Received: 8 November 2019 | Revised: 3 February 2020 | Accepted: 8 February 2020

DOI: 10.1002/ar.24397

COMMENTARY

The Anatomical Record

WILEY

Modern morphometrics and the study of population differences: Good data behind clever analyses and cool pictures?

Andrea Cardini[1,2]

[1]Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Modena, Italy

[2]School of Anatomy, Physiology and Human Biology, The University of Western Australia, Crawley, Western Australia, Australia

Correspondence
Andrea Cardini, Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi, 103, 41125 Modena, Italy.
Email: alcardini@gmail.com, andrea.cardini@unimore.it

Funding information
Leverhulme Trust

Abstract

The study of phenotypic variation in time and space is central to evolutionary biology. Modern geometric morphometrics is the leading family of methods for the quantitative analysis of biological forms. This set of techniques relies heavily on technological innovation for data acquisition, often in the form of 2D or 3D digital images, and on powerful multivariate statistical tools for their analysis. However, neither the most sophisticated device for computerized imaging nor the best statistical test can produce accurate, robust and reproducible results, if it is not based on really good samples and an appropriate use of the 'measurements' extracted from the data. Using examples mostly from my own work on mammal craniofacial variation and museum specimens, I will show how easy it is to forget these most basic assumptions, while focusing heavily on analytical and visualization methods, and much less on the data that generate potentially powerful analyses and visually appealing diagrams.

KEYWORDS

geographic variation, island population, museum data, sampling error, shape coordinates

1 | MODERN MORPHOMETRICS IN STUDIES OF POPULATION DIFFERENCES

Modern morphometrics, or geometric morphometrics (GMM), has largely become a method for the analysis of 2D or 3D digital images. As a flexible, powerful and often low-cost family of techniques for phenotyping animals, plants and other organisms, it has found applications in a variety of fields. Analyses have become increasingly sophisticated and the visualization of results is now even more appealing thanks to the rapid development and spread of hardware and software for digital imaging. However, this analytical and technological progress has often reduced the attention paid to the data and type of variables behind methodologically impressive studies. This happens in all disciplines, but is particularly evident in those requiring large samples and detailed information on individuals, such as most studies of phenotypic variability among closely related populations, a topic which lies at the heart of morphometrics as a discipline born to quantify small differences in morphology (Blackith & Reyment, 1971). Although for brevity I will often refer to this type of research as microevolutionary, I am also considering analyses at the boundary between micro- and macro-evolution (for instance, comparisons of taxa in superspecies and studies of closely related and morphologically similar species). In this context, I will discuss in plain terms some of the potential common pitfalls of modern morphometrics, with a main focus on how data are acquired and morphological variables used. I declare in advance a likely bias in the selection of examples, that largely reflect my personal experience as a biologist working mostly on mammals from museum collections. This is probably the most studied

group of organisms in morphometrics, but, regardless of this, many of the topics I am going to cover have a general relevance, as they are not specific to mammals and may occur in data other than those from museums.

The article is often autobiographical and informal, with the intent to make the issues I discuss less abstract. Thus, the article opens with a very first example from my own work, and a confession: my first big (potential?) mistake in a taxonomic and biogeographical analysis using GMM. After confessing my own sins, I will dedicate the second main part of the article to a number of common problems with sampling and data acquisition (i.e., the study materials). In the third main section, I will briefly discuss a few peculiarities of GMM shape variables (i.e., the methods), mostly well-known and yet still frequently neglected. Finally, I will recap the main considerations and will go back to the first example (my original sin in this field!) and see if I can get absolved for it.

2 | ACCELERATED MORPHOLOGICAL EVOLUTION ON ISLANDS … OR "ADDAMS' FAMILY EFFECT"?

Twenty years ago, I started my career as an amateur morphometrician. Those were the days when "caliper-based" traditional morphometrics (Marcus, 1990) was turning into its modern, geometric morphometric, reincarnation with a strong emphasis on digital images, computerized analyses and effective ways of measuring and visualizing biological shapes (Adams, Rohlf, & Slice, 2004; Rohlf & Marcus, 1993). Over these two decades and half, landmark-based methods using Procrustes superimposition have emerged as the leading approach among a variety of techniques aimed at accurately capturing the geometry of biological forms (Adams, Rohlf, & Slice, 2013). Procrustes employs Cartesian coordinates of corresponding anatomical landmarks (Figure 1) to obtain size and shape data by standardizing their centroid size, centering all individuals in the landmark configuration centroid and minimizing rotational differences among specimens (Rohlf & Slice, 1990). Shape coordinates are then analyzed using multivariate statistics (principal component analysis—PCA, multivariate regression, partial least squares—PLS, multivariate analysis of variance—MANOVA, etc.), and results visualized using various kinds of shape diagrams (deformation grids, displacement vectors, wireframes linking landmarks, and rendering of 2D or 3D images; Klingenberg, 2013). A number of methods for the analysis of outlines is also part of GMM and, although now less widely employed than techniques based on Procrustes, is still fairly popular despite some potential disadvantages (O'Higgins, 1997).

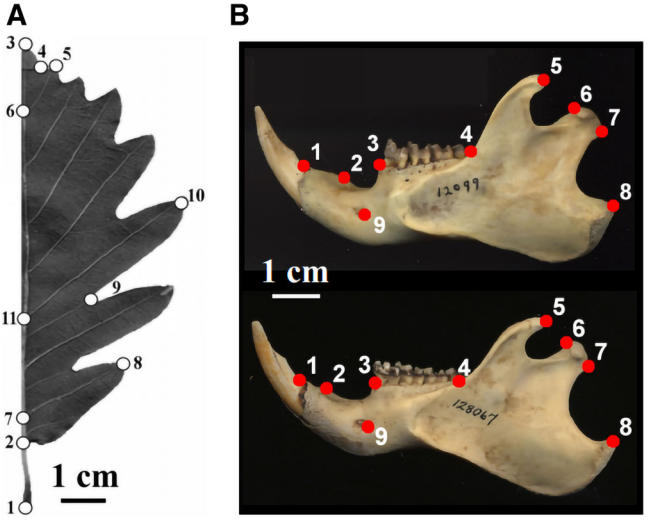

FIGURE 1 Example of anatomical landmarks used to capture size and shape on images: (a) oak leaf (reprinted from Figure 1 of Viscosi & Cardini, 2011 under an open access license: http://journals.plos.org/plosone/s/licenses-and-copyright) and (b) marmot mandibles (*Marmota vancouverensis*, the Vancouver Island marmot, above, and *Marmota caligata*, the hoary marmot, below)

As a self-taught PhD student in biology, with neither strong bases in mathematics nor a propensity for numbers, I struggled learning the principles of this highly quantitative field. Shape analysis is by definition multivariate, as the simplest shape is a triangle, whose vertices are defined by six Cartesian coordinates. Having to perform analyses on a large number of variables introduces a layer of complexity and, of course, multivariate statistics is even harder to learn for a beginner. Thus, most of my focus was on designing my project, acquiring reproducible data and crucially getting the methods right. As I had some experience in morphological cladistics, my original idea was to assess the phylogenetic signal in marmot morphology, but I soon realized that phylogenetic inference using continuous traits, and especially Procrustes shape variables, is hard at best (Adams, Cardini, Monteiro, O'Higgins, & Rohlf, 2011; Rohlf, 1998). The study, thus, progressively shifted towards a comparison of craniofacial similarity relationships in relation to the just published first molecular phylogeny of marmots (Steppan et al., 1999). As things slowly developed, some interesting and somewhat unexpected results started coming out (Cardini, 2003). Mandibular shape was providing a good degree of support for the newly proposed subgeneric division of marmots (Steppan et al., 1999), but there was one main exception. The Vancouver Island

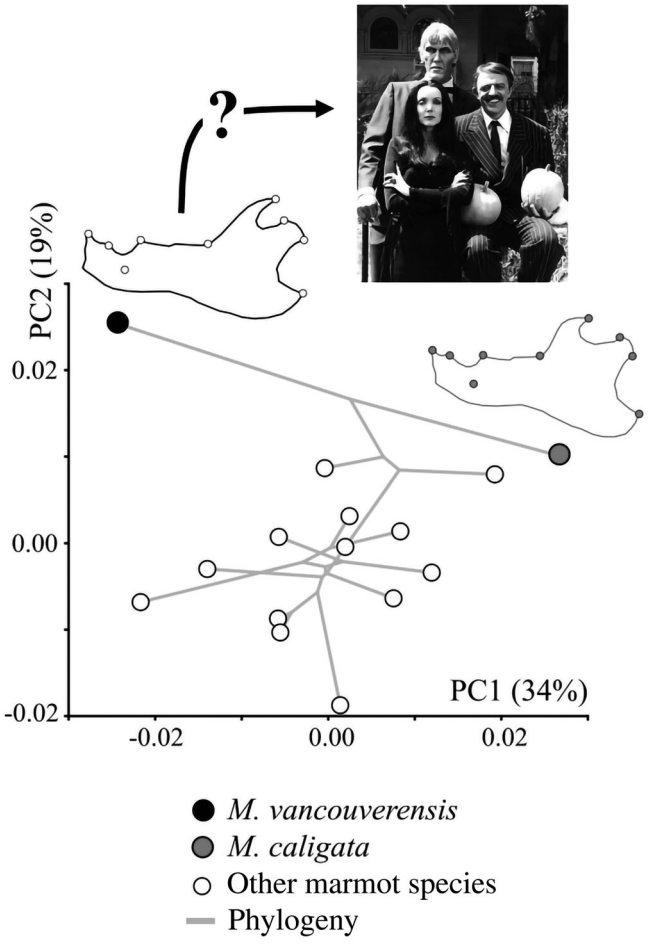

FIGURE 2 First two principal components of marmot mandible variation in species means (in parentheses, the percentages of variance accounted for by each PC). The phylogeny (Steppan et al., 1999) is projected in the Procrustes shape space to emphasize the close relationship of the Vancouver Island and hoary marmots. The shape differences between the two species (magnified ×5 in MorphoJ; Klingenberg, 2011) are visualized using TPS rendering of a mandible outline. The separation and large morphological distance of *Marmota vancouverensis* is very clear. However, is this pattern accurate or an unfortunate by-product of a small sample of closely related and maybe somewhat unusual individuals, as if they are "the Addams Family" of the Vancouver Island population? (Addams Family, NBC Television [Public domain], via Wikimedia Commons at https://upload.wikimedia.org/wikipedia/commons/2/21/Addams_Family_Halloween_1977.JPG)

marmot, a highly endangered, tiny insular population and member of the north American radiation in the Rocky Mountains, had the most distinctive mandible of all 14 (now 15!; Brandler, Lyapunova, Bannikova, & Kramerov, 2010) marmot species (Figure 2), despite DNA differences within or even below the range of within-species variation in all other marmots (Steppan et al., 1999).

My interpretation was that *Marmota vancouverensis* Swarth, 1911, was an example of extremely accelerated morphological evolution in a lineage whose separation from the continent might have occurred little more than 10,000 years ago, at the end of the last glaciation, when the Vancouver Island became separated from the mainland (Cardini, 2003). That would fit well with an established pattern of fast morphological evolution in insular mammals (Millien, 2006), although one that interestingly affected only shape and not size. Indeed, size was fairly similar to that of its sister populations on the continent and thus did not follow the prediction of the "island rule" (Lomolino, 2005). Findings from mandibles (hemi-mandibles, more correctly, as just one side was measured) were later confirmed by the analysis of cranial shape, which also suggested accelerated morphological evolution (Cardini, Thorington Jr, & Polly, 2007). This apparently well supported observation had, we argued, implications for conservation biology as well (Cardini et al., 2009). Regardless of its taxonomic status (a "good" species, a subspecies or just a recently isolated population?), *M. vancouverensis* was a highly distinctive and evolutionary significant unit. Having not had a conspicuously dark brown fur, a unique case for an entire population of marmots (Armitage, 2009), its distinctiveness might have been underestimated by molecular data and its conservation status overlooked. Our main point was that this was a good example of the centrality of integrative taxonomic approaches, combining molecular and phenotypic data to accurately classify populations and provide essential evidence for conservationists (Schlick-Steiner et al., 2007). Importantly (at least for a clearly partisan morphometrician) it showed the usefulness of focusing not only on evident aspects of external morphology, but also subtler differences (Schlick-Steiner et al., 2007) in form, hard to detect without powerful and modern morphometric tools (Cardini et al., 2009).

In fact, we have now begun to appreciate that, even from a purely molecular point of view, the evolution of the Vancouver Island marmot is much more intricate than originally thought (Kerhoulas, Gunderson, & Olson, 2015). Nevertheless, as I gathered and analyzed more data, I kept publishing on marmot morphological evolution with the help of several colleagues. My PhD dataset, based on low cost 2D images of mandibles and crania, was complemented by more accurate analyses of 3D landmarks directly measured on crania (Cardini, Thorington Jr, & Polly, 2007). As I went on working on this subject, I also progressively realized that my whole

story on the Vancouver Island marmot could have been flawed, as I had overlooked an important piece of information. With a living wild population of just 100–200 individuals (Jackson et al., 2015), this species almost inevitably happened to be very rare in museum collections, the source of all my specimens. Indeed, the original sample was less than 10 individuals, that had been kindly shipped to me from Berkeley to Washington DC, where I was doing the main chunk of data collection. That sample, small as it was, was the largest in the USA and the only one I could easily get access to with the limited funds we had. A small sample is never good news, but was there something else that could be really concerning and probably not uncommon in similar studies?

Analyses of small samples are indeed a frequent occurrence in taxonomic and biogeographical applications of GMM and especially typical of studies of large animals from museum collections (Cardini, Seetah, & Barker, 2015). Measuring small anatomical differences in small samples easily leads to very inaccurate estimates of means, variances, and covariances (Cardini et al., 2015; Cardini & Elton, 2007; Polly, 2005). Whether the poor estimates matter or not will depend on the type of analysis and the magnitude of the differences being investigated, but there is no method that can correct for these inaccuracies and, in studies of small geographical and taxonomic differences (the main focus of this article), they almost certainly have a strong impact on the validity and robustness of results. In taxonomy, besides obvious issues with tiny sample sizes, specimens inevitably originate from a few localities of a much larger distribution range. Localities may be clustered and some might be represented by a single individual, while others by multiple specimens. Within each locality, specimens are generally collected at a specific time (year and season), and that may or may not vary across localities. This implies an uneven distribution of data, inadequate spatial resolution and poorly representative samples with non-independent (intrinsically or extrinsically autocorrelated; Beale, Lennon, Yearsley, Brewer, & Elston, 2010) and potentially heterochronic observations. Autocorrelation can in principle be incorporated into statistical models but, especially within species, that may not be so easy (Stone, Nee, & Felsenstein, 2011). In practice, the most commonly employed methods for group comparisons (ANOVA/MANOVA and discriminant analysis, as well as Procrustes ANOVA; Klingenberg & McIntyre, 1998) do not control, at least as they are typically implemented, for autocorrelation and therefore the issue is there virtually all the time. Probably, most morphometricians are aware of these problems, but how often do we really take them into account in modern morphometric analyses of population differences?

Besides naivety and lack of experience, in my marmot studies, I was probably tricked by something that looked rather reassuring in the data from that small Vancouver Island sample: mandibles of virtually all those specimens, showed a characteristic elongated and posteriorly bent coronoid process, and most crania had a V shaped suture between nasals and frontal bones. Previous authors (Hoffmann, Koeppl, & Nadler, 1979) had already spotted this, but, using traditional morphometrics that does not clearly separate size and shape components of form differences, they had not given much weight to just a couple of unusual traits. In fact, these, and some other cranial features, consistently contributed to make the Vancouver Island marmot highly distinctive compared to all other species. They were like a "signature" written in bone of *M. vancouverensis* morphological uniqueness. It did seem to suggest that, regardless of the small sample, these were highly derived and autoapomorphic aspects, consistently found in this species and absent or very uncommon in all others.

Unfortunately, despite taking note of locality and year of collection, I had totally overlooked the importance of when and where those specimens had been collected. Readers familiar with museum collections might have already guessed that those precious specimens were anything but independent observations from the whole species distributional range. Almost all of them came from just two not-so-distant localities, where the animals had been trapped in 1910. Likely, these were members of the same two colonies, probably relatives and therefore strongly autocorrelated observations both in terms of genetics as well as environment. The way I phrased my doubts at the 2008 International Marmot Conference, in Montreux, was by asking whether my original results of high distinctiveness in *M. vancouverensis* were simply explained by what I nicknamed the "Addams family" effect: picking up by chance a sample of a few related animals with very unusual traits (Figure 2), and thus misleadingly interpreting results as evidence of high population distinctiveness.

I will go briefly back to this question later. The fair amount of space I devoted to this very personal example should help to convey the main message of this article, which is about the importance of carefully looking at data that generates the variables we analyze in GMM studies. To strengthen this point, I am going to provide some more examples, that I will again largely borrow from my own research. Sometimes, however, I will also discuss studies by other morphometricians by focusing on very specific aspects, which suggest problems that may or may not have impacted a broader analysis.

3 | MATERIALS: THAT IS, THE DATA GENERATING THE MEASUREMENTS

3.1 | Sample size

GMM taxonomic analyses typically use adults and have a focus on closely related species or geographic populations within a species. In such cases, because differences are generally small and need precise measurements to be quantified and tested, one almost always needs large samples. How large? This is always hard to say and only extensive power analyses might provide accurate estimates of desirable sample size (*N*). However, when I get asked this question by students in introductory lectures on basic statistics, I ask in turn whether they would trust an estimate of, say, female students body height in our university based on a sample of just 10 women, and then confidently use it to compare it with that of another university. They generally agree that this would not be a good idea and larger samples should be used, because differences are likely to be modest, if present, and inter-individual variability within each group may be high. Yet, often populations of a species are represented, within localities, by just a few specimens or even only one individual, as my own study of the Vancouver island marmot exemplifies. The series of papers we published on clinal variation in skulls of African monkeys (Cardini, Dunn, O'Higgins, & Elton, 2013; Cardini & Elton, 2009; Cardini, Filho, Polly, & Elton, 2010; Cardini, Jansson, & Elton, 2007; Dunn, Cardini, & Elton, 2013) employed much larger samples. However, they did so in order to make inferences on geospecies (Grubb, 2006) with a very wide distribution range, typically covering most of Sub-Saharan Africa. Thus, our data, despite measuring hundreds of crania, inevitably had gaps, with many localities being represented by single specimens. A previous seminal work on baboons and their kin (Frost, Marcus, Bookstein, Reddy, & Delson, 2003), despite a large total sample of more than 450 specimens, was, from what one can infer by looking at the maps with the localities of origin of those individuals, similarly affected by an uneven distribution of the observations. Besides, none of these studies could be complemented by a comparison with genetic information, which, especially if originating from the same study specimens, could help to corroborate or refute the patterns suggested by the analysis of phenotypic traits. Thus, as in most morphometric studies in taxonomy but also, for instance, in "evo-devo," ecomorphology or palaeontology, the approach was monodisciplinary and the evidence intrinsically limited by a lack of integrative data from a range of different fields.

Even when focusing only on issues with sampling, the type of research exemplified in the previous paragraph aims at measuring subtle patterns of morphological change in contemporary taxa, whose wild populations may number in tens, if not hundreds, of thousands of individuals (or at least they did it until recent and increasingly rapid declines [Ceballos, Ehrlich, & Dirzo, 2017; Estrada et al., 2017]). Thus, it seems rather implausible to believe that their natural variability in micro-evolutionary studies of geographical differences is adequately represented by samples of a few dozens or, less frequently, few hundreds specimens. Although errors due to measurement and sampling must always be assessed in relation to the specific questions and statistical model, when the questions concerns small differences and crucially depend on accurate estimates of means, variances and covariances, not only sophisticated methods but also high density measurements are no easy fix for very large sampling error. In fact, more variables, as increasingly common in analyses employing semilandmarks, could inflate differences and increase the distortion of between group shape relationships (Bookstein, 2019; Cardini, O'Higgins, & Rohlf, 2019). Again, the problem is neither new nor specific to morphometrics: "Having bucketloads of data only increases the challenges in producing robust and responsible conclusions. A basic humility when building algorithms is crucial" (Spiegelhalter, 2019).

Sampling error in samples of one or just a few dozens of individuals has a strong impact on estimates of means, variance and covariances (Cardini et al., 2015; Cardini & Elton, 2007; Polly, 2005). Luckily, average *N* of 30–40 individuals per group are now typical of many GMM studies (Cardini et al., 2015). Yet, these investigations very often include at least some samples with $N < 10$ (Cardini et al., 2015). In the worst case, a locality or taxon may be represented by just one or two specimens. Statistics based on such a small sample size are not only highly inaccurate, but may also lead to overestimates of evolutionary change, because relatively small inter-individual differences are not "smoothed out" (as it happens in means from larger samples) and might therefore have a disproportionate weight on the analysis. For instance, *Marmota bobak* was the most distinctive species in our 3D cranial shape PCA analysis of marmots (Cardini, Thorington Jr, & Polly, 2007), but was fairly similar to other Euroasiatic species of the subgenus *Marmota* in all previous 2D studies (Cardini, Hoffmann, & Thorington, 2005, and references therein). In fact, this incongruence was simply due to a tiny 3D data sample of just two individuals from the same locality and year of collection, whose average behaved like a strong outlier. Thus, in the end, we excluded this species from the 3D analysis and only included bobak in the larger samples of the 2D studies.

Sample size heterogeneity across localities or taxa further complicates a study design. The problems with

unbalanced samples are clearly mentioned in introductory statistical textbooks (Lane et al., 2017) and not specific to GMM. Yet, they are mostly overlooked in morphometric comparisons of geographical samples and sometimes "fixed" in studies of clines or ecomorphology by simply averaging data within localities (Dunn et al., 2013). Averaging may be a good option to weight equally all data points across a distributional range, but it does not solve the problem of heterogeneous *N* and the inaccuracies of under-sampled localities. Although one cannot easily generalize, the conclusion reached in another paper on sampling error using cranial shape (Cardini & Elton, 2007) may not be atypical in mammals and probably other groups: "in samples of less than 30 specimens, the error in the mean shape estimate can be on average as large as 20–37% of the interspecific distance between mean shapes of *Chlorocebus aethiops* and *Chlorocebus mitis*, two species that diverged about 8 million years ago ... and have profound differences in their ecology and behavior." Sensitivity and power analyses help to better understand these issues, but in many cases one may simply have to acknowledge the very provisional nature of her/his results, and referees should check that this is done, not to penalize those who explicitly state problems to the advantage of others who might overlook the large uncertainties of their findings. Misinterpretations of and overconfidence in *P* values might also lead to misleading conclusions (Smith, 2018), and this could be even more serious using statistically powerful multivariate shape data.

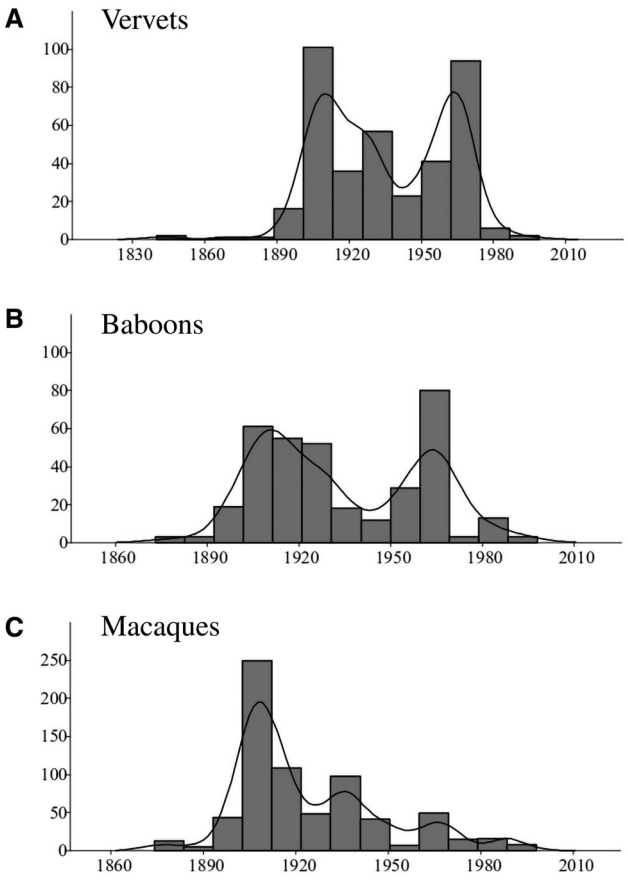

FIGURE 3 Frequency histograms and kernel density plot (computed in PAST 2.17c; Hammer et al., 2001) showing the distribution of the years of collections for the museum specimens in the samples of the primate cranial database (see main text)

3.2 | Nonindependent museum data with patchy distribution, allochronic observations, and scant information?

As anticipated, there are several other problems with sampling, besides *N*: spatial resolution may be poor and data not independent. Hawkins (2012) provides references and a thorough discussion on these topics in biogeographical analyses, and Stone et al. (2011) discusses the problem of within species autocorrelation in the broader context of comparative approaches. We may be all aware of these issues and the potential ways to mitigate their impact, but less often we pay attention to the precise clues that we might get on serious problems with the data by carefully checking where and when specimens were collected. The Vancouver Island study certainly taught me something on this. Morphometric and more generally anatomical and evolutionary studies on mammals, and many other groups, often depend on museum samples (Watanabe, 2019). Working on specimens from collections inevitably implies relying on what is available. Even if one can afford visiting several museums to optimize data collection, for many important and relatively rare groups (the majority, in my experience on mammals), samples will not be huge and one cannot be too picky. The preservation status will also have to be taken into account, which might force users to exclude damaged specimens or, when possible and accurately achievable via a validation, include them by estimating missing landmarks (Arbour & Brown, 2014).

Inevitably, as museum specimens were not collected ad hoc for the specific aims of future researchers, the distribution of the data may be patchy and uneven, and in western museums it might often largely mirror the colonial history of a country. This in fact may not happen only with museum specimens, as time, funds, ethical considerations and problems with accessibility of sites often limits data collection in the wild as well. Nevertheless, it is likely that criteria for the collection of museum specimens have varied over time and especially in the past may have not been up to modern standards: sampling effort seems to often vary across localities, and the

specific sites for collecting specimens may be simply a matter of opportunities and historical contingency. Besides, museum specimens are generally collected over a span of many decades, which adds time to spatial heterogeneity. As an example, Figure 3 shows the distribution of the year of collection for most of the individuals used in some of the biogeographical studies on primate cranial variation we published in a variety of journals (Cardini et al., 2013; Cardini, Jansson, & Elton, 2007; Dunn et al., 2013), as well as an unpublished sample of macaques. About ¾ of these specimens have been collected before the beginning of the 60s and, for macaques, even earlier and mostly before World War II. This skew in the dates of collection towards the first half of the 20th century is at least partly due to changes in ethics and the higher protection status of primates in the last few decades. Regardless of the reasons, if a sample composition is heterochronic, there is the risk of introducing another source of variation, which is hard to control for, unless one decides to limit the analysis to specimens collected within a specific short time span.

In my own work on population variability within species or genera, I typically neglected the time dimension of the data and, as possibly many others, implicitly assumed that changes in morphology and population structure over about a century are negligible on an evolutionary scale. Yet, when for instance we modeled cranial variation in relation to environmental predictors (Cardini, Jansson, & Elton, 2007), our predictors came from a modern database which hardly overlapped with the period of main collection of the specimens in our sample. Considering the fast changes that occurred to habitats and the climate since the beginning of the industrial revolution, and especially after the end of the World War II, we might have been naive in assuming that the mismatch between morphological and environmental datasets was irrelevant. Besides, we only tried to incorporate spatial autocorrelation in our models, without accounting for the potential nonindependence related to the variability in dates of collection of the different specimens. Most ecomorphological and geographical studies of the covariation between morphology and the environment, based on museum samples and especially in large mammals, likely share similar problems, which are not easy to address and should be acknowledged as potential sources of inaccuracy in the results. If, at least for a few taxa, large samples are available, one might perform randomization experiments and rarefaction analyses to explore the robustness of findings in relation to sampling (Cardini et al., 2013), but the outcome will be to some extent biased by the composition of the original sample and certainly cannot solve problems with distributional gaps or the possible mismatch between morphological data and environmental covariates.

Museum data (and not only museum data, in fact!) also contain other potential types of uncertainties. Few morphometricians are trained taxonomists and real specialists of a given group. Thus, most of us rely on museum identifications for taxonomy. A fairly simple expedient to somewhat reduce taxonomic errors, at least for allopatric or parapatric populations, is to double check the classification by plotting the data against the known distribution range of that group. This is a time-consuming task and, as in other cases where precise localities are crucial, a potentially challenging endeavor. Information on localities can be absent, scant or approximate. Knowing that a specimen is from South Africa, for instance, helps a little, but does not allow precise geographical referencing, and old geographical names require some serious investigative work to find the modern correspondent. In this respect, the development of electronic catalogues with detailed information and localities, updated by collection managers and curators, is certainly of very great aid. Nevertheless, many collections still miss complete electronic catalogues and, even when a catalogue is available, taxonomy might change, and big museums cannot constantly keep all their databases updated.

3.3 | Adults only and issues with sex

There are further layers of complexity behind data commonly used in GMM analyses of anatomical variation across populations and species. Most of these studies use adults. There are good reasons for this (including the fact that the majority of museum specimens are adults), but this bias inevitably leaves out large parts of a more interesting life history. Also, focusing on a specific ontogenetic stage is important to avoid mixing up factors (say, age, and geography), but it clearly requires accurate information on age, which is typically missing in museum collections and has, thus, to be estimated. Estimating ontogenetic stages, of course, involves more uncertainty and potential errors, with rare exceptions such as, for instance, in holometabolous insects.

Besides assessing age, as most animals, including all mammals, have separate sexes, one needs to decide whether to pool females and males or perform "split-sex" analyses. This second option has the advantage of controlling by design for another source of variability before comparing groups. However, if *N* is not huge, it also has the undesirable consequence of further reducing sample size. Thus, how to decide what is best? Morphometricians often first test sex differences (Rohlf, Loy, & Corti, 1996; Cardini, 2003) and then, based on results of the

preliminary tests, choose if pooling or not. However, these tests may be less straightforward to implement and interpret than thought. Some groups may have too few specimens of known sex to allow any meaningful test. Others may be bigger, but still too small for adequate statistical power. Rarely samples will be fully balanced between sexes and across geographical samples (be it populations within a species or different species), and heterogeneous *N* makes statistical testing more complicated (e.g., the choice of the type of sum of squares may impact results; unbalanced discriminant analyses are less easy to interpret and require a decision on whether to use weighted or unweighted methods, etc.). Finally, the pattern of sexual dimorphism may vary in magnitude (i.e., absolute difference in, for instance, size) and direction (which sex is bigger?) across groups, making it harder or impossible to control for the effect of sex on comparisons.

Indeed, with multiple taxa, one could test sex differences one group at a time (correcting for multiple testing and inflated type I errors—e.g., Cardini & Elton, 2008) or use more powerful sex by taxon (M)ANOVAs (Cardini, 2003; Frost et al., 2003; Rohlf et al., 1996). All these tests, however, make assumptions (homogeneity of variance–covariance, for instance, besides independence of observations), but these are rarely assessed and often hard or impossible to verify in small samples. Even when assumptions are tested (by the way, remember that tests of assumptions may have their own assumptions!), results of analyses of sexual dimorphism can be easily misinterpreted. For instance, a lack of significance may be indicative of a truly small effect of sex (and, for the interaction between sex and taxon, of similarity of patterns of sex differences across samples) or it could be just an issue with low statistical power. Considering estimates of effect size in these tests (such as R^2 or its multivariate extension) is certainly advisable, but these estimates are themselves affected by sampling.

One might also simply decide whether to pool sexes based on what is already known from the literature (Smith, 2018). However, with the exception of cases where sexual dimorphism is obvious to the point of making tests unnecessary, this could produce unreliable conclusions for the specific structure under study. This is because sex differences may vary in magnitude not only across taxa but also, within a species, among anatomical regions: for instance, in mammals, the pelvis is almost always likely to be highly dimorphic, while other structures might be fairly similar in females and males. Surprisingly, studies are sometimes performed regardless of sex even when there is evident sexual dimorphism and its degree varies across taxa. This is generally justified by arguing that, at least in macroevolutionary analyses, sex differences are small compared to variation across taxa, but this may or may not be true. Alternatively, the problem of potential sex differences is not discussed and analyses are done regardless of that. The impact of ignoring sex differences varies from case to case, but overlooking it makes results harder to interpret and potentially flawed. Using an extreme example, consider the average size or shape of a female and male baboon cranium, a numerical abstraction that likely has no correspondent in nature. Males have very long prognathic snouts and are about twice the size of the short-faced females (Figure 4). Depending on whether samples are more or less balanced, and depending on whether the average is weighted or not, the mean baboon (within locality, region or population) could be a point halfway an almost perfectly separated bimodal distribution or something which is, say, 1/3 female and 2/3 male, if *N* for females is half than *N* for males.

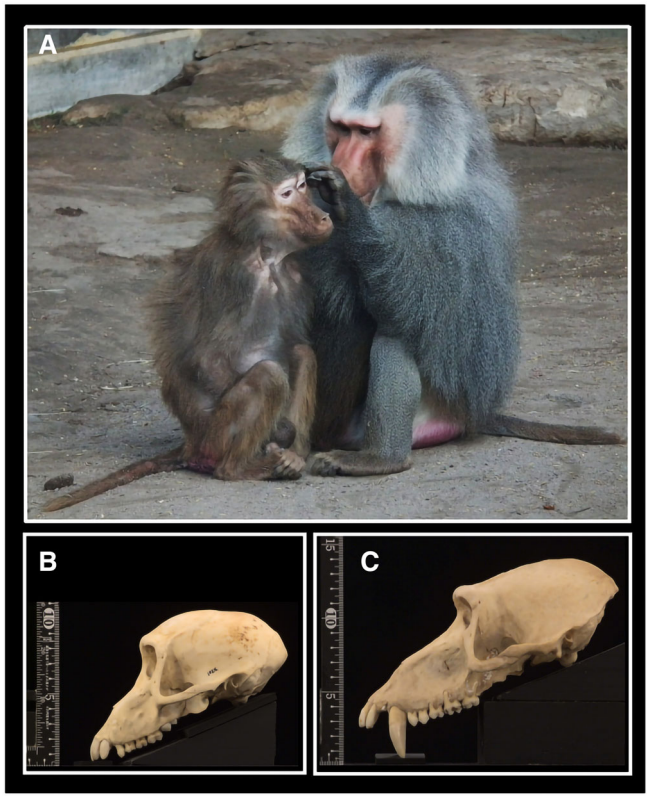

FIGURE 4 Female and male hamadryas baboons: (a) female to the left and the much bigger male to the right (from https://commons.wikimedia.org/wiki/File:Ist_was_los_fragt_sich_der_Mantelpavian_Tiergarten_Worms_2011.JPG) and (b,c) female and male crania in side view (respectively from http://1kai.dokkyomed.ac.jp/mammal/images/large/papio/DKY_1024L.jpg and http://1kai.dokkyomed.ac.jp/mammal/images/large/papio/PRI_6441L.jpg)

How to control for sex differences in dimorphic structures is not easily solved by "sex-corrections" either (Frost et al., 2003; Cardini et al., 2013), as briefly anticipated.

These methods try to account for sex variation by statistically removing it, so that females and males can be pooled and *N* increased. However, the "corrections" make (again) a number of assumptions: estimates of sample means must be accurate; the variance–covariance structure is similar in females and males; and the pattern of sex differences should be the same across populations, something generally tested by the interaction term of an ANOVA model. None of these assumption is easily tested, unless one has very large samples, but then probably the researcher does not need to pool sexes, if dimorphic, and can simply analyze females and males in parallel, with the advantage of being able to compare results between them.

3.4 | Flattening 3D objects in 2D

Finally, to conclude this section on data acquisition, there is at least one other important but largely neglected problem of many GMM studies, including the large majority of my own papers on marmots: three-dimensional structures such as skulls, long bones, most plant seeds and most arthropod body parts etc. are very often analyzed using 2D approximations. This simply means taking *flat pictures* of a specific view of a structure and digitizing 2D landmarks on the photos. This approach dominated the early years of GMM, when tools for 3D landmarking were uncommon and expensive, but it is still very popular today (Cardini, 2014; Cardini & Chiappelli, 2020), because of its simplicity, low costs and the speed of data collection.

However, in doing approximations of 3D anatomies using flat pictures, the loss of information is two-fold. First, the information is strictly dependent on the choice of the side being photographed. Taking pictures of hemi-mandibles on the labial side, as in the Vancouver Island marmot study (Cardini, 2003), the lingual side as well as the anatomical relationship with the other hemi-mandible, are left out of the analysis. Similarly, focusing on the midplane outline of the corpus callosum (Bookstein, Sampson, Connor, & Streissguth, 2002) in lateral view may accurately capture that bidimensional contour, but misses all the other aspects of this 3D nerve tract of the placental brain.

The second type of loss of information in 2D pictures is due to the fact that the flattening of the third dimension not only discards some information, but also potentially introduces a distortion in the relative positions of anatomical points (Roth, 1993). This happens all the times one employs landmarks that are not perfectly coplanar and obviously occurs regardless of whether photos have been carefully taken to avoid perspective distortions, as a perfect 2D photo is still flat! For instance, landmarks on the zygomatic arch of a mammal cranium in side view are on a different plane compared to landmarks on teeth or the midplane (Cardini, 2014). Less obviously, but similarly, landmarks on the various cusps of a cave bear molar photographed from above are at slightly different heights (Seetah, Cardini, & Miracle, 2012). Distances between noncoplanar points are underestimated and shapes inevitably distorted to some degree. The larger the variation in the height of the planes where different landmarks are placed, and the smaller the variation in a study, the stronger the error introduced by the 2D flattening (Cardini, 2014; Cardini & Chiappelli, 2020).

The 2D to 3D approximation is a remarkably obvious type of measurement error. Yet, despite being clearly mentioned in the old GMM literature (Arnqvist & Martensson, 1998; Roth, 1993), the problem is almost constantly ignored. Paradoxically, other and likely much smaller sources of errors, such as landmark digitizing error, are often carefully tested. As with other problems I discussed in this article, this lack of attention to potentially inaccurate 2D measurements is even more problematic in intraspecific studies of small differences. 2D analyses are not necessarily flawed, as 2D data might still produce results congruent with 3D ones (Cardini & Chiappelli, 2020), but seriously omit to consider an important issue, that should at least be mentioned and maybe mitigated by a careful selection of landmarks. In fact, testing the goodness of the 2D to 3D approximation is relatively easy, as exemplified and discussed by Cardini and Chiappelli (2020). More generally, this, as well as other components of measurement error, should be carefully considered and assessed in relation to the study question (Arnqvist & Martensson, 1998; Fruciano, 2016).

4 | METHODS: MULTIVARIATE SHAPE VARIABLES BEHIND APPLICATIONS

Most of the discussion up to this point was on data acquisition and its relevance in the context of studies of geographic variation at microevolutionary level or at the boundary between micro- and macro-evolution. In this section, I consider the type of shape variables used in Procrustes GMM, the most widely employed family of GMM techniques in biology (Adams et al., 2004).

GMM studies can be subdivided (Rohlf, 1990) in three main steps, after data acquisition: (a) feature extraction, which consists in obtaining the morphometric descriptors (e.g., centroid size and Procrustes shape coordinates); (b) statistical analysis of the variables obtained in the previous step (e.g., regressions on geographic coordinates or environmental predictors, comparisons of populations using ANOVAs etc.); (c) visualization of results from the statistical

analyses. I am here focusing mainly on the first step, feature extraction using Procrustes methods. I might briefly touch on a few other aspects, such as specific statistical analyses or visualization methods, but again this will be in relation to the implications of the type of shape variables of Procrustes GMM. Thus, as in the first two main sections, the focus is still on GMM data, but the attention shifts from the material commonly used in taxonomic studies and comparisons of populations and taxa to the peculiarities of the Procrustes shape coordinates and their main linear combinations, which are principal components (PCs) and partial warps (PWs—http://life.bio.sunysb.edu/morph/glossary/gloss2.html). PCs and PWs use different criteria to rigidly rotate the axes on which observations are projected. Because these are rigid rotations that do not alter shape distances in a sample, if all PCs or all PWs are analyzed, they convey is exactly the same information as the full matrix of Procrustes shape coordinates. Informally, when I refer to any of these types of variables, I might therefore use the general term "Procrustes shape variables."

4.1 | Too few specimens in highly multivariate morphospaces?

Shape data are, as mentioned, multivariate by definition. This does not only mean that one has to perform all analyses using multivariate methods, but also that unfavorable ratios between sample size (N) and the number of variables (p) are not unlikely to occur (Cardini et al., 2019). For instance, my first study of the Vancouver Island marmot (Cardini, 2003) was based on a configuration of nine 2D anatomical landmarks and a total of almost 400 adults. Thus, the overall N/p ratios was ca. 28, but within species it ranged from 0.2 to 4.4 (average 2.0). Our later intraspecific study on clines in vervet monkeys (Cardini, Jansson, & Elton, 2007) employed a set of 86 3D landmarks and an overall sample of more than 300 adult specimens, with N/p ratios of 1.2 in total and 0.5–0.7 within sex. This means that in both studies within group N was only slightly larger than, or about the same as, p, and in some cases it was in fact much smaller than p. Statistical textbooks and introductory papers often suggest to have many more individuals than variables in an analysis. Sometimes they offer rules of thumbs, such as N 5–10 times larger than p (Hair, Black, Babin, Anderson, & Tatham, 1998; Howell, 2012; Strauss, 2010). However, there cannot be a rule of thumb for N/p that fits all situations, and the desirable ratio will be clearly context dependent (Howell, 2012; Marcoulides & Saunders, 2006; Strauss, 2010). Also, because of covariance among shape variables, the effective p might be less than apparent, as discussed in the context of between group PCA by Cardini et al. (2019). Nevertheless, as the number of variables in a study increases, while samples remain relatively small, one moves further away from a desirable situation and, even if the method may allow to perform the analysis, the probability of running into problems becomes progressively higher.

Especially with the use of semilandmarks to "discretize" curves or surfaces lacking clearly corresponding anatomical landmarks (Gunz & Mitteroecker, 2013), it is not unusual to have datasets with more than a thousand Procrustes shape variables (e.g., p = 2,805; Neubauer et al., 2018). Common solutions to deal with a large number of variables and relatively small samples are: excluding smallest samples; dimensionality reduction; the use of distance-based resampling statistics in the full shape data space. None of these remedies is perfect and it has been shown that unfavorable N/p ratios might produce spurious patterns in simulated data that only contain random isotropic noise (Bookstein, 2017, 2019; Cardini et al., 2019). Thus, for instance, a PCA might suggest dominant PCs and elliptical scatter on PC1-PC2 (e.g., fig. 4c of Cardini, 2019, but see also Björklund, 2019), when scatterplots should in fact be circular and variance equal across all PCs. With real data, such as covarying shape coordinates, this issue will be less pronounced (Bookstein, 2019; Cardini et al., 2019). However, it might still affect analyses to an extent hard to predict, as the true covariance structure of the data is unknown and simulations of covariance are inevitably specific to the type of variation being simulated, as well as the specific study structure and the landmark/semilandmark configuration.

Despite these problems, and the fact that studies of wild populations are rarely based on huge samples, analyses of large number of landmarks and semilandmarks have become more common with the development of user-friendly software (e.g., the TPS Series for 2D landmarks and semilandmarks; Rohlf, 2015), various GMM R packages (Adams & Otárola-Castillo, 2013; Schlager, 2017, etc.). This, together with cheaper instruments for obtaining 3D images, has led to an increase in data dimensionality, as more points are digitized on curves and surfaces. Although there is disagreement in the morphometric community over the *pros* and *cons* of using large numbers of points (for recent discussions, search at https://www.mail-archive.com/morphmet@morphometrics.org/: semilandmarks AND biology), there are certainly many examples of fruitful use of semilandmark methods and cases where their application is crucial to answer a specific question. For instance, a fragment of fossil cranial vault found in the sea off the coast of the Netherlands was identified as a likely Neanderthal by capturing its shape with a dense set of semilandmarks, as hardly any true landmark was available on this small piece of bone (Hublin et al., 2009).

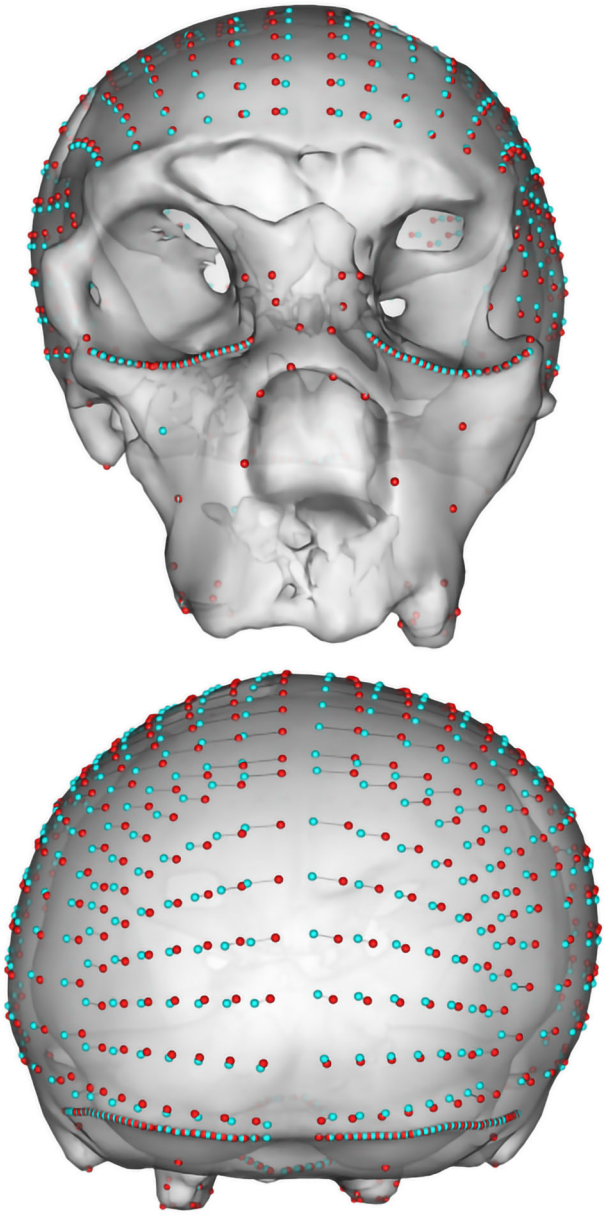

FIGURE 5 Example of application (retrodeformation of a fossil) of semilandmarks on 3D surfaces (reprinted from Figure 4 of Schlager, Profico, Vincenzo, & Manzi, 2018, under an open access license: http://journals.plos.org/plosone/s/licenses-and-copyright)

Semilandmarks may also be crucial in other instances, such as the computerized reconstruction of fossils (Figure 5; Gunz, Mitteroecker, Neubauer, Weber, & Bookstein, 2009; Schlager et al., 2018). However, there are also instances where using more but less precise points leads to larger measurement error (Fruciano et al., 2017) and potentially aggravates the risk of spurious results, because of the altered pattern of covariance introduced by the superimposition (see below) as well as, if present, the mathematical treatment of semilandmarks (Cardini, 2019).

The temptation of always include more points to apparently better capture shape and improve the visualization is strong, and it is easy to be tricked by a misleading belief that more always means better. Semilandmarks may be of real help, and will always produce nicer shape diagrams, but I know of no single demonstration that more points equates to increased accuracy (i.e., results closer to the truth), except if assessed using a tautological and misleading assumption (Watanabe, 2018) that the configuration with the largest number of landmarks/semilandmarks represents the true morphology, regardless of the biological meaning of what is measured and its relationship to the specific study questions. This assumption easily leads, for instance, to the paradox of arbitrarily ignoring anatomy, with corresponding semilandmarks potentially ending up on different bones in different individuals or taxa (Cardini, 2013). Yet, the aim of GMM in biology is not beautiful pictures or digital art, and parsimony is a crucial aspect of the study design.

Indeed, the most important issue to address, before any data collection, is what type of information one needs in relation to the specific study aim: this is not a simple question and will always imply a degree of arbitrariness, but it is a crucial one. On this topic, as well as the problem of "homology" in GMM, I refer the readers to the commentary of Klingenberg (2008) and the careful review of (Oxnard & O'Higgins, 2009). In general, on the same structure, depending on the scientific aim, there can be several different and equally valid configurations of anatomical points: sometimes one will need more measurements, and sometimes less; and the specific measurements will vary with the study question. Practical considerations on sample size and the type of analysis may also be important, so that features of interest are adequately captured, but the impact of unfavorable N/p ratios is reduced.

4.2 | Other "peculiarities" of Procrustes shape data? Examples of common issues in analyses and visualization

Another aspect of the analysis of Procrustes shape data, that must be considered before the statistical analysis and the visual interpretation of shape diagrams, relates to the way these variables are obtained, which make them "peculiar" compared to other types of anatomical measurements. The rescaling and superimposition of landmarks (and, if present, semilandmarks) to remove irrelevant positional differences, and separate size and shape, introduces a degree of covariance in the coordinates of the anatomical points (O'Higgins, 2000, and references therein). This means that, even if one simulates

perfectly circular uncorrelated (isotropic) variation, shape coordinates after the superimposition will covary (Rohlf & Slice, 1990). Procrustes is the preferred option to superimpose landmarks, because of its desirable statistical properties (Adams et al., 2004). However, like other types of superimposition based on pure mathematics and no biological model, Procrustes is biologically arbitrary. As long as the variables generated by this method are used all together in multivariate analyses, Procrustes GMM performs well: at least in simulations using simple configurations of landmarks, it is clearly superior for power and accuracy to the main alternative approaches (Rohlf, 2000a, 2000b, 2003). Nevertheless, the biological arbitrariness of the superimposition means that univariate analyses of Procrustes shape variables cannot be meaningfully performed using any of these variables one at a time (Adams et al., 2011; Rohlf, 1998). Thus, for instance, in multivariate analyses, coefficients such as PC loadings or regression coefficients cannot be used with Procrustes shape data, and variation cannot be interpreted at a specific landmark location, as this is also a function of the choice of superimposition (fig. 9 of Viscosi & Cardini, 2011, and references therein).

Despite being known since the early days of GMM (Rohlf, 1998), it is not uncommon to find studies violating these basic assumptions. On these issues, I will briefly cite a few papers by other authors using GMM in population studies. As anticipated in the introductory section, I am going to focus on examples of misuse of shape data without drawing conclusions on whether that might invalidate an entire analysis. In fact, most of the time, the problem affects just specific aspects of a much broader study and their interpretation.

- Use of partial warps one at a time. In a study comparing populations of the South African-endemic rodent species *Otomys saundersiae*, Procrustes GMM was employed along with other approaches for integrative taxonomy (Taylor, Kumirai, & Contrafatto, 2007). Shape variation in dorsal crania was explored both in relation to size (i.e., allometric shape) and geography. For allometry, although the main regression model was multivariate, and thus correctly employed all shape variables, scatterplots of univariate regressions of selected PWs onto cranial size were shown, together with a table with univariate regressions of each PW onto size. Neither the scatterplots nor the univariate regressions should have been there for the reasons explained in detail by Rohlf (1998). Briefly, the criterion used to rotate Procrustes shape data in order to compute PWs is minimum bending energy (http://life.bio.sunysb.edu/morph/glossary/gloss1.html), which is a quantity used to compute smooth deformation grids using a function (thin plate spline or TPS; Bookstein, 1989) borrowed from studies on the physical properties of thin metal sheets. This method is certainly good for "surface interpolation for computer graphics and computer-aided design" (http://life.bio.sunysb.edu/morph/glossary/gloss2.html), but it is clearly not based on biological insight. Thus, the selection of specific PWs, and any analyses using subsets of these variables, is unlikely to ever make sense in biological and anatomical studies.
- Use of one PC at a time and/or stepwise selection of PCs. Bearing in mind the caveats about highly dimensional shape spaces in relation to *N* and the effect of the superimposition, PCs can be used to summarize shape, because they maximize sample variance (which is a biologically interesting quantity, unlike PWs' bending energy). Besides producing useful scatterplots, by carving out an appropriate subspace of the total Procrustes shape space, a PCA could serve for dimensionality reduction. For instance, one could use all PCs from PC1 to the i-th PC, with the i-th PC selected so that shape distances in the reduced space are almost the same as in the full Procrustes shape space. However, PCs too must be used with some caution. To start, the subspace of the first i-th PCs could be a very good summary of the total shape space and yet miss important information within groups. Also, because the rotation of PCs is sample-dependent, a small change in sample composition could significantly alter the orientation of the PCs, with potentially strong impact on any analysis performed one PC at a time (Adams et al., 2011). In general, as shape is by definition multivariate, the directions of most interesting variation (e.g., in relation to sex, population, geography, environmental variables etc.) may not align well with any PCs and therefore it is better to avoid univariate analyses of PCs (Barčiová, Šumbera, & Burda, 2009), as well as stepwise selection of PCs in discriminant analyses (Dapporto, 2008). This does not mean that a specific PC never captures important and interpretable aspects of shape variation, but, if one is looking for the direction of covariation with other factors (age, sex or taxonomic groups; environmental or genetic covariates, etc.), other methods will likely produce more accurate summaries of shape differences in relation to those factors. For instance, in ontogenetic studies within a species, PC1 often covaries strongly with size and is thus considered the main allometric axis. This may be true but a better approximation is almost certainly obtained by using a multivariate regression of all shape variables onto size (with different proxies for size depending on the context—Hallgrímsson et al., 2019), as this will capture the direction of largest covariance with size and may potentially incorporate curvilinearities in the

allometric trajectory (Mitteroecker, Gunz, & Bookstein, 2005).

- "Recycling" a single dominant direction of variation in other analyses. This is an issue that may involve a variety of methods, where first a vector capturing specific aspects of shape variation is computed and then its scores are reused in another analysis. For instance, when PC1 dominates the pattern of variance in the data, it is tempting to use only this axis and perform simple univariate analyses. As an example, one could plot and explore the covariation of shape using PC1 in relation to latitude (Milne & O'Higgins, 2002). However, PC1 might have been stretched by a few distinctive observations, and thus poorly represent variation among more similar individuals or taxa in the sample. Even more importantly, as already mentioned, PCs are optimized to summarize variance regardless of any other factor and, therefore, are not the best axes to capture covariation with other variables. To this aim, alternative methods may be better, such as multivariate regressions (Cardini, Jansson, & Elton, 2007), PLS (Monteiro, Duarte, & Reis, 2003), and so on. For similar reasons, first extracting an axis that best discriminates groups, such as that of the first discriminant or canonical variate function, and then using it in further analyses in relation to other factors (e.g., latitude, as in Renaud & Michaux, 2003) may not be an optimal choice. It might be better to analyze the multivariate group means in relation to the predictor or to build a model where the covariation of shape with latitude, or other factors, is assessed by simultaneously taking into account group structure. In the specific case of discriminant/canonical variate analysis, a most common tool in studies of biological populations, one should also consider that this type of analysis modifies the similarity relationships in the shape space in a way that tends to overfit the data (Kovarovic, Aiello, Cardini, & Lockwood, 2011). This problem becomes more pronounced using small samples in highly dimensional data spaces (Kovarovic et al., 2011) and is potentially aggravated by strongly unbalanced samples and heteroscedasticity (Albrecht, 1992). In fact, as recently shown (Bookstein, 2019; Cardini et al., 2019), even the main alternative method used to summarize group differences in Procrustes shape data, a between group PCA, is plagued by a potential inflation of group separation when p is very large relative to N. With between group PCA, the problem is potentially aggravated by the lack of computational constraints (unlike in discriminant/canonical variate analysis, where N must be larger than $p - g$, where g is the number of groups), which allows its application even when $p >> N$.
- Allometry and "size-correction." Allometry concerns the covariation of shape with size (Klingenberg, 2016; Hallgrímsson et al., 2019). It is often of interest in itself, as when one estimates the allometric trajectory in a sample using, for instance, a multivariate regression of all shape variables on the centroid size of the landmark configuration (in fact, although centroid size is almost the default option in Procrustean GMM, the choice of the specific size covariate for testing allometry is not trivial and must relate to the specific study questions; Hallgrímsson et al., 2019). However, allometry is also frequently seen as a confounding factor to control for by regressing out size-related shape in order to obtain "size-corrected" residuals for further analyses. When allometric shape is of interest, two common problems are: (a) estimating the allometric trajectories regardless of potential group structure; (b) reusing the allometric vector in other analyses. This second issue (e.g., using allometric regression scores (Maestri et al., 2016) or using PC1 of the form space (Perez, Diniz-Filho, Bernal, & Gonzalez, 2010) in further analyses) is another case of suboptimally recycling a dominant univariate direction of shape change. As I pointed out in the previous two paragraphs, this is likely to lead, at best, to a poor approximation, that misses out important variation in the multivariate shape in relation to factors others than the single one used to obtain that main axis. The first issue (a), which is not specific to shape data, is well known in the context of the analysis of covariance (Howell, 2012). Before controlling for the effect of a covariate such as size (e.g., using "size-corrected" regression residuals; Klingenberg, 2016), one should first test if all groups share a common allometric pattern. This requires demonstrating that the slopes of the group-specific allometric vectors are similar. Only in this case, one can confidently apply the same model to all groups, and across their entire range of sizes, to control for the covariate (Sheets & Zelditch, 2013). Testing for a common slope of shape trajectories can be done with a multivariate analysis of covariance including the interaction between groups and the covariate or by testing for significant differences in vector angles between groups (Sheets & Zelditch, 2013). Resampling methods are available that allow to perform tests even with tiny samples. However, especially when within-group size variation is modest, as typical of studies of adult mammals and birds, accurate estimates of allometric slopes will likely be very inaccurate (Cardini & Elton, 2007) and P values potentially misleading if not interpreted with caution (e.g., suggesting nonsignificance and thus parallel allometries simply because of low power in very small samples or suggesting a significant interaction

because of just one or a few samples having divergent trajectories).

- "Moving" landmarks. As anticipated, because coefficients cannot be interpreted using Procrustes shape coordinates, as one would do in traditional morphometrics, results of shape analyses must be visualized using shape diagrams (Klingenberg, 2013). Figures 2 and 5 provide examples of respectively 2D and 3D rendering of outlines and surfaces as a tool to show shape differences. With some of the shape diagrams, however, it is easy to be mislead, as it often happens when a starting shape is morphed into a target, with the two shapes shown one superimposed on top of the other. This is equivalent to using displacement vectors to illustrate the position of each landmarks in the target relative to the starting shape (fig. 9 of Viscosi & Cardini, 2011). It is an approach that easily gives the impression that one or the other landmark is moving forward, backward, up or down, relative to the same landmark in the other shape (Barčiová et al., 2009). In fact, it is the whole space, whose boundaries are marked by the entire configuration of anatomical points, that changes (widening, narrowing, warping and bending etc.). This risk of misinterpreting shape differences is, nevertheless, easily avoided by either plotting the starting and target shape one next to the other (instead of superimposed) or by using TPS grids.

In conclusion, GMM has some peculiarities in relation to the intrinsically multivariate nature of shape and, for Procrustes methods, to the biological arbitrariness of the superimposition. Problems can range from aspects related to the oversimplification of multivariate shape data (such as restricting analyses to a single dimension of shape variation or neglecting group structures in allometry) to those specific to the type of shape variables (e.g., the misuse of PWs) or the misinterpretation of shape diagrams. Most of these potential issues, however, become nonproblems, as soon as one recognizes and thus avoids them.

5 | BACK TO THE FIRST QUESTION: DID I GET "COOL" RESULTS … BECAUSE OF BAD DATA?

As some of its founding fathers had forecast (Rohlf & Marcus, 1993), GMM, and in particular methods based on Procrustes, has been revolutionary in the field of quantitative morphological analyses. The relative simplicity of the principles behind these techniques, that simply require digitizing corresponding anatomical points to measure differences, have made them highly popular (Adams et al., 2004). Their flexibility has also made GMM an ideal tool for multidisciplinary studies, where it is easily combined with genetic data for evo-devo and phenomic analyses, macroevolutionary studies, integrative taxonomy, phylogeography, biomechanics, etc. This progresses have been accompanied by similar advances in other fields. The "R revolution" (Tippmann, 2015), as well as the development of user-friendly software, has also greatly contributed to the success of new, more accurate and potentially more powerful analytical tools and methodology in the biological sciences. Unfortunately, this might have also led us to pay less attention to the quality of the data we analyze and the assumptions of the methods we use.

Data may be "bad" despite the best efforts of a scientist: samples may be small and heterogeneous even after the most extensive data collection; specimens may cluster locally and be heterochronic across localities; there could be gaps in the distribution, inadequate spatial resolution for the level of the analysis and uneven sampling over a large range; pictures may be taken with the uttermost care and yet large errors be introduced by excessive flattening of the third dimensions in 3D objects; there could taxonomic misidentification; "corrections" for sex differences or allometric variation may be inaccurate; a large number of variables, together with potentially useful information, could bring a lot of noise in the analysis of small samples; dimensionality reduction of multivariate shape data could lead to oversimplification, miss a lot of important details or even be just wrong, as in univariate analyses of PWs.

All these potential issues might give the impression of a pessimistic picture of an otherwise successful field, but I do not see it as such. In terms of samples, a main focus of this commentary, this is most of the time a realistic scenario, which is especially true for microevolutionary analyses dealing with small variation, as in the majority of GMM studies focusing on population differences. It might be that sometimes a study is simply not feasible, because of too many potential sources of inaccuracies in the data themselves. However, most of the time, it is more likely that a scientist will be able to go on with his/her analysis, but might do so with a deeper awareness of the potential issues, and thus avoid obvious pitfalls. Problems will have to be acknowledged and results might turn out to be preliminary, but still important if new or from a poorly studied group. Thus, unless multiple lines of evidence are used and the approach is truly integrative and based on strong biological foundations, GMM will provide useful and potentially very interesting information, but will offer an inevitably partial view of the matter and one which is unlikely to produce conclusive results: hence, for instance, it should not be used on its own to establish new species or subspecies in living taxa, as well as it should not be the

exclusive source of data for assessing modularity and integration in anatomical structures.

I have often heard colleagues saying: this is all I could get and thus I have to live and work with it. It may well be true, but certainly it does not prevent to state clearly the limitations of a study. This is something that should be positively considered by reviewers and editors, and not swept under the carpet by arguing that the paper might look apologetic and be weakened. In this respect, it does seem that there is a tension in the scientific literature between the *pros* and *cons* of different ways of presenting results in a paper. In the same Career Feature in Nature (Gewin, 2018), one contributor suggested to avoid "writing that sounds defensive, with too many caveats ... as if ... to fend off criticism" (p. 129), whereas another said that "authors should avoid being over-confident in their conclusions ... [and by] making clear how robust their findings are, they must convince readers that they've considered alternative explanations" (p. 130). With whom I personally side is as obvious as unimportant. I leave, however, to the readers to decide which view might be more consistent with the preoccupations, expressed by many researchers, of a reproducibility crisis emerging in numerous branches of biology, medicine and other fields of science (Allison, Shiffrin, & Stodden, 2018).

Whereas we could dwell on this much more, the main point is clear: even the most accurate methods, cannot give the right answer, if they are applied to poor quality data or the wrong type of variables. Technological and methodological advances are fundamental for scientific progress, but good data are even more crucial (Spiegelhalter, 2019). This is self-evident, and yet we seem to be going in a different direction. Often, even highly desirable multidisciplinary studies are compartimentalized between specialists of methods, who know or care little about the data they are given, and data collectors, who in turn might overestimate the "magic" that sophisticated numerical analyses might do on their precious samples and measurements. In fact, as data sharing and public depositories become more common (which is definitely good news for science), the gap between methodological wizards and data analysts and the material and measurements they work on is likely to become wider, with potential flaws that will be hard to detect. This trend of increasing emphasis on complex analytical tools probably also explains why it has become increasingly common to have papers reviewed only by experts on methods without anyone looking properly into the biology and data behind a study.

Finally, I have not yet answered the question that opened this paper: did I get it right with my small sample of Vancouver Island marmot hard-tissue data, and this population is really unique, or did I pick up the "Addams family" of marmots (Figure 2)? This question troubled me until a Canadian palaeontologist suggested a follow up of the mandibular analysis on a much larger sample of both modern and subfossil specimens. I readily accepted. Finally I had a chance to find out the (we hope) truth about my original tiny sample of likely nonindependent and simple 2D observations. And, luckily (very luckily!), I seem to have got it right: 2D is a good approximation for three-dimensional but fairly flat hemi-mandible (Cardini, 2014), and the Vancouver Island marmot has indeed a truly distinctive mandibular shape, consistently found, across samples from different periods and localities, for at least the last few thousand years (Nagorsen & Cardini, 2009). The magnitude of the differences, and their apparently fairly ancient origin, fits well with the recent analysis of multiple molecular markers (Kerhoulas et al., 2015) indicating that the evolutionary history of *M. vancouverensis* is longer and more complex than inferred two decades ago using a single mitochondrial gene (Steppan et al., 1999). The Vancouver Island marmot probably represents a reproductively incompletely isolated, but rapidly diverging, monophyletic lineage within a larger continental clade of hoary marmots (Kerhoulas et al., 2015). Despite incomplete isolation, the new genetic data, together with a variety of morphological and behavioral autoapomorphies, makes *M. vancouverensis* even more precious as an evolutionary significant unit and conservation target (Cardini et al., 2009).

Is this the end of story? Certainly not. In fact, even the larger sample we analyzed in 2009 is still fairly small, with gaps and discontinuities across the distribution range and time span we sampled. Yet, the strong congruence of the data makes me feel more optimistic that we are on the right track. It may be hard to largely improve *N* in such a rare island endemism, but I do dream of massively sampling 2D data on marmot mandibles in North America, and maybe the Palaearctic as well, to more accurately explore patterns of phenotypic variation in this genus of large cold-adapted mammals, whose demography and chances of survival are being altered by direct anthropogenic pressures and global warming (Armitage, 2013). I wonder whether this extensive sampling could be achieved by networking with as many museums as possible and by crowd-sourcing data collection. This is indeed one of the potential strengths of GMM: large groups of scientists can easily collect standardized 2D or 3D images, geo-reference specimens through space and time, share the data and use them to provide low-cost but truly huge samples on which to apply modern GMM. In fact, as museums go on digitizing collections, including images of specimens (e.g., https://www.idigbio.org/, https://dissco.eu/, http://dmm.pri.kyoto-u.ac.jp/dmm/WebGallery/index.html), this type of data may become an extraordinary source of eco-morphological,

taxonomic and biogeographical information to quantify trends in space and time, with some groups and structures, such as insect wings and flat plant leaves, especially suited for the purpose. Then, good data, behind clever analyses and cool results, might indeed be round the corner.

ACKNOWLEDGEMENTS

To start, many thanks to Tim Smith, as well as two anonymous referees, for carefully reviewing the original manuscript, suggesting a number of changes that greatly improved the paper. Also, I thank them, and the readers, for putting up with the informal and rather personal style of this commentary, which I hope made the text a bit more entertaining. I would also like to remember Richard Thorington Jr. and Bob Hoffmann (both of whom I greatly miss), and thank a lot Dave Nagorsen, David Polly, Sarah Elton, Paul O'Higgins, Kate Nowak and all the other crucial contributors of my taxonomic and biogeographical studies on marmots and monkeys. Equally crucial, as these colleagues, were all museum curators and collection managers of the many institutions I visited to collect the data for these studies, as well as the funders who provided fundamental economic support for all these projects and mainly the Leverhulme Trust and SYNTHESYS. Finally, as I was revisiting my early years as a morphometrician, I realized once more the huge debt I have with those who helped me most in that early, and often difficult, stage: Horst Seidler, who warmly welcomed me in Vienna, as if I was one of his students, both in the 1998 and 2000 GMM workshops he organized for his group, and whom I hope to meet again sooner or later; Marco Corti (1950–2007), who led the GMM revolution in Italy since its early days and helped me, and a multitude of other young scientists, to learn the ABC of these difficult methods; Paolo Tongiorgi (1936–2018), who patiently put up with all the changes in direction of my PhD (none in the direction he would have liked!) and continued to support my career with extraordinary generosity since those early uncertain steps in 1998. I did not achieve much in science, but, without all these people, I would have achieved nothing.

ORCID

Andrea Cardini https://orcid.org/0000-0003-2910-632X

REFERENCES

Adams, D. C., Cardini, A., Monteiro, L. R., O'Higgins, P., & Rohlf, F. J. (2011). Morphometrics and phylogenetics: Principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. *Journal of Human Evolution*, *60*, 240–243.

Adams, D. C., & Otárola-Castillo, E. (2013). Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, *4*, 393–399.

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution'. *The Italian Journal of Zoology*, *71*, 5–16.

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix Italian Journal of Mammalogy*, *24*, 7–14.

Albrecht, G. (1992). Assessing the affinities of fossils using canonical variates and generalized distances. *Human Evolution*, *7*, 49–69.

Allison, D. B., Shiffrin, R. M., & Stodden, V. (2018). Reproducibility of research: Issues and proposed remedies. *Proceedings of the National Academy of Sciences*, *115*, 2561–2562.

Arbour, J. H., & Brown, C. M. (2014). Incomplete specimens in geometric morphometric analyses. *Methods in Ecology and Evolution*, *5*, 16–26.

Armitage, K. (2013). Climate change and the conservation of marmots. *Natural Science*, *5*, 36–43.

Armitage, K. B. (2009). Fur color diversity in marmots. *Ethology Ecology and Evolution*, *21*, 183–194.

Arnqvist, G., & Martensson, T. (1998). Measurement error in geometric morphometrics: Empirical strategies to assess and reduce its impact on measures of shape. *Acta Zoologica Academiae Scientiarum Hungaricae*, *44*, 73–96.

Barčiová, L., Šumbera, R., & Burda, H. (2009). Variation in the digging apparatus of the subterranean silvery mole-rat, *Heliophobius argenteocinereus* (Rodentia, Bathyergidae): The role of ecology and geography. *Biological Journal of the Linnean Society*, *97*, 822–831.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology Letters*, *13*, 246–264.

Björklund, M. (2019). Be careful with your principal components. *Evolution*, *73*, 2151–2158.

Blackith, R. E., & Reyment, R. A. (1971). *Multivariate morphometrics*. London, England, UK: Academic Press.

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*, 567–585.

Bookstein, F. L. (2017). A newly noticed formula enforces fundamental limits on geometric morphometric analyses. *Evolutionary Biology*, *44*, 522–541.

Bookstein, F. L. (2019). Pathologies of between-groups principal components analysis in geometric morphometrics. *Evolutionary Biology*, *46*, 271–302. https://doi.org/10.1007/s11692-019-09484-8

Bookstein, F. L., Sampson, P. D., Connor, P. D., & Streissguth, A. P. (2002). Midline corpus callosum is a neuroanatomical focus of fetal alcohol damage. *The Anatomical Record*, *269*, 162–174.

Brandler, O. V., Lyapunova, E. A., Bannikova, A. A., & Kramerov, D. A. (2010). Phylogeny and systematics of marmots (*Marmota*, Sciuridae, Rodentia) inferred from inter-SINE PCR data. *Russian Journal of Genetics*, *46*, 283–292.

Cardini, A. (2003). The geometry of the marmot (Rodentia: Sciuridae) mandible: Phylogeny and patterns of morphological evolution. *Systematic Biology*, *52*, 186–205.

Cardini, A. (2013). Geometric morphometrics. Dictionary entry in: UNESCO-EOLSS, Biological science fundamental and systematics, p. 1–52.

Cardini, A. (2014). Missing the third dimension in geometric morphometrics: How to assess if 2D images really are a good proxy for 3D structures? *Hystrix Italian Journal of Mammalogy*, *25*, 73–81.

Cardini, A. (2019). Integration and modularity in Procrustes shape data: Is there a risk of spurious results? *Evolutionary Biology*, *46*, 90–105.

Cardini, A., & Chiappelli, M. (2020). How flat can a horse be? Exploring 2D approximations of 3D crania in equids. *Zoology*, *139*, 125746. https://doi.org/10.1016/j.zool.2020.125746

Cardini, A., Dunn, J., O'Higgins, P., & Elton, S. (2013). Clines in Africa: Does size vary in the same way among widespread sub-Saharan monkeys? *Journal of Biogeography*, *40*, 370–381.

Cardini, A., & Elton, S. (2007). Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology*, *126*, 121–134.

Cardini, A., & Elton, S. (2008). Variation in guenon skulls (II): Sexual dimorphism. *Journal of Human Evolution*, *54*, 638–647.

Cardini, A., & Elton, S. (2009). Geographical and taxonomic influences on cranial variation in red colobus monkeys (Primates, Colobinae): Introducing a new approach to 'morph' monkeys. *Global Ecology and Biogeography*, *18*, 248–263.

Cardini, A., Filho, J. A. F. D., Polly, P. D., & Elton, S. (2010). In E. AMT (Ed.), *Biogeographic analysis using geometric morphometrics: Clines in skull size and shape in a widespread African arboreal monkey* (pp. 191–217). Berlin, Heidelberg, Germany: Springer Berlin Heidelberg.

Cardini, A., Hoffmann, R. S., & Thorington, R. W. (2005). Morphological evolution in marmots (Rodentia, Sciuridae): Size and shape of the dorsal and lateral surfaces of the cranium. *Journal of Zoological Systematics and Evolutionary Research*, *43*, 258–268.

Cardini, A., Jansson, A., & Elton, S. (2007). A geometric morphometric approach to the study of ecogeographical and clinal variation in vervet monkeys. *Journal of Biogeography*, *34*, 1663–1678.

Cardini, A., Nagorsen, D., O'Higgins, P., Polly, P. D., Thorington, R. W., & Tongiorgi, P. (2009). Detecting biological distinctiveness using geometric morphometrics: An example case from the Vancouver Island marmot. *Ethology Ecology and Evolution*, *21*, 209–223.

Cardini, A., O'Higgins, P., & Rohlf, F. J. (2019). Seeing distinct groups where there are none: Spurious patterns from between-group PCA. *Evolutionary Biology*. *46*, 303–316. https://doi.org/10.1101/706101

Cardini, A., Seetah, K., & Barker, G. (2015). How many specimens do I need? Sampling error in geometric morphometrics: Testing the sensitivity of means and variances in simple randomized selection experiments. *Zoomorphology*, *134*, 149–163.

Cardini, A., Thorington, R. W., Jr., & Polly, P. D. (2007). Evolutionary acceleration in the most endangered mammal of Canada: Speciation and divergence in the Vancouver Island marmot (Rodentia, Sciuridae). *Journal of Evolutionary Biology*, *20*, 1833–1846.

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, *114*, E6089–E6096.

Dapporto, L. (2008). Geometric morphometrics reveal male genitalia differences in the Lasiommata megera/paramegaera complex (Lepidoptera, Nymphalidae) and the lack of a predicted hybridization area in the Tuscan Archipelago. *Journal of Zoological Systematics and Evolutionary Research*, *46*, 224–230.

Dunn, J., Cardini, A., & Elton, S. (2013). Biogeographic variation in the baboon: Dissecting the cline. *Journal of Anatomy*, *223*, 337–352.

Estrada, A., Garber, P. A., Rylands, A. B., Roos, C., Fernandez-Duque, E., Fiore, A. D., et al. (2017). Impending extinction crisis of the world's primates: Why primates matter. *Science Advances*, *3*, e1600946.

Frost, S. R., Marcus, L. F., Bookstein, F. L., Reddy, D. P., & Delson, E. (2003). Cranial allometry, phylogeography, and systematics of large-bodied papionins (primates: Cercopithecinae) inferred from geometric morphometric analysis of landmark data. *The Anatomical Record*, *275A*, 1048–1072.

Fruciano, C. (2016). Measurement error in geometric morphometrics. *Development Genes and Evolution*, *226*, 139–158.

Fruciano, C., Celik, M. A., Butler, K., Dooley, T., Weisbecker, V., & Phillips, M. J. (2017). Sharing is caring? Measurement error and the issues arising from combining 3D morphometric datasets. *Ecology and Evolution*, *7*, 7034–7046.

Gewin, V. (2018). How to write a first-class paper. *Nature*, *555*, 129–130.

Grubb, P. (2006). Geospecies and superspecies in the African Primate Fauna. *Primate Conservation*, *20*, 75–78.

Gunz, P., & Mitteroecker, P. (2013). Semilandmarks: A method for quantifying curves and surfaces. *Hystrix Italian Journal of Mammalogy*, *24*, 103–109.

Gunz, P., Mitteroecker, P., Neubauer, S., Weber, G. W., & Bookstein, F. L. (2009). Principles for the virtual reconstruction of hominin crania. *Journal of Human Evolution*, *57*, 48–62.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate statistics*. Pearson International Ed, Harlow, UK.

Hallgrímsson, B., Katz, D. C., Aponte, J. D., Larson, J. R., Devine, J., Gonzalez, P. N., ... Marcucio, R. S. (2019). Integration and the developmental genetics of allometry. *Integrative and Comparative Biology*, *59*, 1369–1381.

Hawkins, B. A. (2012). Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, *39*, 1–9.

Hoffmann, R. S., Koeppl, J. W., & Nadler, C. F. (1979). The relationship of the amphiberigian marmots (Mammalia, Sciuridae). *Occasional papers of the Museum of Natural History, the University of Kansas*, *83*, 1–56.

Howell, D. C. (2012). *Statistical methods for psychology* (p. 796). Belmont, CA: Cengage Learning.

Hublin, J.-J., Weston, D., Gunz, P., Richards, M., Roebroeks, W., Glimmerveen, J., & Anthonis, L. (2009). Out of the North Sea: The Zeeland Ridges Neandertal. *Journal of Human Evolution*, *57*, 777–785.

Jackson, C., Baker, A., Doyle, D., Franke, M., Jackson, V., Lloyd, N., ... Traylor-Holzer, K. (Eds.). (2015). *Vancouver Island Marmot population and habitat viability assessment workshop. Final report*. Apple Valley, MN: IUCN SSC Conservation Breeding Specialist Group Retrieved from https://portals.iucn.org/library/node/46138

Kerhoulas, N. J., Gunderson, A. M., & Olson, L. E. (2015). Complex history of isolation and gene flow in hoary, Olympic, and endangered Vancouver Island marmots. *Journal of Mammalogy*, *96*, 810–826.

Klingenberg, C. P. (2008). Novelty and "homology-free" morphometrics: What's in a name? *Evolutionary Biology*, *3*, 186–190.

Klingenberg, C. P. (2013). Visualizations in geometric morphometrics: How to read and how to make graphs showing shape changes. *Hystrix Italian Journal of Mammalogy*, *24*, 15–24.

Klingenberg, C. P. (2016). Size, shape, and form: Concepts of allometry in geometric morphometrics. *Development Genes and Evolution*, *226*, 113–137.

Klingenberg, C. P., & McIntyre, G. S. (1998). Geometric morphometrics of developmental instability: Analyzing patterns of fluctuating asymmetry with Procrustes methods. *Evolution*, *52*, 1363–1375.

Kovarovic, K., Aiello, L. C., Cardini, A., & Lockwood, C. A. (2011). Discriminant function analyses in archaeology: Are classification rates too good to be true? *Journal of Archaeological Science*, *38*, 3006–3018.

Lane, D. M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2017). *An introduction to statistics*. Citeseer.

Lomolino, M. V. (2005). Body size evolution in insular vertebrates: Generality of the island rule. *Journal of Biogeography*, *32*, 1683–1699.

Maestri, R., Fornel, R., Gonçalves, G. L., Geise, L., de Freitas, T. R. O., & Carnaval, A. C. (2016). Predictors of intraspecific morphological variability in a tropical hotspot: Comparing the influence of random and non-random factors. *Journal of Biogeography*, *43*, 2160–2172.

Marcoulides, G. A., & Saunders, C. (2006). Editor's comments: PLS: A silver bullet? *MIS Quarterly*, *30*, iii–ix.

Marcus, L. F. (1990). *Traditional morphometrics*. In Proceedings of the Michigan morphometrics workshop. Vol. 2. pp. 77–122.

Millien, V. (2006). Morphological evolution is accelerated among Island mammals. *PLOS Biology*, *4*, e321.

Milne, N., & O'Higgins, P. (2002). Inter-specific variation in macropus crania: Form, function and phylogeny. *Journal of Zoology*, *256*, 523–535.

Mitteroecker, P., Gunz, P., & Bookstein, F. L. (2005). Heterochrony and geometric morphometrics: A comparison of cranial growth in *Pan paniscus* versus *Pan troglodytes*. *Evolution and Development*, *7*, 244–258.

Monteiro, L. R., Duarte, L. C., & Reis, S. F. (2003). Environmental correlates of geographical variation in skull and mandible shape of the punaré rat *Thrichomys apereoides* (Rodentia: Echimyidae). *Journal of Zoology*, *261*, 47–57.

Nagorsen, D. W., & Cardini, A. (2009). Tempo and mode of evolutionary divergence in modern and Holocene Vancouver Island marmots (*Marmota vancouverensis*) (Mammalia, Rodentia). *Journal of Zoological Systematics and Evolutionary Research*, *47*, 258–267.

Neubauer, S., Gunz, P., Leakey, L., Leakey, M., Hublin, J.-J., & Spoor, F. (2018). Reconstruction, endocranial form and taxonomic affinity of the early Homo calvaria KNM-ER 42700. *Journal of Human Evolution*, *121*, 25–39.

O'Higgins, P. (1997). Methodological issues in the description of forms. *Fourier Descriptors and their Applications in Biology*, 74–105.

O'Higgins, P. (2000). The study of morphological variation in the hominid fossil record: Biology, landmarks and geometry. *Journal of Anatomy*, *197*, 103–120.

Oxnard, C., & O'Higgins, P. (2009). Biology clearly needs morphometrics. Does morphometrics need biology? *Biological Theory*, *4*, 84–97.

Perez, S. I., Diniz-Filho, J. A. F., Bernal, V., & Gonzalez, P. N. (2010). Alternatives to the partial mantel test in the study of environmental factors shaping human morphological variation. *Journal of Human Evolution*, *59*, 698–703.

Polly, P. D. (2005). Development and phenotypic correlations: The evolution of tooth shape in *Sorex araneus*. *Evolution & Development*, *7*, 29–41.

Renaud, S., & Michaux, J. R. (2003). Adaptive latitudinal trends in the mandible shape of *Apodemus* wood mice. *Journal of Biogeography*, *30*, 1617–1628.

Rohlf, F. J. (1990). Morphometrics. *Annual Review of Ecology and Systematics*, *21*, 299–316.

Rohlf, F. J. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology*, *47*, 147–158.

Rohlf, F. J. (2000a). On the use of shape spaces to compare morphometric methods. *Hystrix Italian Journal of Mammalogy*, *11*, 1–17.

Rohlf, F. J. (2000b). Statistical power comparisons among alternative morphometric methods. *American Journal of Physical Anthropology*, *111*, 463–478.

Rohlf, F. J. (2003). Bias and error in estimates of mean shape in geometric morphometrics. *Journal of Human Evolution*, *44*, 665–683.

Rohlf, F. J. (2015). The tps series of software. *Hystrix Italian Journal of Mammalogy*, *26*, 9–12.

Rohlf, F. J., Loy, A., & Corti, M. (1996). Morphometric analysis of Old World Talpidae (Mammalia, Insectivora) using partial-warp scores. *Systematic Biology*, *45*, 344–362.

Rohlf, F. J., & Marcus, L. F. (1993). A revolution morphometrics. *Trends in Ecology & Evolution*, *8*, 129–132.

Rohlf, F. J., & Slice, D. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, *39*, 40–59.

Roth, V. (1993). On three-dimensional morphometrics, and on the identification of landmark points. *Contributions to Morphometrics. Madrid: Museo Nacional de Ciencias Naturales*, 41–61.

Schlager, S. (2017). Morpho and Rvcg – Shape analysis in R. In G. Zheng, S. Li, & G. Szekely (Eds.), *Statistical shape and deformation analysis* (pp. 217–256). Academic Press. https://www.sciencedirect.com/science/article/pii/B9780128104934000110

Schlager, S., Profico, A., Vincenzo, F. D., & Manzi, G. (2018). Retrodeformation of fossil specimens based on 3D bilateral semilandmarks: Implementation in the R package "Morpho". *PLOS ONE*, *13*, e0194073.

Schlick-Steiner, B. C., Seifert, B., Stauffer, C., Christian, E., Crozier, R. H., & Steiner, F. M. (2007). Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in Ecology & Evolution*, *22*, 391–392.

Seetah, T. K., Cardini, A., & Miracle, P. T. (2012). Can morphospace shed light on cave bear spatial–temporal variation? Population dynamics of Ursus spelaeus from Romualdova pećina and Vindija, (Croatia). *Journal of Archaeological Science*, *39*, 500–510.

Sheets, H., & Zelditch, M. (2013). Studying ontogenetic trajectories using resampling methods and landmark data. *Hystrix Italian Journal of Mammalogy*, *24*, 67–74.

Smith, R. J. (2018). The continuing misuse of null hypothesis significance testing in biological anthropology. *American Journal of Physical Anthropology*, *166*, 236–245.

Spiegelhalter, D. (2019). *The art of statistics: Learning from data* (p. 290). Penguin Random House, UK: Pelican Books.

Steppan, S. J., Akhverdyan, M. R., Lyapunova, E. A., Fraser, D. G., Vorontsov, N. N., Hoffmann, R. S., & Braun, M. J. (1999). Molecular phylogeny of the marmots (Rodentia: Sciuridae):

Tests of evolutionary and biogeographic hypotheses. *Systematic Biology*, *48*, 715–734.

Stone, G. N., Nee, S., & Felsenstein, J. (2011). Controlling for non-independence in comparative analysis of patterns across populations within species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*, 1410–1424.

Strauss, R. E. (2010). Discriminating groups of organisms. In E. AMT (Ed.), *Morphometrics for nonmorphometricians* (pp. 73–91). Berlin, Heidelberg, Germany: Springer Berlin Heidelberg.

Taylor, P. J., Kumirai, A., & Contrafatto, G. (2007). Species with fuzzy borders: The taxonomic status and species limits of Saunders' vlei rat, *Otomys saundersiae* Roberts, 1929 (Rodentia, Muridae, Otomyini). *Mammalia*, *69*, 297–322.

Tippmann, S. (2015). Programming tools: Adventures with R. *Nature News*, *517*, 109–110.

Viscosi, V., & Cardini, A. (2011). Leaf morphology, taxonomy and geometric morphometrics: A simplified protocol for beginners. *PLOS ONE*, *6*, e25630.

Watanabe, A. (2018). How many landmarks are enough to characterize shape and size variation? *PLOS ONE*, *13*, e0198341.

Watanabe, M. E. (2019). The evolution of natural history collections: New research tools move specimens, data to center stage. *Bioscience*, *69*, 163–169.