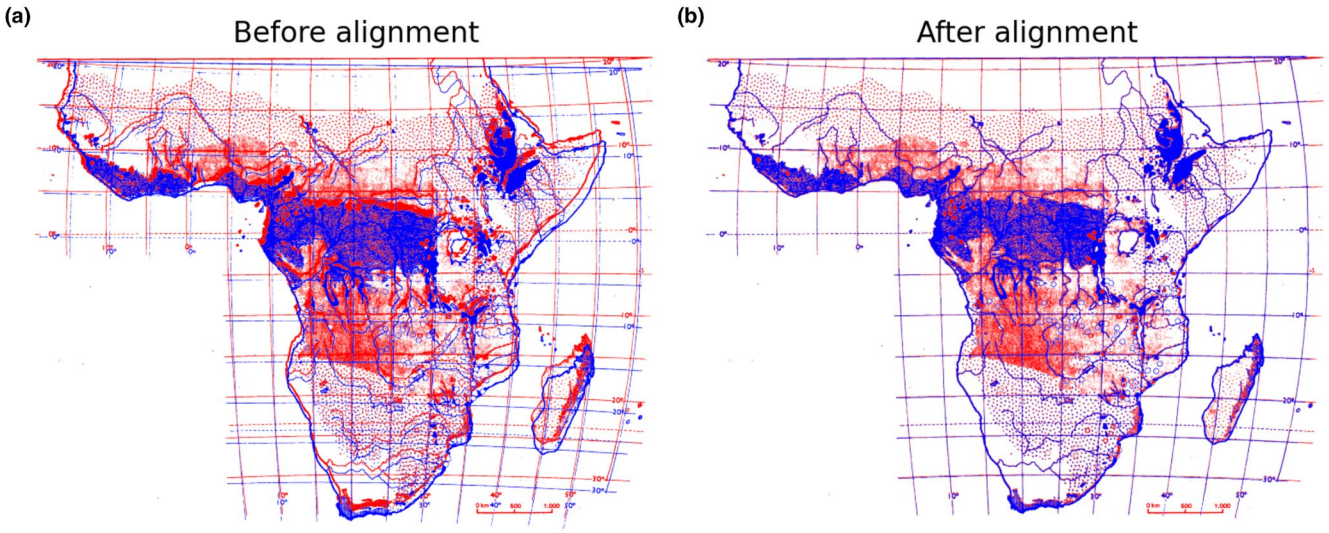

FIGURE 4 (a) Disagreement between two historical occurrence maps due to scanning/digitisation inaccuracies. The blue map represents a baseline reference map against which the red map is being aligned. Both maps are sourced from Snow (1978). (b) The results of the map alignment process, where a tailored Bayesian optimisation function ('get_transformation') is used to find the optimal transformation parameters (translation, rotation and scaling) to align all visible features to the reference map.

is at a minimum (0) when *B* is identical to *A*, and increases as the maps become more different.

2.1.5 | Coordinate conversion model

After aligning all input maps to the reference map and transforming the occurrence marker pixel coordinates using the same map-specific transformation, the pixel coordinates must be converted into geographically relevant coordinates (i.e. latitude and longitude). The default conversion used by HOGS is decimal degrees in the World Geodetic System 1984 (WGS84). The default can be altered based on user preference within the protocol.

First, a dataframe of manually selected known georeferenced points is created using the Python OpenCV v4.10.0 package (Bradski, 2000). Pixel positions can be chosen at points of interest such as graticule intersections, distinctive geographic features (e.g. mountain peaks, peninsula tips) or other features with known latitude and longitude, obtained from map content or external sources. Each point thus needs and has both geographic and pixel coordinates. A minimum of four points is required as the training input for a third-order polynomial regression model (Rawlings et al., 2006) to predict geographic coordinates (lat, lon) from pixel coordinates (*x*, *y*), although HOGS recommends ≥10 for improved accuracy.

Our protocol uses the 'scikit-learn' v1.5.2 package in Python (Pedregosa et al., 2011), which implements a third-order polynomial to model nonlinear relationships between the coordinates. Using our 'transform_points' function, the marker positions with *x*-*y* pixel coordinates are transformed to align with a reference map using the Bayesian parameters determined for map alignment ('Reference Map Alignment'). Their geographic coordinates are then predicted using the trained model, and the output is saved in .csv format along with information on marker type and original map identity.

3 | CASE STUDY

We tested HOGS using two historical distribution atlases for sub-Saharan Passerine and Non-Passerine birds (Hall & Moreau, 1970; Snow, 1978). A total of 159,852 markers were identified, categorised, and exported in an easily readable data table format from the 751 maps across the two bird atlases. The use of the compound marker splitting step accounted for 38,933 of these occurrence markers, or ~25.44% of all data extracted. The required hardware and computing time was low (~11 min, Table S5), making it accessible without large-scale computing. All analyses were run on an M2 MacBook CPU with 16 GB of RAM. Importantly, the marker extraction from multiple images is inherently parallel, making it well-suited for distribution across large systems when required. Manual digitisation averaged 15–20 min per map, whereas automated marker identification and splitting for all 751 maps took ~12 min, plus ~3 min for georeferencing and data frame creation, and ~10 min for selecting points for coordinate conversion.

We assessed HOGS' performance against manual georeferencing using four metrics (see supplement: 'Evaluating HOGS – Case Study'). HOGS performed well compared to manual georeferencing across all four metrics. Across the 10 maps, it detected 4.8% fewer markers on average (range: 0%–11.31%; Table S1), with most omissions caused by overlaps of ~90% or more, which cannot be separated by any known method. Species counts matched exactly between methods for all maps.

Positional agreement between datasets was high. NN similarity values were low (0.17°–0.87°; Table S2), averaging 0.37° for

Non-Passerine maps and 0.43° for Passerine maps, indicating that HOGS reliably reproduces manual spatial patterns. Mean spatial deviations were also low: 5.87% (longitude) and 4.37% (latitude) for Passerine maps, and 5.18% and 1.73% for Non-Passerine maps (Table S3).

Mean centre discrepancies averaged 0.14° longitude/0.37° latitude for Passerines and 0.65°/0.22° for Non-Passerines, with most values falling within marker dimensions (Table S4). Standard distance differences averaged 4.65% and 1.70%, respectively, with only Passerine map 180 showing a notably higher deviation (13.89%), consistent with its higher discrepancies across other metrics.

See supplement 'Testing Protocol Accuracy' for additional details on tests of our HOGS protocol.

4 | DISCUSSION

HOGS enables rapid generation of spatial occurrence data from historical maps. In our case study using sub-Saharan birds, it delivered a time saving of nearly three orders of magnitude with minimal data loss or alteration.

Future improvements could enhance accuracy and reduce user input. Additional testing of compound marker splitting would help to identify the most challenging combinations of shapes, overlaps, marker numbers, resolutions and orientations. Manual selection of ground control points for coordinate conversion slows processing and requires expertise; automated boundary and control point detection (e.g. Burt et al. (2020)) could remove this step. Similarly, automatic detection of map colouration (Chesneau, 2011) and resolution could streamline threshold settings for size and compound marker discrimination.

Although birds provide an ideal test case given the richness of their historical records, HOGS is broadly applicable to any taxa or region where maps contain point markers and identifiable features for georeferencing. By enabling efficient digitisation and georeferencing, HOGS makes previously inaccessible distribution data available for analysis.

This is particularly important in the context of the shifting baseline syndrome: without reliable historical baselines, present-day patterns risk being misinterpreted against already degraded conditions (Bonebrake et al., 2010). By unlocking historical biodiversity data, HOGS contributes to more accurate ecological baselines and strengthens the foundation for ecological niche modelling, biogeographical and evolutionary research, and conservation planning aligned with pre-anthropogenic conditions.

AUTHOR CONTRIBUTIONS

Harri Ravenscroft and Utku Perktas conceived the ideas. Harri Ravenscroft and John Gardner designed the methodology. Harri Ravenscroft collected the data. Harri Ravenscroft and John Gardner analysed the data. Harri Ravenscroft and Erin Saupe led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Ravenscroft, H acknowledges a NERC studentship (NE/S000/7474/1), and support from the Department of Earth Sciences, University of Oxford.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicting interests.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70166.

DATA AVAILABILITY STATEMENT

Code for execution of this protocol is available via: https://doi.org/10.5281/zenodo.17017908 (Gardner & Ravenscroft, 2025). Data is available via: https://doi.org/10.5281/zenodo.15025024 (Ravenscroft & Gardner, 2025).

ORCID

Harri Ravenscroft https://orcid.org/0009-0000-1084-1709

REFERENCES

Adobe. (2025). Hi res meaning: What is a high-resolution image? https://www.adobe.com/uk/creativecloud/photography/discover/high-resolution.html

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *arXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/1907.10902

ArcGIS Pro. (2024). https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*(January), 10–15.

Bonebrake, T. C., Christensen, J., Boggs, C. L., & Ehrlich, P. R. (2010). Population decline assessment, historical baselines, and conservation. *Conservation Letters*, *3*(6), 371–378.

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Burt, J. E., White, J., Allord, G., Then, K. M., & Zhu, A.-X. (2020). Automated and semi-automated map georeferencing. *Cartography and Geographic Information Science*, *47*(1), 46–66.

Bush, M. B. (2023). Bridging the divide from paleoecology to neoecology. *Bulletin of the Florida Museum of Natural History*, *60*(2), 57.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in knowledge discovery and data mining* (pp. 160–172. Lecture Notes in Computer Science). Springer Berlin Heidelberg.

Chapman, A., & Wieczorek, J. (2020). Georeferencing best practices. https://doi.org/10.15468/DOC-GG7H-S853

Chesneau, E. (2011). A model for the automatic improvement of colour contrasts in maps: Application to risk maps. *International Journal of Geographical Information Science*, *25*(1), 89–111.

Gardner, J., & Ravenscroft, H. (2025). HOGS: Geo-marker-extraction. *Zenodo*. https://doi.org/10.5281/ZENODO.17017908

GBIF. (2024). https://www.gbif.org.

Hall, B. P., & Moreau, R. E. (1970). *An atlas of speciation in African passerine birds*. Trustees of the British museum (Natural history).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern,

R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Heitzler, M., Gkonos, C., Tsorlini, A., & Hurni, L. (2018). A modular process to improve the georeferencing of the siegfried map. http://www.e-perimetron.org/Vol_13_2/Heitzler_et_al.pdf

Huang, J., & Frimpong, E. A. (2015). Using historical atlas data to develop high-resolution distribution models of freshwater fishes. *PLoS One*, *10*(6), e0129995.

Kemp, M. E. (2023). Characterizing past communities to build future ones: Lessons from caribbean conservation Paleobiology. *Bulletin of the Florida Museum of Natural History*, *60*(2), 58.

Kiessling, W., Raja, N. B., Roden, V. J., Turvey, S. T., & Saupe, E. E. (2019). Addressing priority questions of conservation science with Palaeontological data. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1788), 20190222.

Marcer, A., Chapman, A. D., Wieczorek, J. R., Picó, F. X., Uribe, F., Waller, J., & Ariño, A. H. (2022). Uncertainty matters: Ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. *Ecography*, *2022*(9), e06025.

Marcer, A., Haston, E., Groom, Q., Ariño, A. H., Chapman, A. D., Bakken, T., Braun, P., Dillen, M., Ernst, M., Escobar, A., Fichtmüller, D., Livermore, L., Nicolson, N., Paragamian, K., Paul, D., Pettersson, L. B., Phillips, S., Plummer, J., Rainer, H., … Wieczorek, J. R. (2021). Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. *Diversity and Distributions*, *27*(3), 564–567.

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205.

Nelson, G., & Ellis, S. (2018). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1763), 20170391.

Nowak, M. M., Słupecka, K., & Jackowiak, B. (2021). Geotagging of natural history collections for reuse in environmental research. *Ecological Indicators*, *131*(108131), 108131.

Parker, E. J., Weiskopf, S. R., Oliver, R. Y., Rubenstein, M. A., & Jetz, W. (2024). Insufficient and biased representation of species geographic responses to climate change. *Global Change Biology*, *30*(7), e17408.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science (New York, N.Y.)*, *344*(6187), 1246752.

QGIS. (2024). http://www.qgis.org.

Ravenscroft, H., & Gardner, J. (2025). HOGS: Passerine and non-passerine map dataset. *Zenodo*. https://doi.org/10.5281/ZENODO.15025024

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2006). Polynomial regression. In *Springer texts in statistics* (pp. 235–268). Springer-Verlag.

Seltmann, K., Lafia, S., Paul, D., James, S., Bloom, D., Rios, N., Ellis, S., Farrell, U., Utrup, J., Yost, M., Davis, E., Emery, R., Motz, G., Kimmig, J., Shirey, V., Sandall, E., Park, D., Tyrrell, C., Thackurdeen, R. S., … Nyberg, B. (2018). Georeferencing for research use (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers. *Research Ideas & Outcomes*, *4*, e32449. https://doi.org/10.3897/rio.4.e32449

Snow, D. W. (1978). *An atlas of speciation in African non-passerine birds*. British Museum Press.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292.

Szabo, J. K., Khwaja, N., Garnett, S. T., & Butchart, S. H. M. (2012). Global patterns and drivers of avian extinctions at the species and subspecies level. *PLoS One*, *7*(10), e47080.

Thorn, C. (2022). Georeferencing historic collection data. *Biodiversity Information Science and Standards*, *6*, e91578. https://doi.org/10.3897/biss.6.91578

Tobias, J. A., Sheard, C., Pigot, A. L., Devenish, A. J. M., Yang, J., Sayol, F., Neate-Clegg, M. H. C., Alioravainen, N., Weeks, T. L., Barber, R. A., Walkden, P. A., MacGregor, H. E. A., Jones, S. E. I., Vincent, C., Phillips, A. G., Marples, N. M., Montaño-Centellas, F. A., Leandro-Silva, V., Claramunt, S., … Schleuning, M. (2022). AVONET: Morphological, ecological and geographical data for all birds. *Ecology Letters*, *25*(3), 581–597.

Venceslau, L., & Lopes, L. (2019). Comparison of automated georeferencing tools using insect collection data. *Biodiversity Information Science and Standards*, *3*(July). https://doi.org/10.3897/biss.3.37345

Zatelli, P., Gabellieri, N., & Besana, A. (2022). Digitalization and classification of Cesare Battisti's Atlas of 1915. *ISPRS International Journal of Geo-Information*, *11*(4), 238.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1. The number of occurrence markers identified per test map by automated (HOGS) and manual georeferencing, and the counts of the number of species (identified by a change in marker shape) for the same maps.

Table S2. Nearest neighbour similarity in degrees, derived from marker longitude-latitude positions (WGS84) for each of 10 test maps using automated (HOGS) and manual georeferencing.

Table S3. The decomposed standard distance derived from marker longitude-latitude positions (WGS84) per test map using automated (HOGS) and manual georeferencing.

Table S4. Mean centre and standard distance from centre comparisons for each of the 10 test maps using automated (HOGS) and manual georeferencing.

Table S5. Processing times for each protocol stage.

How to cite this article: Ravenscroft, H., Gardner, J., Perktas, U., & Saupe, E. (2025). Release the HOGS: An unsupervised marker extraction, classification and georeferencing approach for biodiversity data. *Methods in Ecology and Evolution*, *00*, 1–8. https://doi.org/10.1111/2041-210X.70166