Methods in Ecology and Evolution

DR ARDERN HULME-BEAMAN (Orcid ID : 0000-0001-8130-9648)

Article type : Research Article
Editor : Dr Laura Graham

GeoOrigins: A new method and R package for trait mapping and geographic provenancing of specimens without categorical constraints

Ardern Hulme-Beaman*†[1,2], Anna Rudzinski*[3], Joseph E. J. Cooper[4], Robert F. Lachlan[5], Keith Dobney[1,6,7,8], Mark G. Thomas[3,9]

* these authors contributed equally

† Corresponding author

Ardern Hulme-Beaman

Email: ardernhb@gmail.com

Address: Department of Archaeology, Classics and Egyptology, University of Liverpool, 12–14 Abercromby Square, Liverpool, L69 7WZ, UK

Telephone number: +44 7783400278

[1] Department of Archaeology, Classics and Egyptology, University of Liverpool, 12-14 Abercromby Square, Liverpool, L69 7WZ, UK

[2] Research Centre in Evolutionary Anthropology and Palaeoecology, School of Natural Sciences and Psychology, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK

[3] Research Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK

[4] School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK

[5] Department of Psychology, Royal Holloway University of London, Surrey, TW20 0EX, UK

[6] Department of Archaeology, University of Aberdeen, Aberdeen, AB24 3UF

[7] Department of Archaeology, Simon Fraser University, Burnaby, British Columbia, Canada

[8] School of Philosophical and Historical Inquiries, University of Sydney, NSW 2006 Australia.
[9] UCL Genetics Institute, University College London, London, WC1E 6BT, UK

Running title:
GeoOrigins: Trait mapping and provenancing

Abstract

1. Biologists often seek to geographically provenance organisms using their traits. This is typically achieved by defining spatial groups using distinct patterns of trait variation.
2. Here we present a new spatial provenancing and trait boundary identification methodology, based on correlations between geographic and trait distances, that requires no *a priori* group assumptions. We apply this to three datasets where spatial provenance is sought: morphological rat and vole dentition data (human translocation datasets); and birdsong data (cultural transmission dataset). We also present the results of cross-validation testing.
3. Spatial provenancing is possible with differing degrees of accuracy for each dataset, with birdsong providing the most accurate geographic origin (identifying an average spatial region of 0.22km^2 as the area of origin with 99.9% confidence).
4. Our method has a wide range of potential applications to diverse data types — including phenotypic, genetic and cultural — to identify trait boundaries and spatially provenance the origin of unknown or translocated specimens where trait differences are geographically structured and correlated with spatial separation.

Keywords: Spatial mapping; Trait mapping; Identification; Provenancing; Morphology; Biogeography; Phylogeography

1. Introduction

Tracking changes in the spatial distribution of organisms and their traits is a central feature of biogeographical research. Such studies include exploring human-mediated translocation and/or natural dispersal of organisms (e.g. Cucchi, 2008; Cucchi et al., 2014; Frantz et al., 2018; Lachlan et al., 2013) and establishing the geographic origin of human introduced invasive and commensal species (e.g. Gargan et al., 2016; Hooten and Wikle, 2008; Hunt et al., 2018; Jones et al., 2013). These studies regularly make use of quantifiable character traits, such as genotypes and phenotypes (including morphologies and behaviours, e.g. Lachlan et al., 2013) of specimens with known geographic provenance. Here we introduce the GeoOrigins approach and R package, which provide new spatial provenancing and trait boundary identification methods that consider continuous patterns of variation rather than imposing discrete groups.

Most spatial provenancing methods require separation of reference material into discrete groups (taxonomic units or populations), with little consideration of admixture between groups or a continuum of trait variation across a range. Consequently, these discrete groups are often synthetic and of questionable biological validity, particularly in the case of populations. Furthermore, summarising geographic location information (i.e. latitude/longitude) for grouped individuals results in lost information, as geographic groupings artificially collapse ranges of varying sizes, which may be bounded by different geographic features (e.g. mountains or rivers) that do not represent equally strong barriers to gene-flow. Therefore, spatial provenancing methods that do not require assignment to specified groups are needed to avoid information loss of trait and geographical data.

A number of methods for group assignment are well established. Posterior probabilities of a specimen belonging to *a priori* defined groups extracted from linear discriminant analyses (LDA) developed by Fisher (1936) are among the most common approaches for assigning specimens to a given geographic location (e.g. Evin et al., 2013). These methods can be susceptible to overestimation if the number of variables is greater than the number of individuals in the smallest group (Mitteroecker and Bookstein, 2011), although this can be assessed using leave-one-out correct cross-validation (CCV) approaches (Tukey, 1958). This presents a problem for datasets that have many variables (e.g. geometric morphometrics), which necessitates dimensionality reduction and therefore loss of information (Mitteroecker and Bookstein, 2011 and references

therein) — although dimensionality reduction can provide other benefits such as the removal of variation stemming from data collection biases, and various sources of noise (e.g. Claude, 2013).

Distance based methods (such as *k*-nearest neighbour [*k*-NN] classification methods) provide a non-parametric alternative (Altman, 1992). *K*-NN methods assign group membership of unknown specimens using the majority vote of a set number (*k*) of nearest neighbours with known group membership (Ripley, 2007). For example, for a dataset with reference specimens representing two assumed groups (A and B), with a *k* of 10, the *k*-NN approach will assign an unknown specimen to group A if greater than 5 of its nearest neighbours are known members of group A (e.g. see SI R code for example). However, like LDA, *k*-NN approaches make similar discrete group assumptions and require user-defined classes for the reference data. Furthermore, the user-defined *k* can dramatically affect outcomes (e.g. Baylac and Friess, 2005; Guillaud et al, 2016 and references therein), particularly due to different sampling densities of trait spaces. Therefore, as they require discrete category approaches, *k*-NN and LDA remain limited.

Here we present GeoOrigins, a new R package containing functionality required to implement a novel provenancing and boundary finding method. Our correlation-based method provides an alternative to discrete group based methods (e.g. *k*-NN and LDA) and does not require *a priori* categorisation of specimens. However, our method can also integrate well with those existing discrete group based methods. We apply our new methods to three different datasets that include two shape datasets and one birdsong dataset. We empirically test our methods with specimens of known origins and propose ways the methods might be integrated into future studies.

2. Materials and Methods

2.1 A new biogeographical provenancing method

Our method first requires the calculation of a distance or dissimilarity vector based on one or more quantifiable traits between samples of known (reference specimens) and unknown (specimen of interest) geographic provenance. We avoid specifying a distance measure here, as many are available, and should be chosen according to the data considered. The trait distance vector can be based on continuous and/or discrete character trait data using Euclidian, Jaccard, or a range of other means of summarising difference (hereafter 'trait distances') between the known georeferenced samples and the test sample. If a similarity score is used (for example, Jaccard

indices as might be used in the quantification of similarities among cultures, e.g. Shennan et al., 2015; or birdsong repertoires, e.g. Lachlan and Slater, 2003) the corresponding distance must be calculated accordingly (i.e. low values indicate similarity and high values indicate difference). Under the assumption that there is a correlation between trait difference and spatial distance — as expected under a wide-range of dispersion models including simple isolation by distance (Nei, 1972; Wright, 1943) — the geographic location where that correlation is maximised (when comparing known georeferenced samples and test samples) marks the most likely origin location of the test sample.

To identify this location, a spatial grid can be defined within which all reference specimens are present — including plausible origin regions for the test sample (Figure 1) — and where the latitude/longitude position of each reference specimen is noted as ($\boldsymbol{x}_n$, $\boldsymbol{y}_n$), where n corresponds to the nth reference specimen. This spatial grid is a matrix A, where rows $\boldsymbol{i}$ and columns $\boldsymbol{j}$ represent latitude and longitude values respectively, and should be of sufficient size to gain good spatial resolution without making exploration prohibitively expensive in computational resources. The vectors of $\boldsymbol{i}$ and $\boldsymbol{j}$ are defined as containing at minimum:

$$\{\boldsymbol{i} \mid \boldsymbol{y}_{min} \ldots \boldsymbol{y}_{max}\}$$
$$\{\boldsymbol{j} \mid \boldsymbol{x}_{min} \ldots \boldsymbol{x}_{max}\}$$

For each element of matrix A, a vector $\boldsymbol{g}$ of n geographic distances is calculated from the point at A_{ij} to the latitude/longitude position of each reference specimen ($\boldsymbol{x}_n$, $\boldsymbol{y}_n$). As these distances are spatial, $\boldsymbol{g}_1 \ldots \boldsymbol{g}_n$ are estimations of distances across the curved surface of the Earth and should be calculated using the haversine formula (Robusto, 1957). Elements of this matrix A are populated by calculating the correlation coefficient (either Spearman's *rho* or Pearson's *r,* depending on assumptions of linearity see SI. 1) for the correlation between these spatial distances ($\boldsymbol{g}$) and trait distances ($\boldsymbol{d}$) (Figure 1, following the correlation plotting method described in Frantz et al., 2018):

$$A_{ij} = corr(\boldsymbol{g}, \boldsymbol{d})$$

This calculation can be made for each reference specimen at that specimen's true latitude/longitude location in a CCV approach. The resulting distribution of r values can be

examined to set the threshold *r* value required to correctly spatially provenance a specimen of interest with a given level of confidence. For example, if the desired confidence level for spatial provenancing is 95% then the *r* value that correctly provenances 95% of the reference specimens can be extracted. In this way the method uses an empirical approach to spatially provenance specimens and, therefore, estimate trait boundaries. We then make the assumption that the correlation of trait and geographic distances will be equal to or greater in the region of the grid covering the unknown specimen's origin. We found the results of Pearson's and Spearman's to be approximately the same and for brevity we present just the results generated from Pearson's *r* here (results can be compared with those in the vignette that uses Spearman's).

2.2 Mapping trait boundaries

Once the threshold *r* value is set, we can apply that threshold to the reference material to generate a set of intersecting polygons. Where the edges of those provenancing regions show substantial overlap among individuals, a trait boundary can be defined. This can be mapped by taking the vector of all the correlation values for every specimen at each grid location and counting how many times that grid location is at the boundary of our chosen *r* threshold (See SI.2). The resulting counts can be given a grey/colour scale and broad boundaries can then be interpreted from the resulting map. Note that the plotted result of this approach will show boundaries of varying strength, which can be influenced by uneven regional sampling. This is because as the boundary plotting approach is a relative scale, if one region with an associated trait is more thoroughly sampled than others, the boundary for the thoroughly sampled region will be more confidently identified.

2.3 Assessing method performance

To illustrate the utility of our spatial provenancing methodology, we analysed three datasets at different geographic resolutions. To summarise and assess the results we calculated the area of overlap between the estimated origin range at 95% confidence and the convex hull (i.e. maximum parameter) of the range represented by the distribution of the reference specimens. We consider the convex hull of the reference material distribution as a conservative approximation of the distribution range of the examined taxa. We then calculated the percentage of total reference distribution that was a likely origin for unknown specimens. This CCV procedure provides an empirical metric of the ability to spatially provenance an unknown specimen. Where appropriate

and where results of provenancing estimations provided high precision (i.e. the Tenerife blue chaffinches [*Fringilla teydea*] dataset), we expanded the CCV test to assess overfitting. This was carried out by iteratively subsampling 75% of specimens and treating that subset as the reference specimens for training the spatial provenancing method; the remaining 25% of specimens were then treated as those to be spatially provenanced. The percentage of specimens of interest correctly provenanced was then calculated and the distance from the true point of origin to both the nearest edge of the provenancing region and its centroid was calculated to assess how the method performed when the estimated provenancing region does not include the true location.

2.4 Comparison with nearest neighbour and grouping approaches

The new spatial correlation methods described here inherently avoid making *a priori* group assumptions; as such, direct comparison with those methods that do so (e.g. LDA and *k*-NN) is not possible. However, the trait boundary identification methods presented here can inform potential groupings for consideration as evolutionary units or for use in subsequent LDA and *k*-NN classification. Therefore, the spatial trait groups identified were compared with the CCV% achieved from LDAs and *k*-NN. For LDA comparisons, we calculate the mean CCV% result from a resampling procedure to equal sample size (1000 times) for each stepwise combination of principal components for the shape data (following Evin et al. 2013) and multi-dimensional scaling variables (using the vegan package; Oksanen et al. 2017) for the birdsong data. We then report the maximum of these mean stepwise CCV% values. *K*-NN methods are applied to the Procrustes distances for the shape datasets and are applied directly for the birdsong dataset. *K*-NN analyses are carried out on groups of equal sample size (resampled 1000 times) with the package KnnDist (Hulme-Beaman, 2020) and applied with a stepwise increase in *k*. The maximum mean CCV% calculated from the stepwise increase in *k* is reported in the same way as for the results of the stepwise LDA.

2.5 Test datasets

The three worked examples comprise two different data forms: shape and birdsong recording data. All specimens are of known origin, with known sampling locations and associated latitude and longitude data; therefore all results presented here are in effect CCV exercises. For shape, we used two geometric morphometric datasets of dental morphology: 48 New Guinea large spiny rat (*Rattus praetor*) specimens (Hulme-Beaman et al. 2018); and 553 common vole (*Microtus arvalis*)

specimens (Cucchi et al., 2014). For these datasets we aligned, processed and generated Procrustes distances between shape configurations using R and the package 'shapes' (Dryden, 2016). For birdsong, log transformed dynamic time-warping dissimilarities between song type and repertoires were generated from recordings of 116 Tenerife blue chaffinches using Lucinia v2.16.10.29.01 (Lachlan et al., 2013, Lachlan, 2016 http://rflachlan.github.io/Luscinia/). For packages used to plot these maps and those that were used to construct this package see the supplementary information (SI.3). All distance matrices, code and functions written for this paper are published in the supplementary information and the corresponding R package "GeoOrigins" (See SI.2).

3. Example applications

3.1 Dental morphology: Rattus praetor, *a possible species complex within Sahulian* Rattus

The large spiny rat, *R. praetor,* is distributed across New Guinea and the neighbouring islands, including the Bismarck Archipelago and the Solomon Islands. Recent shape analyses of their teeth revealed geographic structure with a general east–west cline (Hulme-Beaman et al., 2018). This presents an interesting dataset for future studies into human migration since *R. praetor* was introduced to remote Oceania by humans (White et al., 2000). Applying our spatial provenancing method to the dental morphology reveals a similar east–west pattern of geographic structure with Pearson's *r* with specimen origins generally identified to approximately either side of the 145th meridian east (Figure 2A & 2B). At the 95% threshold the true location of three specimens (5%) fell outside the provenanced region. These specimens were located an average of 280km from their true location to the nearest boundary of the estimated provenance region. Two of the three specimens, whose true location fell outside the estimated range, were located within ~150km of the 95% confidence boundary (SI.4.1). The third specimen's true origin was ~500km away from the closest provenancing boundary edge (SI.4.1). A further six instances returned a range encompassing almost the entirety of the region. The method provided an approximate origin with reduced spatial area to an average of 63% of the total range represented by the reference material (Figure 2C). With so few specimens being incorrectly provenanced it is difficult to discern a trend, but it is notable that the misidentifications are within the central distribution of the species across the possible west to east morphological gradient. Given that six specimens returned the entire range, it is likely that the method does not have sufficient numbers or evenly distributed reference material to build a more confident model. As a result, it is possible that given more sampling of

the central region of the range, a better definition of the morphological trait boundaries and/or gradient would be achieved.

To integrate and compare these results with LDA and *k*-NN analyses we created two groups east and west of the 145th meridian as identified by the trait boundary identification exercise; each had a sample size of 24 specimens. Maximum discrimination was achieved at 90% with 11PCs using LDA and 83% with 6 weighted nearest neighbours. For comparison, when the dataset was grouped by specimens from Bougainville Island (to the east) versus those from New Guinea, the LDA and *k*-NN CCV% were improved to 92% and 87%. This improved rate of identification to 92% for LDA does not reach the heatmap result we achieved with 95% confidence. However, when set to 95% confidence, our method returned the entire examination region for a number of specimens; as the LDA approach includes groups of different spatial areas it is in some regards more precise, e.g. Bougainsville Island is smaller than the ranges returned by our method, but less precise in other instances, e.g. the area of the entirety of New Guinea is much greater than the area returned by our identification method. The discrepancies between methods likely result from the presence of poorly sampled population(s) in central and eastern New Guinea. The trait boundary found here is likely to be largely influenced by the extensive sampling at both ends of the range. This is likely a common problem for museum specimens, particularly for human commensal species, where many specimens might be collected from one location and, as a result, have a single latitude/longitude value. However, our method highlights this, since specimens in central regions prove to be more difficult to correctly provenance, e.g. the central New Guinea specimens, and if a trait is poorly represented in the reference dataset the method returns the entire region.

*3.2 Dental morphology: Common voles (*Microtus arvalis*) and Orkney voles (*M. a. orcandensis*)*

M. a. orcandensis colonised the Orkney Islands around 5Kya, likely arriving with Neolithic farmers (Cucchi et al. 2014). These island populations have since rapidly diverged in both size and shape from each other, as well as from their ancestral European counterparts (Cucchi et al., 2014). This dataset provides an example of how this method can be used to assess: 1. if a trait boundary is formed by the divergence of island and mainland dental morphologies; 2. if geographic structuring of continental populations exists; 3. whether future studies of ancient mainland European common voles would be informative if assessed for similarities to Orkney populations.

We applied the method at two different spatial resolutions: across the entire species range, and within the distribution of each respective subspecies (i.e. continental European and Orkney populations).

When considered as percentages of the total species distribution, the CCV results of our method are strongly bimodal at the species-wide resolution. This is unsurprising given the highly localised and morphologically distinct populations of *M. a. orcandensis* when compared with the much more widely distributed populations of continental European *M. arvalis* (Figure 3A). This pattern is consistent with the findings of previous morphological and genetic analyses (Martínková et al., 2013). Splitting the dataset into Orkney versus European voles found the maximum CCV% from LDA was 98% with 27PCs, and 96% from *k*-NN classification with 17 weighted nearest neighbours.

Subsetting the data to examine the 131 *M. a. orcandensis* specimens increased provenance resolution within the Orkney archipelago, and narrowed provenance down to an average of 51% of the total Orkney Island distribution (Figure 3B). Splitting the Orkney dataset into northern versus southern islands along 59.1º latitude found the maximum CCV% from LDA was 97% with 18PCs and 96% from *k*-NN classification with 15 weighted nearest neighbours; this compares with 92% LDA CCV% and 86% *k*-NN CCV% when individual islands are used as grouping categories. This demonstrates that our method can be used to inform group assignment exercises (e.g. those using LDA or *k*-NN) and although those methods continue to perform less well compared with our method at 95% confidence, the trait boundary output of our method can be used to merge groups and, therefore, improve confidence in classification by those other methods, if required.

Compared with the Orkney analysis, the same level of spatial provenancing was not possible for most of the 387 mainland European *M. arvalis* specimens, where 26% of specimens could not be provenanced to any location beyond a convex hull of the entire distribution of the reference samples. Of the remaining specimens, the provenancing method returned between 60–90% of the total distribution (Figure 3C and SI.5). This illustrates the tendency for our approach to identify origin location to different degrees when different geographic scales are considered. At higher geographic resolutions, the method still performs well if geographic structure persists at that level and if sufficiently detailed latitude/longitude data are available. However, not all island

populations are equally similar/dissimilar, with the strongest difference between northern and southern populations (Figure 3D). Again, as with much of the *R. praetor* location data, precise latitude/longitude data for voles within the Orkney archipelago were not available, and this hampered fine resolution analyses. As a result, it is difficult to assess whether more precise data would assist in refining spatial identifications and thus trait boundaries. In this case study, the complex life-history of *M. arvalis* has likely rendered poor geographic structuring (and thus resolution) within mainland Europe and the better resolution of its insular forms may instead represent a 'snap-shot' in time of their past continental diversity (Martínková et al., 2013). Applying our method on different scales does, therefore provide different levels of information.

*3.3 Birdsong: Tenerife blue chaffinch (*Fringilla teydea*)*

F. teydea colonised Tenerife approximately 2 Mya (Lifjeld et al., 2016). Males learn songs through imitation of neighbours and errors in imitation result in localised innovations in song structure. Characterising structural change across a landscape is highly desirable for understanding cultural evolution of song and also dispersal ecology. This dataset illustrates the application of our method on a culturally transmitted trait, as opposed to genetically inherited ones. We applied the method at three resolutions: 1) low resolution across the entire island (to form *sensu lato* isoglosses); 2) medium resolution within regional variants identified from the low resolution analyses; and 3) at a high resolution looking at densely sampled sub-regions of wider isoglosses to identify each bird's most likely tutor location. Origin ranges generated from the CCV procedure varied widely (Figure 4). Three main regions were identifiable as having accumulated sufficient song variants to make them distinguishable from each other. As the specimens from central Tenerife did not appear to fall into a clear group, possibly due to low sampling, these were removed for comparison with LDA and *k*-NN methods. The *k*-NN and LDA CCV%s for the three spatial groups were 80% with 2 weighted nearest neighbours and 100% with two multidimensional scaling axes.

At the highest resolution, it was possible to correctly provenance a bird's song to an average of $0.22km^2$ by log transforming the dissimilarity matrix. This extremely high degree of accuracy was tested for over-fitting by training the method on a randomly selected subset of 75% of specimens and testing with the remaining 25%. This procedure was run 600 times on both the north-eastern

subpopulations and southern subpopulations; the north-western subpopulation was too sparsely sampled to examine in this way. This meant that in each iteration of the resampling procedure, the number of specimens being treated as of unknown origin in the north-eastern subpopulation was seven and in the southern population was 11. As a result, the over-fitting resampling procedure made a total of 5,400 spatial provenancing identifications. The method continued to perform well, and accurately provenanced the location of 83% of specimens considered unknown. Of the 17% that were not successfully provenanced, the true origin of 5% of those specimens was not found (i.e. no region met the *r* threshold for provenancing) and 12% were incorrect (i.e. the true origin was outside the provenancing boundary). However, in the cases of incorrect provenancing results, the distance between the true origin and the provenancing area was often shorter than the diameter of a bird's territory (~92–112m; Carrascal, 1987; García del Rey and Cresswell, 2005). Considering the precision of learning exhibited in these birds, although the provenancing is incorrect in some cases, the likelihood of a bird being in the identified provenancing location having a near identical song is extremely high. This demonstrates the possible predictive power of this method where trait and geography are highly correlated. Comparison between our method and other discrete group-based methods was not possible or appropriate at this high resolution.

For incorrectly provenanced specimens, the distance from the true location to the region identified by the method as the most likely area of provenance can be estimated in two ways: 1. the shortest distance to the boundary; 2. the distance to the centroid of the suggested region of provenance. As such small regions were returned at this high resolution, the difference between the distance to closest boundary and the centroid of the area was very small (Figure SI.6.1); the median distance to the nearest edge was 35.8m, whereas the median distance to the centroid was 37.0m. The maximum incorrect distance from the true location to the centroid of the proposed location was 217.0m, which is a shorter distance than 91% of distances among locations of birds in each respective subpopulation dataset.

This indicates that given good sampling and high geographic structure, this method should be useful for identifying likely tutors, possible long-distance dispersers and understanding differences in localised adaptation of birdsong. Not all specimens conform to the geographic distribution, thus violating the monotonic assumption, so could either not be identified or the entire investigation grid was returned as a likely origin (see Discussion for hypothetical scenario of assumption

violations). These specimens could result from factors such as convergence, innovation or otherwise unrecognised long-distance dispersal.

4. Discussion

4.1 Assumptions of monotonicity

The blue chaffinch song dataset provided the most accurate and precise results because of the extremely strong monotonic spatial distance correlation with song similarity. Results were improved with log transformation to below 500m^2 in some cases. The high level of monotonicity in these data in the boundary finding exercise proved to be problematic because in some instances, particularly at the small regional scale, the spatial grid had to be sampled to such a high level that the amount of computational time required to locate the boundaries became prohibitive. Therefore, if (as in the blue chaffinch song case) the signal is so strong that a very precise region is identifiable, then the results can provide such specific origin locations that general boundary trends become unidentifiable (i.e. each reference specimen exists within its own trait unique boundary). This effect is also more likely to be seen where the assumption of linearity is not required, and the correlation method uses Spearman's *r*. In some cases where this occurs, origin identification can also be missed because the high level of accuracy and precision means the predicted origin region can be smaller than the grid square. In such a case the provenancing threshold may not be reached at the nearest sampled grid square.

Application of our spatial provenancing approach to the mainland European voles provides an example of the method's response to breaking the assumption of monotonicity. There is little to no consistent geographic structure in the mainland European vole populations and, as a result, spatial provenancing is often not possible with useful levels of confidence. This can be observed by plotting traits distances versus their corresponding geographic distances at the true location of a specimen being treated as unknown (Figure SI.5.1). Each dataset had different levels of isolation by distance characteristics. Of the case studies, the mainland European vole example had the poorest ability to reduce the species range to a likely area of provenance (Figure SI.5.1A). In contrast, the inclusion of Orkney Island voles to the dataset creates a clear pattern of isolation by distance with a largely monotonic relationship at the place of origin (e.g. see Figure SI.5.1B). Of the case studies, the blue chaffinch song dataset has the clearest isolation by distance pattern with the strongest monotonic relationship between trait and geographic distance (Figure SI.5.1C). The

provenancing output for mainland European specimens tends to return large proportions of the region being examined (e.g. Figure SI.5.2). This demonstrates that with varying degrees of monotonicity and geographic structure to traits, the method will respond differently. However, when provenancing is not possible the method returns most if not all the region, making erroneous provenancing or identification less likely.

4.2 Hypothetical problematic trait scenarios

The provenancing and boundary identification method presented here will struggle when trait distance does not have a monotonic relationship with geographic distance. This is because the method will only return one likely origin region and so if there are multiple and spatially distant regions that an unknown specimen appears similar to in trait comparisons, then all regions will be returned in a single large polygon. Here we expand on this and provide a hypothetical example of where trait boundaries may exist but are unlikely to be detected by this method. A likely scenario is where a trait is the result of local climatic adaptation. In such a scenario it may be the case that where climatic conditions are matched in distant geographical regions, the same trait will occur in both populations. As a result, any provenancing approach using this method will be unable to distinguish between the two spatially distant locations and, as a result, the trait distance distribution will not be monotonic.

A hypothetical example scenario is as follows: a species distribution is bounded by mountain ranges at the opposite extremes of the species' distribution. The trait of interest is associated with cold and high elevation adaptations, e.g. coat characteristics. The far eastern and western mountain range populations will share similar traits and therefore have short trait distances. The methods presented here will be unlikely to distinguish between reference specimens occupying these two very spatially distant locations. There may in fact be a trait boundary at a certain elevation, but because this trait is shared by two spatially distant populations, this will blur the trait boundary and the method will likely fail to identify the elevational trait boundary.

4.3 Method performance and integration with existing methods

Our method performed well in comparison to conventional classification methods (*k*-NN and LDA). The methods are very different in their approach, thus direct comparison is not possible.

Inherently our method is better suited to identification of a handful or few individuals, since provenancing is achieved empirically using a heat map. As such the area returned for provenancing an individual will vary in size depending on the strength of signal. In contrast, a discrete group based classification method can be both more precise in some instances (where groups occupy small patches of a region) but also less precise (where some groups may represent large regions), for example in the *R. praetor* case study. As our method can have a set confidence level (which can be set to 1), the area returned has the potential to ensure maximum confidence in classification and will provide an empirically derived, though potentially large, region of likely origin.

Integration of our method may prove most useful when provenancing many unknown individuals at once, or when identification of evolutionary units is desirable. In cases where multiple specimens are to be provenanced either the mean or median values for the sample of unknowns could be analysed. However, a likely more robust and efficient approach is to use the trait boundary finding method to define discrete spatial trait groups in the reference data and then use these trait boundary defined groups in methods such as LDA and *k*-NN for classification of multiple specimens at once.

Where traits are shared among spatial populations but at different frequencies, it might be desirable to assess the different trait frequencies spatially by carrying out the trait boundary finding exercise, but lowering the required correlation value to the required confidence. Here we have set the required confidence to 95%, which means boundaries returned in the case studies here require trait frequencies between populations to be different by 95% or higher before a boundary will be plotted. However, if traits are shared among populations at a lower frequency, and this is of interest to the user, the confidence level can be adjusted; this requires further investigation and exploration as such questions were not relevant to the datasets we examined here. In this way the methods we present should integrate well with existing frameworks, particularly where geographic divisions are desired, but would otherwise need to be constructed arbitrarily or subjectively.

5. Conclusions

Our method provides a useful and robust geographic provenancing tool that takes into consideration the confidence with which a given specimen of unknown origin can be spatially located. As it can be applied to any set of distances constructed between any set of traits, our method has a wide range of potential uses in multiple different fields where spatial provenancing is desired. This goes beyond applications in ecology and evolution and, as demonstrated in the instance of birdsong, can also be applied to spatially provenance organisms based on cultural traits and characteristics (e.g. human material culture) given sufficient reference data. Furthermore, this method provides insights into geographic structuring of traits, with the possibility of identifying particularly different populations that are geographically well-bounded and this can be integrated with other existing methods. Our method should, therefore, prove valuable to future geographic studies across multiple fields of research.

Acknowledgements

AHB was funded by a Leverhulme ECR fellowship (ECF-2017-315). AR was supported by a Marie Curie Initial Training Network (BEAN – Bridging the European and Anatolian Neolithic, GA No. 289966) awarded to MGT. MGT was supported by a Wellcome Trust Senior Research Fellowship, grant 100719/Z/12/Z.

Data availability

The distance matrices used in this paper are available as part of the package hosted on the Github repository (https://github.com/ArdernHB/GeoOrigins —DOI: 10.5281/zenodo.3919325). The raw landmark data for the *Rattus praetor* dataset is also sorted with the package repository; the raw landmark data for *Microtus arvalis* is available in the online supplementary information for the original publication Cucchi et al. 2014.

Author contributions

AHB, AR and MGT conceived of and designed the methodology; AHB wrote the R package and carried out the analyses; JC, RL and KD collected and provided the datasets used; AHB led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

Baylac, M., Friess, M., (2005). Fourier Descriptors, Procrustes Superimposition, and Data Dimensionality: An Example of Cranial Human Populations, in: Slice, D.E. (Ed.), Modern Morphometrics in Physical Anthropology. Springer-Verlag, New York, 145–165.

Carrascal, L.M., (1987). Relacion entre avifauna y estructura de la vegetacion en las repoblaciones de coniferas de Tenerife (Islas Canarias). *Ardeola* 34, 193–224.

Claude, J. (2013). Log-Shape Ratios, Procrustes Superimposition, Elliptic Fourier Analysis: Three Worked Examples in R. *Hystrix, the Italian Journal of Mammalogy*, *24*(1), 94-102.

Cucchi, T., (2008). Uluburun shipwreck stowaway house mouse: molar shape analysis and indirect clues about the vessel's last journey. *Journal of Archaeological Science*, 35, 2953–2959. doi.org/10.1016/j.jas.2008.06.016

Cucchi, T., Barnett, R., Martínková, N., Renaud, S., Renvoisé, E., Evin, A., … Dobney, K.M., (2014). The changing pace of insular life: 5000 years of microevolution in the orkney vole (*Microtus arvalis orcadensis*). *Evolution* (N. Y). 68, 2804–2820. doi.org/10.1111/evo.12476

Dryden, I., (2016). shapes: Statistical Shape Analysis. R package version 1.1-13. https://CRAN.R-project.org/package=shapes

Evin, A., Cucchi, T., Cardini, A., Strand Vidarsdottir, U., Larson, G., Dobney, K., (2013). The long and winding road: identifying pig domestication through molar size and shape. *Journal of Archaeology Science*, 40, 735–743. doi.org/10.1016/j.jas.2012.08.005

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179-188.

Frantz, L.A.F., Rudzinski, A., Nugraha, A.M.S., Evin, A., Burton, J., Hulme-Beaman, A., … Larson, G., (2018). Synchronous diversification of Sulawesi's iconic artiodactyls driven by recent geological events. *Proceedings of the Royal Society B Biological Sciences*, 285(1876). doi.org/10.1098/rspb.2017.2566

García del Rey, E., Cresswell, W., 2005. Density estimates, microhabitat selection and foraging behaviour of the endemic Blue Chaffinch Fringilla teydea teydea on Tenerife (Canary Islands). *Ardeola* 52, 305–317.

Gargan, L.M., Cornette, R., Yearsley, J.M., Montgomery, W.I., Paupério, J., Alves, … McDevitt,

A.D., (2016). Molecular and morphological insights into the origin of the invasive greater white-toothed shrew (*Crocidura russula*) in Ireland. *Biological Invasions* 18, 857–871. doi.org/10.1007/s10530-016-1056-y

Guillaud, E., Cornette, R., & Béarez, P. (2016). Is vertebral form a valid species-specific indicator for salmonids? The discrimination rate of trout and Atlantic salmon from archaeological to modern times. *Journal of Archaeological Science*, *65*, 84-92.

Hooten, M.B., Wikle, C.K., (2008). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15, 59–70. doi.org/10.1007/s10651-007-0040-1

Hulme-Beaman, A., Rudzinski, A., Cooper, J.E.J, Lachlan, R.F., Dobney, K., Thomas, M.G. (2020) GeoOrigins Beta. *Zenodo* DOI: 10.5281/zenodo.3919325

Hulme-Beaman, A., (2020) KnnDist: Knn classification with distance inputs. https://github.com/ArdernHB/KnnDist

Hulme-Beaman, A., Cucchi, T., Evin, A., Searle, J.B., Dobney, K., (2018). Exploring *Rattus praetor* (Rodentia, Muridae) as a possible species complex using geometric morphometrics on dental morphology. *Mammalian Biology*, 92, 62–67. doi.org/10.1016/j.mambio.2018.04.002

Hunt, H. V, Rudzinski, A., Jiang, H., Wang, R., Thomas, M.G., Jones, M.K., (2018). Genetic evidence for a western Chinese origin of broomcorn millet (Panicum miliaceum). *The Holocene*, 28(12), 1968–1978. doi.org/10.1177/0959683618798116

Jones, E.P., Eager, H.M., Gabriel, S.I., Jóhannesdóttir, F., Searle, J.B., (2013). Genetic tracking of mice and other bioproxies to infer human history. *Trends in Genetics*, 29, 298–308. doi.org/10.1016/j.tig.2012.11.011

Lachlan, R. F., & Slater, P. J. B. (2003). Song learning by chaffinches: how accurate, and from where? *Animal Behaviour*, *65*(5), 957-969.

Lachlan, R.F., Verzijden, M.N., Bernard, C.S., Jonker, P.P., Koese, B., Jaarsma, … Ten Cate, C., (2013). The progressive loss of syntactical structure in bird song along an Island colonization chain. *Current Biology* 23, 1896–1901. doi.org/10.1016/j.cub.2013.07.057

Lifjeld, J.T., Anmarkrud, J.A., Calabuig, P., Cooper, J.E.J., Johannessen, L.E., Johnsen, … Garcia-del-rey, E., (2016). Species-level divergences in multiple functional traits between the two endemic subspecies of Blue Chaffinches (*Fringilla teydea*) in Canary Islands. *BMC*

Zoology. 1–19. doi.org/10.1186/s40850-016-0008-4

Martínková, N., Barnett, R., Cucchi, T., Struchen, R., Pascal, M.M., Pascal, M.M., … Searle, J.B., (2013). Divergent evolutionary processes associated with colonization of offshore islands. *Molecular Ecology*, 22, 5205–5220. doi.org/10.1111/mec.12462

Mitteroecker, P., Bookstein, F., (2011). Linear Discrimination, Ordination, and the Visualization of Selection Gradients in Modern Morphometrics. *Evolutionary Biology*, 38, 100–114. doi.org/10.1007/s11692-011-9109-8

Nei, M., (1972). Genetic Distance between Populations. *American Naturalist*, 106, 283–292.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H., (2017). vegan: Community Ecology Package 2.4-5.

Ripley, B.D., (2007). Pattern recognition and neural networks. Cambridge University press.

Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly*, 64(1), 38–40

Shennan, S. J., Crema, E. R., & Kerig, T. (2015). Isolation-by-distance, homophily, and "core" vs."package" cultural evolution models in Neolithic Europe. *Evolution and Human Behavior*, *36*(2), 103-109.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

White, J.P., Clark, G., Bedford, S., (2000). Distribution, present and past, of *Rattus praetor* in the Pacific and its implications. *Pacific Science*, 54, 105–117.

Wright, S., (1943). Isolation by Distance. *Genetics* 28, 114–138.

Data availability

The data used in this paper are available in the package. The package is currently available on https://github.com/ArdernHB/GeoOrigins and will be made available on CRAN shortly.

Figure captions

Figure 1. Cartoon panel of GeoOrigin algorithm. Panel A) First, calculate the trait distances from the known reference specimens (coloured dots) to the unknown specimen (black dot) to create vector ***d***. Panels B–G are examples of populating values of the spatial grid *A*. Panels B, D, and F) Second, calculate the geographic distance (dotted coloured lines) from the centre of each grid cell to each reference specimen (coloured dots corresponding to those in Panel A to create vector ***g***. Panels C, E, G) Third, calculate the correlation *r* between ***d*** and ***g***. This process is carried out for every grid cell to populate the matrix *A*. Panel B depicts a grid cell with poor correlation, as shown in Panel C, and therefore can be assumed to be an unlikely origin for the unknown specimen. Panel D depicts a grid cell with a highly negative correlation, as shown in Panel E, making this location among the least likely to be the origin. Panel F depicts a grid cell with a highly positive correlation, as shown in Panel G, and is among the most likely origins for the unknown specimen.

Figure 2. Spatial identification by distance of *R. praetor*. A) Example output of result — the polygon encompasses the region of correlation values at the 95% *r* threshold. B) Boundary finder output demonstrating identification gradient and boundaries around the 145th meridian. C) Histogram showing the percentage of the species range returned by the provenancing method at a 95% confidence level.

Figure 3. *M. arvalis* results at three different geographic scales: A–C) Combined European and Orkney archipelago range; D–F) Orkney archipelago; G–H) Mainland Europe. The results of the boundary finder method are presented in A & D. Example identification outputs are presented in B, E & G. Histogram showing the percentage of the species range returned by the provenancing method at a 95% confidence level in C, F & H.

Figure 4. *F. teydea* provenancing and boundary finding at different geographic scales. A–C) The results from island wide analyses on the raw dissimilarity data; D–E) The results from the Northern sub-region; F–G) The results from most densely sampled NE area. At the highest resolution (F–G) the results of all the individual cross-validation identifications are superimposed as polygons around the specimens true location (note that almost all points fall within a spatial provenancing polygon). The proportion of reference distribution identified is calculated at each

resolution level separately and as a result the analyses run on subsets of the data is not the proportion of the total species range.