European Journal of Taxonomy 934: 1–92
https://doi.org/10.5852/ejt.2024.934.2527
ISSN 2118-9773
www.europeanjournaloftaxonomy.eu
2024 · Cardini A.

Monograph

urn:lsid:zoobank.org:pub:7D2C81B2-737F-441E-AC44-C743ED4063C7

A practical, step-by-step, guide to taxonomic comparisons using Procrustes geometric morphometrics and user-friendly software (part A): introduction and preliminary analyses

Andrea CARDINI

Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi, 103 - 41125 Modena, Italy.
School of Anatomy, Physiology and Human Biology, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia.
E-mails: alcardini@gmail.com, andrea.cardini@unimore.it
urn:lsid:zoobank.org:author:A63AF653-521E-4EE0-9623-7B9273DF7D0A

Table of contents

Abstract. Taxonomy lays the foundations for the study of biodiversity and its conservation. Procrustean geometric morphometrics (GMM) is a most common technique for the taxonomic assessment of phenotypic population differences. To measure biological variation and detect evolutionarily significant units, GMM is often used on its own, although it is much more powerful with an integrative approach, in combination with molecular, ecological and behavioural data, as well as with meristic morphological traits. GMM is particularly effective in taxonomic research, when applied to 2D images, which are fast and low cost to obtain. Yet, taxonomists who may want to explore the usefulness of GMM are rarely experts in multivariate statistical analyses of size and shape differences. In these twin papers, I aim to provide a detailed step-by-step guideline to taxonomic analysis employing Procrustean GMM in user-friendly software (with tips for R users). In the first part (A) of the study, I will focus on preliminary analyses (mainly, measurement error, outliers and statistical power), which are fundamental for accuracy, but often neglected. I will also use this first paper, and its appendix (Appendix A), to informally introduce, and discuss, general topics in GMM and statistics, that are relevant to taxonomic applications. In the second part (B) of the work, I will move on to the main taxonomic analyses. Thus, I will show how to compare size and shape among groups, but I will also explore allometry and briefly examine differences in variance, as a potential clue to population bottlenecks in peripheral isolates. A large sample of North American marmot mandibles provides the example data (available online, for readers to replicate the study and practice with analyses). However, as this sample is larger than in previous studies and mostly unpublished, it also offers a chance to further explore the patterns of interspecific morphological variation in a group, that has been prominent in mammalian sociobiology, and whose evolutionary divergence is complex and only partially understood.

Keywords. Group differences, marmot mandible, measurement error, outliers, replicability, P values, sampling error, shape analysis.

Cardini A. 2024. A practical, step-by-step, guide to taxonomic comparisons using Procrustes geometric morphometrics and user-friendly software (part A): introduction and preliminary analyses. *European Journal of Taxonomy* 934: 1–92. https://doi.org/10.5852/ejt.2024.934.2527

Introduction

Taxonomy and Procrustes

Taxonomy is often seen as an old-fashioned discipline with little appeal in the era of the big 'omics' (genomics, proteomics, transcriptomics etc.). Few universities, if any, have a course specifically teaching taxonomy. Although there is a wide range of views on the problem, taxonomy is seen by many as a field in a state of crisis (Giangrande 2003; Evenhuis 2007; Wheeler 2014; Kotov & Gololobova 2016). Yet, we cannot study organisms we have not described and named (a paradox known as 'taxonomic impediment' (Taylor 1983)), and thus taxonomy is crucial in providing the bricks for building biological science (May 1990). Genetic methods are now dominant in taxonomic research, as in many other fields of biological inquiry, but most taxonomists embrace a more pluralistic view. Thus, integrative taxonomy merges genetics with knowledge from disciplines as different as ecology, ethology and morphology (Dayrat 2005). In this broader context, GMM, a type of quantitative analysis of morphological variation (Rohlf & Marcus 1993; Adams *et al.* 2004), is one of the most used approaches. For taxonomic assessment, GMM rarely provides conclusive evidence on its own, but it offers a low cost and relatively simple tool for exploratory studies and phenotypic analyses complimentary to genetics (Cardini *et al.* 2022). Indeed, GMM analyses, often combined with molecular data, have become common in journals specifically devoted to taxonomy, such as the *European Journal of Taxonomy*, *Zookeys* and *Zootaxa* (e.g., in gerbils (Yazdi *et al.* 2014), lizards (Marín *et al.* 2016), fish (Craig *et al.* 2017), snails (Perez *et al.* 2021; Whelan *et al.* 2023; Miller *et al.* 2023), wasps (Schwarzfeld & Sperling 2014), fossil bees (Dehon *et al*. 2019), fruit flies (Hendrichs *et al.* 2015), beetles (Qubaiová *et al.* 2015; Su *et al.* 2015; Sasakawa 2016), hemipterans (Armendáriz-Toledano *et al.* 2023), spiders (Valdez-Mondragón *et al.* 2019), and centipedes (Gutierrez *et al.* 2011)).

The basic idea behind the GMM 'revolution' (Rohlf & Marcus 1993; Corti 1993) is to extend traditional morphometrics (TMM (Marcus 1990)), the quantitative study of, often subtle, morphological differences, by accurately measuring the size and geometric shape of biological structures (O'Higgins 2000; Adams *et al.* 2004). In the large family of GMM techniques (Sneath 1967; Rohlf 1990; O'Higgins 1997, 2000; Reyment 2010), Procrustes-based landmark analysis, the method I will be using, represents the leading approach (Adams *et al.* 2013). With this method, a configuration of 'homologous' (Smith 1990) anatomical landmarks is chosen on a study structure. The choice of the landmarks must be functional, and specific, to the research aim, so that different configurations might be available on a given structure depending on the research question (Klingenberg 2008; Oxnard & O'Higgins 2009; Cardini 2020b). For instance, in an evolutionary study, a valid landmark could be the meeting point of the sutures of homologous bones in a vertebrate cranium or the intersection of homologous wing veins or body segments in an insect. Of course, a landmark is meaningless on its own, but able to provide useful information in relation to other landmarks. Thus, using the Cartesian coordinates of all the digitized landmarks and a sample of individuals, a researcher can quantify the variation in size and shape of a structure.

More precisely, size in Procrustean GMM is a function of the sum of the distances between the landmarks and their centroid (i.e., the average of the raw coordinates), and it is, therefore, called centroid size (CS). The original raw coordinates are divided by CS, to standardize size[1], and new coordinates, measuring shape, are computed by centroid[2] centering the configurations and least-square superimposing all

[1] Standardizing size does not 'remove' allometric variation, if present. This is because, even if all individuals have the same CS = 1 after the division, any change in the proportions of a structure correlated to size is still there. To provide an intuitive analogy to understand why allometry is not removed by using standardized size, I use an example of ontogenetic variation, where size and shape changes are more pronounced and, thus, easier to see. In marmot ontogeny (Cardini & Tongiorgi 2003), younger individuals are smaller and also tend to have slender mandibles; in contrast, older individuals and especially large adults, progressively develop proportionally shorter, deeper and more robust mandibles. If one imagines an ontogenetic series of mandibles drawn to the same size (analogous to setting CS = 1 in a sample) and ordered from the smallest to the largest, the series will suggest the trend in increasing mandibular robustness despite having standardized size.

individuals. This procedure is named Procrustes superimposition (Sneath 1967; Gower 1975; Rohlf & Slice 1990; Goodall 1991). However, because the Procrustes shape space is a curved multivariate data space (Kendall 1989; Slice 2001) and most conventional statistical analyses require data to be in a Euclidean space (see Marcus *et al.* 2000, and references therein), data are projected into a flat Euclidean space tangent (usually in the sample mean shape) to the Procrustes shape space. This step, called the tangent space approximation, is well explained in TPSSmall (Rohlf 2015) and generally is performed automatically by GMM programs. It rarely makes a practical difference for the relatively small amount of shape variation typical of most biological datasets (Marcus *et al.* 2000). Yet, it is theoretically important, because it guarantees that differences between any two shapes can be computed as a straight line connecting them in the multivariate space, instead of requiring a distance metric that follows the curvature of the Procrustes shape space. This approximation allows the application of the most common statistical techniques without recurring to methods specifically tailored for curved non-Euclidean spaces.

Like all methods based on a biologically arbitrary mathematical superimposition, Procrustes shape analysis has some limitations (Richtsmeier *et al.* 2002; Cardini & Verderame 2022), but also a number of desirable statistical advantages (Adams *et al.* 2004). Procrustean GMM also allows an effective visualization of shape differences using a variety of diagrams (e.g., wireframes and deformation grids) and image rendering methods (Klingenberg 2013). Shape is inherently multivariate and Procrustes shape coordinates must be analysed all together (Rohlf 1998; Cardini & Verderame 2022). Thus, Procrustean GMM requires not only an understanding of its basic principles, but also a degree of knowledge of multivariate statistics.

Why a step-by-step guide with user-friendly software?

Because of the complexity of the mathematical treatment of the data, GMM is still perceived as a challenge by many biologists, including most taxonomists. The morphometric-statistical literature is, now, vast, but requires expertise to select the readings and a good amount of time to familiarize with the methods. Users that do not aim at specializing in GMM often prefer example papers and simplified guidelines rather than technical theoretical papers or extensive manuals. This is probably why the simplified 'protocol' for beginners by Viscosi & Cardini (2011, henceforth abbreviated as V&C) has been broadly cited. Indeed, V&C is mentioned in papers from a wide range of disciplines: from comparisons of sibling species in rodents (Jojić *et al.* 2014) and research on allometry in evo-devo (Klingenberg 2016) or palaeontological analyses of dinosaurs (Marugán-Lobón *et al.* 2013) to museum conservation and archaeology (Kuzminsky & Gardiner 2012; Okumura & Araujo 2019), the analysis of shell coiling (Polly & Motz 2016) and even studies of tool use in animals (Sugasawa *et al.* 2017). The analytical framework of V&C was general, but the article was originally aimed at botanists with a main interest in group comparisons. The current study, in contrast, has a specific focus on animal taxonomy and represents both an update and an extension of V&C. The descriptions of methods are more detailed and some of the analyses I added, were not included in V&C. Overall, the two new papers (A and B) should cover most of the topics a taxonomist, with little expertise in Procrustean GMM, needs to have a basic understanding of. Despite the length, readers interested in specific issues should be facilitated in selecting the relevant text by the subdivision of the articles in 'chapters' specific to different research questions. For instance, one might focus in part A on measurement error and outliers, and skip the section on statistical power, or, in part B, read only the 'chapter' on the discriminant analysis, when specifically and strictly interested in group classification.

[2] The centroid of a set of landmarks is the point defined by the mean of the landmark coordinates along the X, Y (and Z, for 3D data) axes.

For the same reasons why beginners may be seeking a simplified but detailed introduction to the field, they might also adopt user-friendly GMM software rather than scripts in a powerful, flexible, but complex statistical environment such as R (Claude 2008; R Core Team 2023). Free programs for statistics and GMM, with an intuitive windows interface, including PAST (Hammer *et al.* 2001), the TPS Series (Rohlf 2015) and MorphoJ (Klingenberg 2011), which I will be using in this work, offer a smaller range of analytical tools. Yet, these programs are available to everyone, including scientists from less wealthy countries, who might not have funds to buy commercial software, and they are still largely adequate for the majority of taxonomic analyses. However, the guidelines I provide are general and users can freely select different programs to implement the methods I exemplify. For instance, a researcher who is a good R-coder will easily be able to replicate the analyses using commands in R base (R Core Team 2023), plus additional functions in packages such as Morpho (Schlager 2017), Momocs (Bonhomme *et al*. 2014), car (Fox & Weisberg 2019) and vegan (Oksanen *et al.* 2022). Thus, in Table 1, together with the list of user-friendly software employed in A and B and a concise reference to the specific analyses performed in the different programs, I have included a few tips on R packages. Later, in part B, I briefly discuss the advantages of R and, in a few cases, exemplify R functions for specific analyses.

Compared to V&C, another difference is that in the current papers I dedicate more space to issues, that are often neglected in GMM publications, including: the complexity and impact of measurement error (ME); what outliers are and how to detect them; why sexual dimorphism (SDM) in gonochoric species cannot be overlooked; how sample size (N) and the number of variables (p) potentially affect findings; and how some fairly basic sensitivity analyses help assessing the robustness of the results. In both papers, I also occasionally digress to discuss some of the recent methodological developments, but I keep these introductory discussions brief, while providing references for those who want to learn more.

Inevitably, in a didactic application of methods, there is a bias that reflects my personal interests and experience in the last '20-plus' years of work in GMM. The study design I suggest is neither the best nor a solution to all problems and questions one might want to address using GMM in taxonomy. The general framework I have adopted is largely that of Rohlf *et al.*'s (1996) comparison of Old World moles using cranial data, which has become a model for many studies of GMM applied to taxonomy (see Introduction in part B). However, compared to this and several other 'old' papers, I avoid partial warps, which are concisely discussed in the Introduction of part B. Briefly, partial warps are a type of linear combination of the Procrustes shape coordinates, that are rarely needed in statistical tests and can easily lead to errors, unless users have a very clear understanding of the theory (Rohlf 1998; Cardini 2013, 2020a). Partial warps are, in fact, still found in some of the programs I will be using, but they are there for contingent historical reasons (Rohlf 2015) and should be computed only as an intermediate step (imposed by the software, but otherwise unnecessary) to get to the final output. This issue in the software implementation of some analyses is emphasized, whenever it may occur, in the short instructions I provide on how to perform a test in a specific program. The instructions are another new addition, compared to V&C. However, with the exception of more detailed information for a few analyses (B2 and B6) in TPSRegr (Rohlf 2015), these instructions are mostly concise tips to aid inexperienced users. Neither of the two papers is meant to be the equivalent of a software tutorial.

Background on the study group: North American marmots

The work I am publishing is mainly didactic and methodological, but the material is largely unpublished. Thus, the scientific findings are new, although they concern questions already investigated using smaller samples (Cardini 2003; Cardini *et al.* 2009; Nagorsen & Cardini 2009) in a group whose evolutionary history is still partly unclear (Kerhoulas *et al.* 2015; Mills *et al.* 2023). Specifically, I will be comparing North American marmots (see Table 2 for a list of species with scientific and common names) using a 2D dataset of hemi-mandibles in labial side view (henceforth simply referred to as 'mandibles'). The total sample size (Table 2) is almost twice that of the previous studies (citations above). The number of

Table 1 (continued on next page). List of user-friendly programs used in part A and B of the study with tips on R packages in which to perform analyses.

software	version used	webpage	OS*	main analyses done in this study	tips on R packages**
TPSUtil	1.84	https://sbmorphometrics.org/soft-utility.html	Windows; Linux with WINE	create TPS files from images; randomize specimen order; convert TPS to NTS or CSV; can also delete/ reorder specimens or landmarks, make wireframes etc.	-
TPSDig	2.31	https://sbmorphometrics.org/soft-dataacq.html	as above	digitization of landmarks on photographs (can also take linear measurements and has tools for image enhancement)	**Momocs** (comprehensive package for morphometrics)
TPSSmall	1.36	https://sbmorphometrics.org/soft-tps.html	as above	tangent space approximation (and 2D Procrustes superimposition, as in other TPS Series programs, but here also for 3D landmarks); Procrustes shape distances to mean shape or between individuals pairwise	**Morpho** (comprehensive package for GMM)
TPSRelw	1.75	as above	as above	PCA (in this software called Relative Warps analysis)	**base** *prcomp*(data, retx=TRUE, center= TRUE, scale=FALSE ...); **Morpho**; **Momocs**
TPSRegr	1.45	as above	as above	multivariate regression of shape onto independent variables (e.g., CS for allometry); MANOVA and MANCOVA using dummy variables	**car** for parametric tests using type III SS as in TPSRegr; **vegan** (*adonis* function) can do the same using permutations and type I SS
TPSPower	1.06	https://sbmorphometrics.org/soft-tutorial.html	as above	power analysis simulation for Procrustes shape data	-
G*Power	3.1	https://www.psychologie.hhu.de/arbeitsgruppen/ allgemeine-psychologie-und-arbeitspsychologie/gpower	Windows; Mac OS; Linux with WINE	power analysis for univariate data such as CS (which was not exemplified because easier and not specific to GMM)	many packages: search for "power" at https://cran.r-project.org/web/packages/ available_packages_by_name.html
MorphoJ	1.07d	https://morphometrics.uk/MorphoJ_page.html	Windows; Linux; Mac OS	Procrustes superimposition; search for outliers; ME ANOVA; PCA; regression; pairwise permutation tests of mean differences; angle between vectors; can also exclude landmarks or specimens, split into subsets, compute mean shapes, show confidence ellipses, visualize shape variation etc.	**Morpho** for superimposition, outliers, PCA, vector angles, visualization etc. (however, there may be small differences and Morpho currently does not do regressions and ME ANOVA†)

Table 1 (continued). List of user-friendly programs used in part A and B of the study with tips on R packages in which to perform analyses.

software	version used	webpage	OS*	main analyses done in this study	tips on R packages**
PAST	2.17††	https://www.nhm.uio.no/english/research/resources/past/index.html	Windows; Mac OS; Linux with WINE	search for outliers using plots (box and jitter plots, scatterplots with convex hulls etc.); PCA and bgPCA; cluster analysis; species by sex ANOVA of CS; permutation tests of mean size and shape differences; cross-validated CVA with classification table ('confusion matrix'); Levene's test for equal variance of univariate CS (see main text in B7 for its equivalent for shape); visualize shape variation using expansion factors	**stats** *hclust()* for cluster analysis; **car** for parametric ANOVA (one-way, two-way etc.); **Morpho** for cross-validated CVA (or bgPCA): **vegan** *adonis2()* function for tests of mean differences using permutations
Morpheus et al.	beta version	https://sbmorphometrics.org/morphmet/morpheus_vienna_2006.zip	Windows; Linux with WINE	superimposition, visualization of shape differences between pairs of cases (individuals, means etc.); not exemplified here, but can estimate missing landmarks (commands *list pmiss options* and *set pmiss imputation ...* etc.)	**Momocs**, **Morpho** and **shapes** are some of the packages that may have similar options for shape diagrams

* Windows software may run also in Mac OS using emulators such as WineBottler: https://winebottler.kronenberg.org/

** I am listing a few of the packages (emphasized in bold) I am more familiar with. There are others, such as, for instance, **geomorph** (Adams & Otárola-Castillo 2013) for GMM. However, users must be aware that there may be differences in methods and, therefore, they should carefully check the help files of the different R packages. As customary with R, I use italics for the name of specific functions.

† The ME ANOVA can be done in vegan with the *adonis2()* function using permutations, but the analysis requires two steps for results to be obtained with the same design as in MorphoJ (i.e., comparing species and sex variance to individual variance, and individual variance to variance between replicates): 1) a species+sex+individual ANOVA using replicates, as in MorphoJ, but only employed to obtain the result for the individual factor (totally ignore those for species and sex!); 2) after replicates are averaged within each individual (as for the analyses in part B), a species+sex ANOVA to have for these two factors results that should be virtually identical to those of the single ME ANOVA done in MorphoJ.

†† This version is available upon request (to me or the PAST mailing list: https://sympa.uio.no/nhm.uio.no/subscribe/past-users?previous_action=info); the link in the table, however, is to version 4, which is the latest and most updated version that unfortunately does not work well in Linux OS but has versions for both Windows and Mac OS.

Table 2. Sample composition by species and sex. In this and other tables, species names are abbreviated using the first three letters of the scientific names (see column for acronyms). Abbreviations: F = females; M = males; U = unknown sex.

Names				All				No. outliers				
Scientific	**Authority**	**Common**	**Acronym**	**F**	**M**	**U**	**Total**	**F**	**M**	**U**	**Total**	**% outliers**
M. (*M.*) *broweri*	Hall & Gilmore, 1934	Alaskan	bro	5	3	8	16	5	3	8	16	0%
M. (*P.*) *caligata*	Eschscholtz, 1829	hoary	cal	42	40	30	112	41	40	27	108	4%
M. (*P.*) *flaviventris*	Audubon & Bachman, 1841	yellow-bellied	fla	76	74	12	162	73	72	11	156	4%
M. (*M.*) *monax*	Linnaeus, 1758	woodchuck	mon	53	40	12	105	51	38	12	101	4%
M. (*P.*) *olympus*	Merriam, 1898	Olympic	oly	7	7	0	14	7	7	0	14	0%
M. (*P.*) *vancouverensis*	Swarth, 1911	Vancouver isl. marmot	van	9	11	33	53	9	10	31	50	6%
		all north Am. marmots		192	175	95	462	186	170	89	445	4%

anatomical landmarks used to measure mandibular size and shape has also been increased from nine to 15, even if I demonstrate that half of the 'new' landmarks have low precision and, thus, only 12 in total are eventually included in the main analysis. Before focusing on the research questions and methods, however, I wish to provide some basic information on marmots and their morphological and molecular evolution.

Marmots belong to a large Holoarctic radiation of ground-squirrels (Herron *et al.* 2004; Mills *et al.* 2023). They are the largest true-hibernating mammals, range from solitary to highly social species with extended families and helpers, and, being cold-adapted, mostly live in high altitude/latitude open prairies or steppe environments, where they dig complex systems of burrows (Armitage 2000, 2014). Our understanding of the evolutionary history of this lineage is limited, but has improved over the last two decades. The genus originated during the Miocene in North America, probably in cool, moist lowland grassland or periglacial habitats, and only later, in the late Pliocene, colonized the Palaearctic, where today marmot diversity is highest with nine of the 15 modern marmot species (Armitage 2014; Mills *et al.* 2023). Since the pioneering molecular study of Steppan and colleagues (Steppan *et al.* 1999), we have also learnt that four of the six species of North American marmots likely constitute a monophylum, the subgenus *Petromarmota*, whereas the remaining two species, the woodchuck and the Alaskan marmot, belong to the mostly Eurasiatic subgenus *Marmota*. Steppan *et al.* (1999) also showed that, at least using the mitochondrial cytochrome b gene, the critically endangered Vancouver Island marmot (VAN) shows a degree of divergence from its continental sister species, the hoary marmot, which is similar to or smaller than found within other species of marmots. This finding has potential implication for deciding whether VAN, possibly the most endangered Canadian mammal (Roach 2017) and one of the most iconic North American species at risk of extinction (Wilson 2002), truly represents a top conservation priority.

Later, in fact, Steppan and colleagues (Steppan *et al.* 2011) found inconsistencies between mitochondrial and nuclear DNA (with the latter failing to support the subgenus *Marmota* and instead favouring a monophyletic origin of all North American species) and suggested a longer history of isolation (~ 0.4-1.2 Ma – million of years) for VAN. More recently, Kerhoulas *et al.* (Kerhoulas *et al.* 2015) also supported a longer separation of VAN, but they demonstrated that it was incomplete and suggested an intricate scenario of molecular evolution. Thus, mitochondrial genes recovered two separate, largely allopatric clades of hoary marmots, with VAN sister to the coastal clade. Yet, nuclear genes failed to support the two clades, leaving open the question of the exact evolutionary history of the VAN-hoary marmot species complex. Likely, part of the incongruence between types of DNA happens because introgression brought VAN to acquire mitochondrial genes from the hoary marmot. Kerhoulas and colleagues also found evidence of a potential gene flow between the VAN-hoary marmot lineage and the Olympic marmot, which is also a member of *Petromarmota*. However, all molecular data supported a much longer separation of the Olympic marmot that today survives in a small population on the Olympic Peninsula, just south of the Vancouver Island, and an even older divergence for the more southerly distributed yellow-bellied marmot, the fourth species of *Petromarmota*.

A recent study (Rankin *et al.* 2019), using mitochondrial markers from a large sample of *Petromarmota* marmots, has corroborated the phylogeny of Kerhoulas *et al.* (2015). This latest research also suggested that yellow-bellied marmots not only represent a deeper split in the radiation of *Petromarmota*, but also has a genetic diversity approximately six times larger than found in either clade of hoary marmots. Hoary marmots, likely, had their range fragmented and went through one or more strong bottlenecks, as glacial conditions sharply reduced habitat availability in the north of the continent. Yellow-bellied marmots, in contrast, could survive in the milder climate of the south in large and well-connected populations, which preserved more diversity.

Also, the most recent analysis of mitogenomes and ultraconserved molecular regions (Mills *et al.* 2023) supports divergent lineages within hoary marmots, although with incongruences between the two types of data, and shows that, as in Kerhoulas *et al.* (2015), VAN and the Olympic marmot are nested within the radiation of the hoary marmot superspecies complex. Mills *et al.* (2023) also strengthen the hypothesis of a North American origin of marmots, with woodchucks and Alaskan marmots originating before the single migration to Eurasia that started the marmot radiation on this continent. However, they acknowledge that more data and better sampling of geographic variability will be needed to further clarify the evolutionary relationships of *Marmota*.

The first molecular phylogenies of marmots (Kruckenhauser *et al.* 1999; Steppan *et al.* 1999) were published precisely at the time when, as a PhD student, I started my own study of the morphological differences in this group. Over the years, using GMM, I showed that the molecular subgenera were partially supported by mandibular morphology, but, in sharp contrast to the results from DNA, VAN appeared as the most distinctive species in the entire genus (Cardini 2003). This incongruence, I speculated (Cardini 2003), was indicative of a highly accelerated morphological change in an insular population of mammals (Millien 2006), likely as a consequence of genetic bottlenecks (Nagorsen & Cardini 2009) and differences in environmental pressures on the island after its isolation from the continent, due to sea level rise at the end of the last glaciation. In a series of papers with several co-authors, we confirmed that, consistent with findings from studies of alarm calls (Blumstein 1999), VAN is not just a dark-brown 'version' of the hoary marmot, its lightly coloured parental lineage on the mainland. VAN is almost unique among marmots for several aspects of cranial (Cardini *et al.* 2005, 2007) and mandibular morphology (Cardini 2003), a finding confirmed by palaeontological data (Nagorsen & Cardini 2009). Thus, by complementing genetic findings with phenotypic evidence, we documented distinctive aspects of evolutionary variability overlooked by early DNA analyses (Steppan *et al.* 1999). Also, using GMM, we anticipated some of the conclusions of later molecular studies, which showed an intricate and likely deeper history of the 'hoary marmot species complex' (Kerhoulas *et al.* 2015; Rankin *et al.* 2019; Mills *et al.* 2023). By documenting its large phenotypic disparity, we suggested that the evolutionary potential of this clade of closely related populations might be in fact larger than inferred using a limited number of genetic markers from a progressively larger but still inadequate geographic sampling, a conclusion supported by the latest molecular research (Mills *et al.* 2023).

Study structure: why hemi-mandibles?

The main discoveries on marmot morphological evolution came from studies of mandibles and crania (see above) that, together with teeth (Caumul & Polly 2005; Evans 2013), are a main source of morphometric data for taxonomic and evolutionary research on mammals and other terrestrial vertebrates. Skull and teeth also represent a primary source of diagnostic characters (for examples in mammals, see Kelt & Patton 2020). Cranial and dental data may, thus, seem almost a default option in vertebrates. Indeed, this study also focuses on mandibular differences among species of marmots. However, regardless of the group, the choice of an appropriate study structure in morphometric research is in fact a fundamental step, that must be carefully considered.

The anatomical region one decides to measure should be tailored to the specific research question (Oxnard & O'Higgins 2009). A taxonomist must use her/his expertise to make this important decision. Similarly, once the structure has been chosen, the researcher will have to carefully consider what are the landmarks that might better describe taxonomic variation using this anatomical region in the specific group she/he is investigating. Searching the morphometric literature for examples is useful and it would be good to find a common configuration, on the same anatomical features, in different but related taxa. However, examples might be rare in groups that are less studied, and configurations used in previous work should be assessed with a critical eye: they may or may not be appropriate examples and, often, they need to be at least partly modified to suit specific aspects of a study. As I discuss below (and elsewhere in these

and other papers (Cardini 2020a)), a deep understanding of anatomy and biology provides the main guide for deciding the morphometric descriptors. Thus, it is the expert of the taxonomic group who must choose what to measure and how. Nonetheless, practical considerations and a basic understanding of the methods are also important.

In vertebrate taxonomy, skulls represent complex and highly informative structures. Dental anatomy is simpler, but highly relevant to understand function and ecology (in relation to diet, but also as a weapon in agonistic interactions). Teeth are characterized by correlated evolutionary patterns (Kangas *et al.* 2004) and can be particularly useful in taxonomic and evolutionary research. This is partly because they might evolve more slowly and be less plastic than other skeletal parts (Caumul & Polly 2005; Hulme-Beaman *et al.* 2019; Karagic *et al.* 2020), although this is not a rule. Post-cranial material is also potentially important for morphometric studies in mammals and other vertebrates and, for taxonomy, is often analysed in combination with cranial data (in primates, for instance, see Sargis 2002; Sargis *et al.* 2008, 2017; Kenyon-Flatt *et al.* 2020).

The tendency to preferentially focus on skulls and teeth in mammals and, more generally, vertebrates is, however, often largely a matter of convenience. Teeth and crania are well represented in museum collections. Mandibles, or more accurately hemi-mandibles, also have another practical advantage, which has contributed to make them popular (Corti *et al.* 1996; Duarte *et al.* 2000; Klingenberg & Leamy 2001; Caumul & Polly 2005; Bastir *et al.* 2007; Grossnickle 2017) in GMM[3]: they tend to be relatively flat, and thus potentially suitable for 2D analyses on photographs. 2D data collection is faster, which reduces costs. A rapid data collection, on an abundant museum material, helps to obtain large samples, essential for quantitative comparisons of small differences. This practical advantage is one of the reasons why my first morphometric work was on mandibles. It is also a reason for their use in this study, as I already had a large, mostly unpublished, mandibular dataset of adult marmots. Thus, my aim is first and foremost to provide a step-by-step guideline for taxonomic comparisons using GMM, but also to extend the analysis of Nagorsen & Cardini (2009) on *Petromarmota* to a larger dataset of all North American marmot species.

Outline of part A and B

As already explained, because a careful taxonomic study requires a fairly long series of steps and analyses, I have split the work in two parts: A) preliminary analyses; B) group comparisons. For the same reason, after the general introduction and the description of the study material, and before the main conclusions, both papers are also subdivided in 'chapters', with methods, results and discussion for each research question. This type of organization should make the articles easier to read and consult.

In the methodological sections, readers can also find tips on how to implement the analyses in the free user-friendly software I will be using. I opted for this format, instead of moving the brief instructions on software to a dedicated appendix, because I hope this may stimulate beginners to try the analyses. This could be done, using, first, my example data (available as supplementary files), as soon as a researcher reads the relevant theoretical introduction, which is similar to the organization of most morphometric workshops, where an introductory lecture is followed by a practical session.

In the preliminary analyses, together with some general considerations on statistical methods (mainly in Appendix A), I have included the assessment of ME, the detection of outliers and an example of power analysis. The power analysis anticipates a few results of part B, which is dedicated to the main taxonomic analyses. Readers can, as anticipated, skip a chapter and focus on a specific topic or, for instance, if interested, could go back to the chapter on statistical power in this paper after having read those on sex and species differences in part B.

[3] More references on GMM studies using mandibles can be found in Cooke & Terhune (2015).

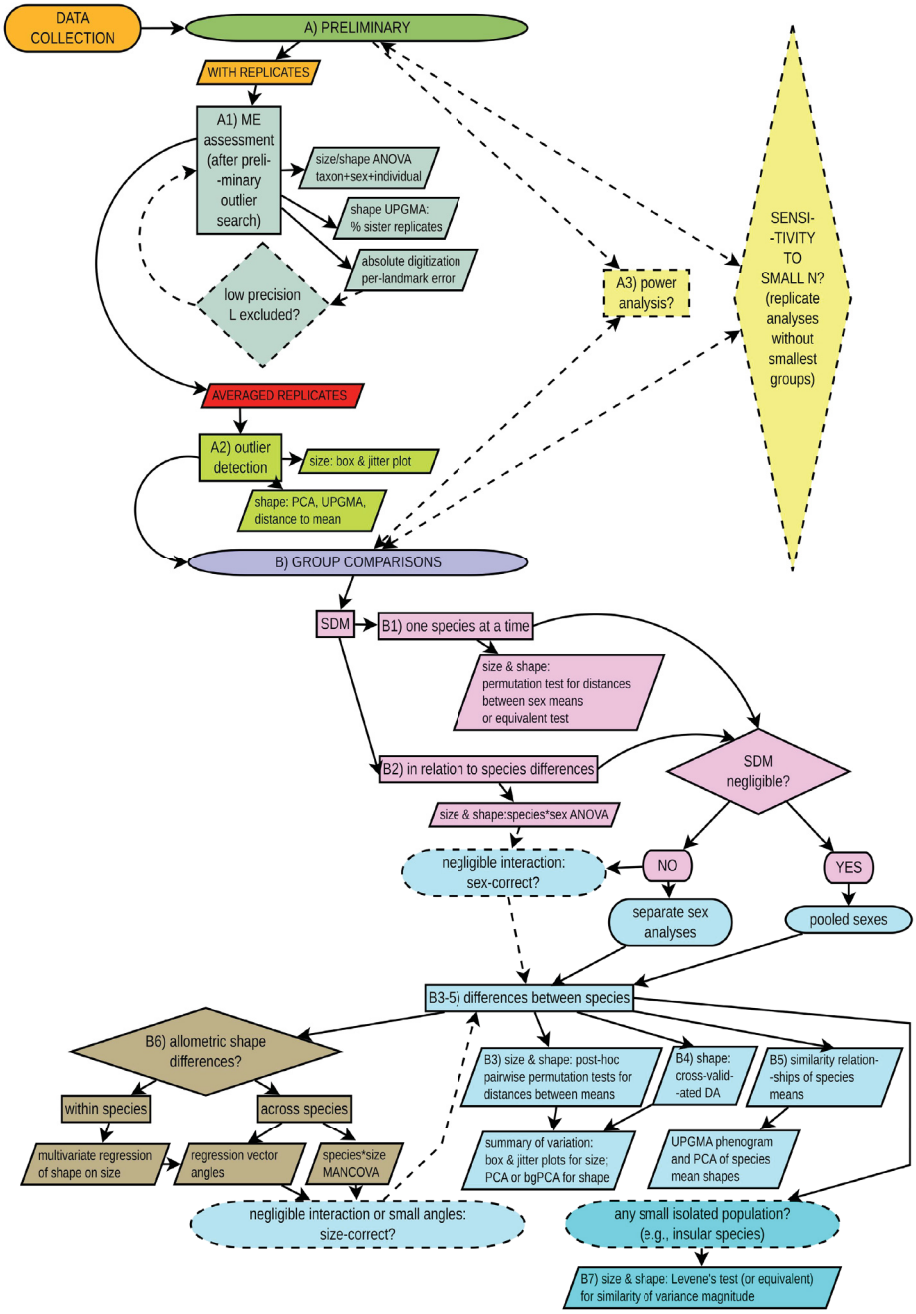


Fig. 1. Study flowchart for both preliminary (A) and main (B) analyses. The flowchart can be used as a reminder for the main analytical steps in a taxonomic study using GMM. To the same aim, at the end of part B, I added a checklist (Appendix B). In the flowchart, I have included the power analysis and few other analyses, which are optional (dotted lines). The power analysis is shown here connected to both the preliminary steps and the group comparisons, because it can be either prospective or retrospective. The sensitivity of results to the inclusion or exclusion of the smallest samples is also connected to both preliminary and main analyses, because it can be used at any step in the analysis. Sometimes (e.g., in the discriminant analysis (DA) of shape), if p is large relative to N and dimensionality reduction is needed, one could also assess the sensitivity of results to the inclusions of different number of PCs. I stress that, as discussed in the main text of both parts A and B, if a taxon is known to have a large SDM, which may vary in pattern depending on the species or subspecies, even preliminary analyses (such as those for ME or outlier detection) should probably be run with separate sexes.

Below, as well as in the flowchart of Figure 1, I summarize the study outline for the entire work. However, the detailed description of the steps of part B is postponed to the second paper. Before focusing on the study questions, I stress that in both papers, unless specified otherwise, the same tests are done in parallel on mandibular size and shape. In contrast, I will not perform any type of form analysis. What is form? Form is simply a way of simultaneously combining size and shape information in the same set of variables. For instance, a type of form space is obtained by appending the natural logarithm of CS to the matrix of Procrustes shape coordinates and another type by multiplying shape coordinates by CS or by simply omitting the standardization of size in the Procrustes superimposition (Klingenberg 2016, 2022). Form space analyses can be interesting in various ways and contexts, but probably most of the time taxonomists prefer to avoid that group differences are inflated or confounded by mere size variation. However, if one desires to explore taxonomic groups in a form space, he/she can easily analyse form variables by applying mostly the same multivariate methods I will be using on shape data in this and its twin paper.

A) Preliminary analyses

A1) *Measurement error.* Data on each individual are collected twice and the difference between duplicates used to assess ME in relation to the magnitude of biological variation in the samples. In this study, for simplicity, I will only be considering digitization error, but the same analytical design can be used to investigate other sources of error (specimen orientation, measuring device, 2D to 3D approximation etc.).

A2) *Outlier detection.* Outliers are looked for and interpreted, as with ME, in relation to the amount of variation observed within and across groups. This analysis, as well as all the next ones (A3 and part B), are done on the averages of the two digitizations of each individual used in A1.

A3) *Example of power analysis.* The estimate of statistical power (the probability of detecting an effect, such as group differences in size and shape, when real) is exemplified for pairwise tests of sex and species mean differences.

B) Group comparisons (see part B for details)

B1) *Sexual dimorphism within species.*

B2) *Sexual dimorphism in relation to species differences.*

B3) *Pairwise tests of species mean differences.*

B4) *Species discriminant analysis (DA).*

B5) *Summary and visualization of species mean shape differences*.

B6) *Relationship between shape and size within and across species.*

B7) *Species comparisons of the magnitude of size and shape variance.*

Material and methods

In this section, I describe the software, material and techniques for data acquisition. Methods specific to a study question will be explained at the beginning of the chapter dedicated to that question. In the methods, there may be short comments on aspects that are secondary to the main discussion and would be a distraction if placed there. Readers can also find, in Appendix A, some introductory considerations on methodological issues with small samples, the frequentist statistical approach, and some of the common assumptions of the tests I will be using, as well as a brief discussion on what semilandmarks are, their advantages and disadvantages, and an informal glossary of selected terms.

Software

Analyses are run in three main sets of programs: the TPS Series (Rohlf 2015), MorphoJ (Klingenberg 2011), and PAST (Hammer *et al.* 2001), which are freely available online (https://sbmorphometrics.org/toc-software.html or Table 1 for direct links to software webpages). These programs are complimentary, because specific analyses may be available in just one or the other software. Often, however, the same test can be performed in more than one program. When a method is implemented in multiple programs, readers are free to decide the software they prefer and, in general, they might want to look for alternatives beyond the selection I made for this work. Especially with programs that are not specific to Procrustean GMM, however, I suggest caution and will exemplify, in both A and B, some of the issues one might occasionally encounter. Furthermore, even if I provide short instructions for the implementation of most analyses, readers are invited to consult the help files of the programs. Often, the help file also includes brief explanations and references on methods. Finally, searching online for 'MorphoJ tutorials' or 'TPSDig tutorial', for example, might allow to find introductory videos on how to start using these programs.

In the Introduction I have explained why I chose free user-friendly software for all analyses. Yet, in spite of several advantages for beginners, this choice implies some inevitable limitations (see Conclusions in part B) that also concern the availability of the versions for the different operating systems (Table 1). MorphoJ is an exception in this respect. Like R (R Core Team 2023), MorphoJ is multiplatform and runs on Windows, Mac OS and Linux. Other programs, in contrast, are more specific. The TPS Series is only Windows, while PAST has both Windows and Mac versions. Despite these constraints, Windows software can generally be run in other operating systems using an emulator (e.g., WINE, in Linux). For PAST, Linux users, like me, are unfortunately bound to use an old version (2.17c). Despite some recent improvements, later versions (PAST 3 and 4) do not seem to work properly with emulators in the main distributions of Linux. PAST 2.17c is no longer maintained, but it is still made available by subscribers to the PAST email list (past-users@nhm.uio.no). However, for Mac and Windows users, as well as for those willing to install Windows in a virtual machine in Linux, I strongly suggest employing the latest version of PAST. Analyses will be mostly the same as in PAST 2.17c and commands will be also similar, but the windows interface is new, older bugs have been fixed and some new types of analyses have been added. In terms of file format, the older and newer versions of PAST seem to be fairly compatible, but users may have to do some small tweaks for full compatibility.

With all programs, as a convention in both part A and B, I will use italics for the tips on commands as they appear in the scroll-down window menu of a specific software, and commas to separate submenus from the main menu.

Samples

All six north American marmot species are included in the study. Scientific and common names, taxonomic authority and sample sizes are shown in Table 2. In this table, readers can also find the abbreviations for the species names used in all figures and tables of both papers. Specimens originate mostly (81% of the total) from the collection of the National Museum of Natural History (USNM, Washington, D.C., USA), but some are from the Field Museum of Natural History (FMNH, Chicago, USA), University of Kansas Natural History Museum (KUNH, Lawrence, USA), Museum of Vertebrate Zoology (MVZ, Berkeley, USA), Royal British Columbia Museum (RBCM, Victoria, Canada), University of British Columbia (UBC, Vancouver, Canada), Zoological Museum of the University of Montana (UMZ, Missoula, USA), as well as from the private collections of R.L. Rausch and D.W. Nagorsen.

All individuals are adults with fully erupted permanent dentition. Using the same age class is necessary, because both size and shape change during marmot mandibular ontogeny (Cardini & Tongiorgi 2003; Cardini & O'Higgins 2005) and, therefore, variability in age might confound interspecific comparisons.

The vast majority of the specimens were collected in the wild between 1874 and 1994. Approximately 3/5 of the VAN sample, however, consists of subfossils, as detailed in Nagorsen & Cardini (2009). The subfossil sample is included to increase sample size in this species, which is poorly represented in museum collections. As I discuss further, later, allochronic samples are generally better kept as separate groups in taxonomic comparisons. However, previous work (Nagorsen & Cardini 2009) demonstrated that differences in mandibular morphology between the archaeological samples and modern VAN are much smaller than interspecific variation, a finding corroborated by this study (part B).

Digital images and landmark configuration

Standardized digital images of the mandible labial side were captured using a flatbed scanner at a resolution of 300 dpi, as described in Nagorsen & Cardini (2009). For 98% of the specimens, I scanned the left hemi-mandible. For 10 individuals of VAN, however, only the right hemi-mandible was available and, thus, scanned. Before landmarking right hemi-mandibles, I flipped their images to make them look as if they were all left sides. This operation reduces the risk of unintentional biases in the landmark digitizations (Nagorsen & Cardini 2009). Mandibular size and shape differences between left and flipped right marmot hemi-mandibles have been shown to be negligible and, therefore, the choice of the side does not affect results (Nagorsen & Cardini 2009). In a previous work (Cardini 2017), I have discussed whether to measure one or both sides of bilaterally symmetric structures in relation to the study aims. In taxonomy, most of the time, one is not interested in small asymmetries and measuring only one side makes data collection faster, at the cost of a generally negligible reduction in accuracy. However, if this is done with structures with object symmetry (Klingenberg *et al.* 2002), such as crania, which consist of mirror halves, I suggest to 'symmetrize' the half that has been landmarked (Cardini 2017). For structures with matching symmetry (e.g., the left and right hand or the two disarticulated hemi-mandibles), when there is no directional asymmetry, one can measure one or the other side, but each individual must be represented either by its left or by its right side (never both, as they would be pseudo-replicates). Alternatively, a researcher can measure both left and right hemi-mandibles, but the size and shape of each specimen must be then averaged between sides.

Generally, regardless of the type of symmetry, if I am not measuring both sides, I prefer to consistently use the same side in all individuals, so that I do not have to test for subtle directional differences between left and right, before using the data in the taxonomic comparisons. There might be rare exceptions in which consistently measuring the same side does not avoid potential biases. For instance, if, in a population of crabs, the left claw is larger than the right but in another population the asymmetry is reversed, a researcher will have to take this difference into account. A simple solution might be to compare larger claws with larger claws and smaller with smaller, but the decision on how to deal with this type of issues cannot rely on a general rule and should be adapted specifically to the study group using the best available knowledge on its biology.

I digitized landmarks on the 2D images using TPSDig 2.31, which is part of the TPS Series (Rohlf 2015). The TPS Series includes other programs used in this study (Table 1), such as TPSUtil 1.82, TPSSmall 1.34, TPSPower 1.08, TPSRelw 1.69 and TPSRegr 1.45. The two main file formats of the TPS Series are TPS and NTS. They mainly store the landmark coordinates and are simple ascii text files whose format is well described in the help of the software. In both papers (part A and B), I indicate file extensions using uppercase (e.g., TPS) or the conventional notation *.extension (e.g., *.tps), where the asterisk is the file name. When files are created or manually edited in a text editor, users simply need to rename the *.txt extension with *.tps or *.nts.

A single TPS file to load all the images in TPSDig is easy to make in TPSUtil, which has a very intuitive menu and allows a variety of other data manipulations. For instance, the operation to create a TPS file is called *Build tps file from images*. After creating the TPS file, I generally randomize the order of the individuals in TPSUtil (*Randomize order of specimens*). This is done to avoid other potential unintentional biases, that may occur if, for instance, one has organized data in groups (e.g., first all females of species A, then all males, then females of species B etc.). Having data organized in groups is, in fact, a good idea and, after digitizing the landmarks on randomized data, one can go back to the original order using the TPSUtil command *Restore original order*.

Using the randomized dataset, I opened the TPS file in TPSDig, set the scale factor to convert coordinates from pixel to mm, and digitized the landmarks twice, with an interval of at least one week between the first and second digitization. Because all the digital images in my dataset have been acquired with the same magnification, the scale factor can be set only once in the first image. TPSDig will reuse this specific scale factor in all other images, unless an operator specifies a different one. It is easy to check that a scale factor is correct by measuring a known distance using the *make linear measurement* tool in TPSDig [4]. The scale factor can also be manually replaced with the average of multiple measurements of scale (e.g., re-measuring more than once several cm, both horizontally and vertically, on the millimeter paper of many images taken with the same zoom). The average scale factor will be more accurate. Of course, one has to pay a lot of attention when manually editing files and, in general, check carefully that the zoom used during the image acquisition was indeed the same in all individuals [5]. If not, the TPSDig scale factor must be reset in each image.

For the analysis, I initially selected fifteen homologous landmarks (L) (Fig. 2a). Most of them have already been used in previous studies (see Introduction), including Cardini & Tongiorgi (2003), from which I borrow most definitions: (L1) upper extreme anterior part of the incisor alveolus; (L2) anterior top of the mandibular symphysis; (L3) anterior extremity of the maxillary toothrow (premolar alveolus); (L4) contact point between the premolar and first molar projected lingually onto alveolar margin; (L5) contact point between the second and third molar on the lingual alveolar margin; (L6) intersection of the dental ridge with the dorsal portion of the masseteric ridge (base of the coronoid process); (L7) tip of the coronoid process; (L8 and L9) anterior and posterior tips of the condyle; (L10) posterior extremity of the angular process; (L11) anterior extremity of the scar marking the insertion of the superficial masseter on the margin of the angular process; (L12) incisura vasorum facialum; (L13) most anterior point on the masseteric ridge; (L14) mental foramen and (L15) lower extreme anterior margin of the incisor alveolus. Because the assessment of ME showed that L11, L12 and L13 have a much lower precision than other landmarks (see below), I later excluded these three landmarks and used the remaining 12 (Fig. 3a) for all main analyses.

[4] For instance, from time to time, I re-measured a distance of 40 mm either horizontally or vertically on the millimeter paper frame placed around each mandible and checked that the measurements were accurate.

[5] To avoid misunderstanding, I stress that here I am talking about the magnification used when the image was acquired (with a camera, scanner etc.). For instance, the zoom could be ×1, ×10, ×25 or something else. In that case, unlike in my marmot dataset with all images acquired with the same magnification and resolution, one has to set up a different scale factor depending on the camera zoom. With a variable zoom during data collection, a researcher should vary the distance of the camera also when replicate pictures are taken, so that this source of error too is included in the analysis. However, once the correct scale factor is set in each photograph, I may use the zoom tool of TPSDig like a magnifying lens to see more clearly specific landmarks. When zooming in (or out), TPSDig does not change the scaling of the coordinates, as easily verified with the *make linear measurements* tool to measure the same distance (e.g., 1 cm) before and after zooming.

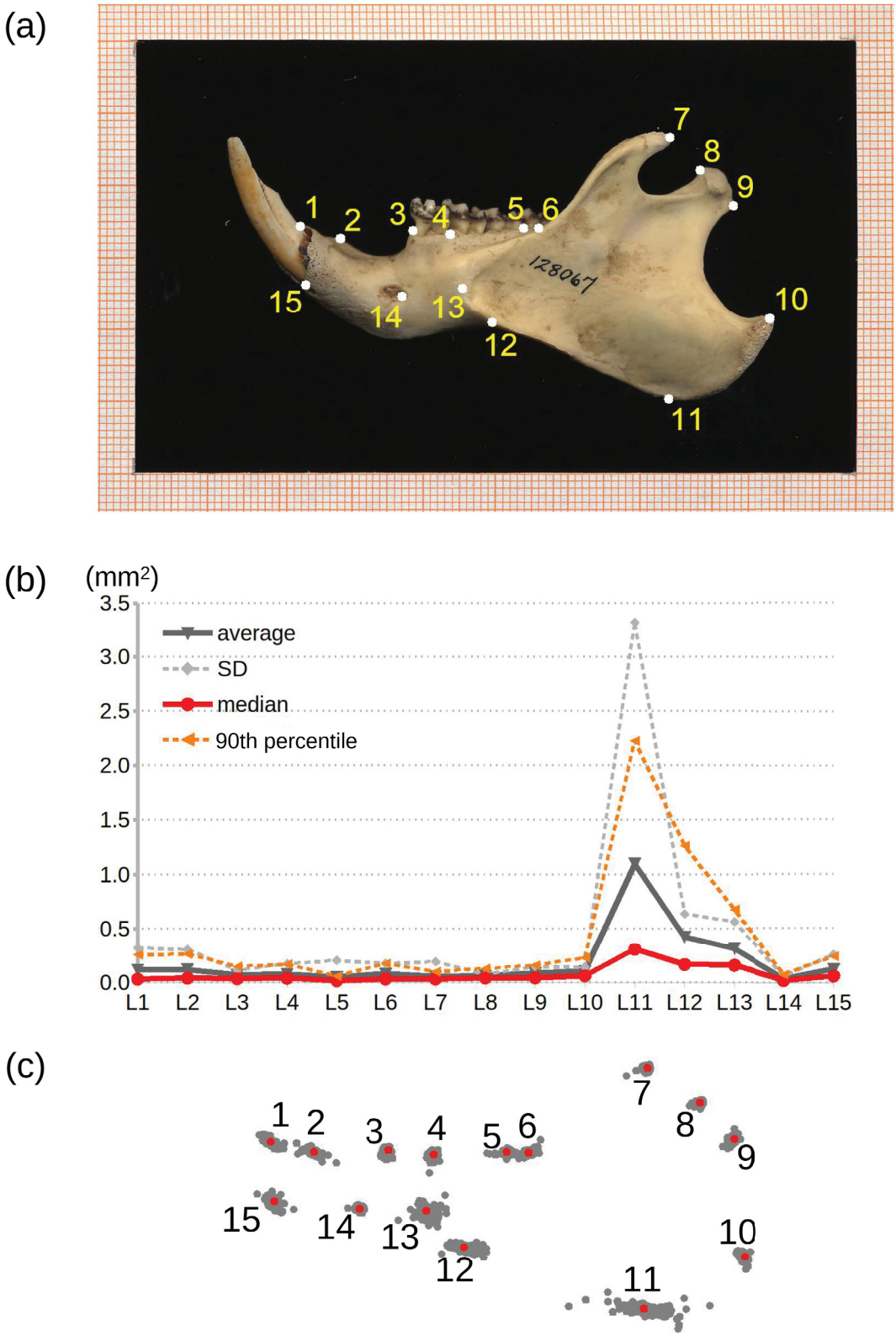


Fig. 2. Initial configuration with 15 landmarks (a) and analysis of absolute per-landmark imprecision (b, c). Figure 2b shows the profile plot for the summary statistics of per-landmark variance in the two digitizations. Figure 2c shows the scatter of landmarks purely due to digitization error (red landmarks mark the mean form, to which the differences between the first and second digitization were added).

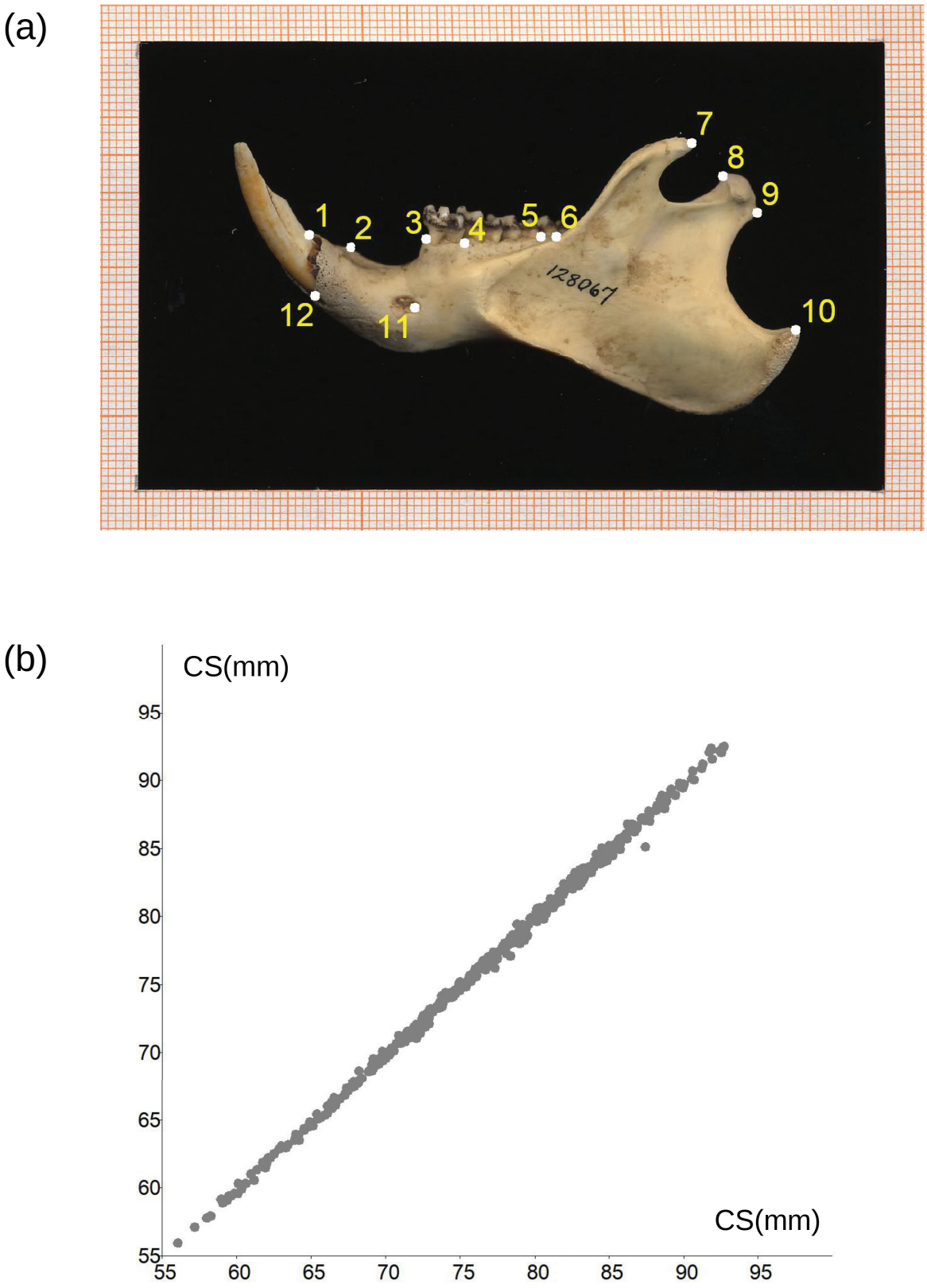


Fig. 3. a. Final configuration, reduced to 12 landmarks after excluding low precision landmarks. **b**. Graphical examination of size replicability (12 landmarks configuration) using a plot of CS in the second duplicate against CS in the first duplicate.

Methods, results and discussion subdivided by study question

A1) Measurement error (ME)

Methods (A1)

For assessing ME, I am employing the approach of V&C, which is modified from Arnqvist & Martensson (1998), Klingenberg *et al.* (2002) and, although univariate and for traditional morphometrics, Yezerinac *et al.* (1992). ME can be random or systematic (i.e., a directional bias) (Arnqvist & Martensson 1998), and both types of error can potentially vary in magnitude across groups. An accurate assessment of ME is complex and should take into account all steps that produced the measurements (Arnqvist & Martensson 1998; Fruciano 2016). The basic idea, in the approach of V&C and others, is to scale ME ('noise') in relation to the magnitude of the biological variation ('signal') and verify if the signal is much larger than the noise. If that happens, assuming one has carefully considered all potential sources of error, both random and systematic errors are unlikely to impact results. Thus, data precision (which I refer to, in the articles, as 'replicability') will be adequate to the study aim. However, I stress that one can be confident about this conclusion, only if ME is very small and all contributions to ME are considered, which is often difficult (Arnqvist & Martensson 1998). The expectation that total ME accounts for a small amount of variance in the data at the level specific to the study question is assessed using replicate measurements. The variation among replicates of each individual should be negligible compared to the differences among individuals. An individual is estimated by averaging its replicates. Thus, using the marmot mandible data, with their two replicates, a negligible ME requires that the first and second digitization of a specimen are very close in size or shape space, but clearly separated from pairs of digitizations of other individuals.

Before focusing on ME, I anticipate a consideration on outliers. Outliers in GMM are specimens of relatively unusual size or shape. If numerous and strong, they can alter results, including those of the assessment of ME. This is why, before examining ME, even if data include replicates and the landmark configuration may not be the definitive one, a researcher should at least rapidly check that there are no extreme outliers. Thus, using all 462 specimens and their duplicates, for a total of 924 observations, I briefly screened summary plots of size and shape variation to detect any strongly isolated specimen. Methods are the same that are described in detail in the chapter on outlier detection (A2). Because I spotted none, I proceeded with the analysis of ME, which led to the decision of excluding a few low precision landmarks. Later (A2), using the reduced configuration with only precise landmarks, and having averaged the duplicates of each individual to obtain the data for the taxonomic study, I carefully looked again for outliers in size and shape. This deeper investigation confirmed no extreme outlier, but suggested a few moderate ones. Thus, I went back to the ME analysis and redid it without these potential outliers, as well as both using all landmarks or the final reduced configuration. In all analyses, results were the similar regardless of the outliers, but the magnitude of ME in shape was halved after excluding low precision landmarks (see below).

Identification of low precision landmarks

As in V&C, I assessed ME using an analysis of variance (ANOVA), plus a graphical examination of differences between replicates. These methods are described in the next two subsections. In this study, I added another step, which allows to explore whether any landmark has a much lower precision than others. I call it 'assessment of per-landmark digitization error' or 'per-landmark absolute imprecision'. If this analysis suggests the occurrence of one or a few low precision landmarks, that does not necessarily imply that ME will be too large for a meaningful study of size and shape differences. Yet, it may be useful to identify potentially problematic landmarks, which add dubious information and might be excluded to make the data less 'noisy'.

An accurate analysis or visualization of per-landmark shape (or form) variation is not possible in Procrustean GMM, as in all other types of shape analysis based on superimposition methods. This has been known (Moyers & Bookstein 1979) since before the GMM revolution (Rohlf & Marcus 1993). I am not covering the topic, which is extensively discussed by Cardini & Verderame (2022). It is relevant here only because it means that one cannot assess ME at specific landmarks after the superimposition. If that was done, it might appear that one or the other landmark varies more, but this is inaccurate and potentially misleading (Cardini & Verderame 2022). A researcher can, nevertheless, examine the absolute imprecision of 2D landmarks, if at least two digitizations are made on the same identical image for each individual in the sample or in a representative subsample. Using landmarks on 3D images, the approach is the same: a researcher has to re-digitize multiple times the same 3D image and has to do it for many or all individuals. In contrast, with landmarks directly measured on a real specimen (e.g., a cranium) using a 3D digitizer (e.g., https://gomeasure3d.com/microscribe/), one needs two or more digitizations done without changing the position of the structure (mandible, cranium, a post-cranial bone etc.) between digitizations. With this specific type of 2D or 3D replicates, the analysis of per-landmark absolute imprecision is done before superimposing the data (Cardini & Tongiorgi 2003; von Cramon-Taubadel *et al.* 2007). Results, however, will not include any cause of ME (e.g., variability in specimen preparation, differences in photographs after repositioning a specimen, 2D flattening of 3D structures etc.) other than the digitization error of specimens held in the same exact position. If other sources of ME are considered, they will be used in the overall assessment of ME after the superimposition.

For assessing per-landmark absolute imprecision, because I am only measuring digitization error as a source of ME, the data are, in my case, the same as in all other steps of the ME analysis. Thus, using the raw coordinates (no superimposition!) of the two TPSDig digitizations of the same image of an individual, I computed in a spreadsheet [6] the variance of the X and Y coordinates of L1 and summed them up. Because the image is the same, any difference between the digitizations represents the digitization error of L1 for this individual. The same computations are done for landmark 2, 3 and so on, and the whole procedure is repeated on the second individual in the sample, then the third one, fourth one etc. In my dataset, the result was a matrix of 15 raw landmark variances for each of the 462 individuals in the sample. The variances are the estimate of landmark digitization imprecision in each individual and can be summarized using medians and/or means. I employed both types of averages. Medians and means suggested the same pattern (see Results), but the median is generally less strongly influenced by extreme observations. For instance, few individuals with a very large variance for a specific landmark might lead to an overestimate of its imprecision using the mean, but not with the median. Besides using averages to summarize the main trend in per-landmark imprecision, I computed (also in a spreadsheet) the 90th percentile and the standard deviation (SD) of each of the 15 columns of per-landmark variances. A 90th percentile is a type of trimmed range, where the top 10% highest values (variances, in my case) are excluded. Like the median, the trimmed range mitigates the potential impact of extreme observations. By including in the analysis the 90th percentile and SD of the per-landmark individual variances, I thus captured also the variation in individual estimates of per-landmark imprecision. Finally, with these four statistics, I made a profile plot of the median, mean, SD and 90th percentile of per-landmark variances. A pronounced spike in the profile plot, if found, suggests something unusual on average for a certain landmark (an unusually low precision or a large variability in precision).

The values shown in the profile plot were also reported in a table (see Table 3 in the Results). In this table, I added the ratio obtained by scaling the median variance of each landmark (e.g., 0.16 mm^2 for L13) by the median of the medians of all 15 landmarks (which is 0.04 mm^2). Thus, for example for

[6] To import the data in a spreadsheet, the TPS file with the duplicates can be converted to *.csv in TPSUtil or opened in PAST, saved as *.dat and then imported in a program such as Gnumeric (http://www.gnumeric.org/) or LibreOffice calc (https://www.libreoffice.org/). A similar format conversion can be done in MorphoJ by importing the TPS file as a new dataset and, then, re-exporting the raw coordinates as TXT to be opened in a spreadsheet.

L13, the median ratio showed that the variance due to digitization error was on average four times (0.16 / 0.04 = 4) larger than generally found across the entire configuration of 15 landmarks. I computed this type of ratio to aid the detection of low precision landmarks, but did it only with the medians for simplicity and because, as mentioned, they are less strongly affected by extreme observations. Finally, in the table, I ordered the values according to the median ratios (from the lowest to the highest), so that the most and least precise landmarks are respectively at the top and bottom of the table.

The absolute digitization error can be visualized on the configuration itself. This is easy, but needs some preliminary work in an spreadsheet. For each individual, I calculated the difference of the raw coordinates between the second and first duplicate. Then, I added these differences, that measure imprecision, to the raw coordinates of an individual (anyone can be chosen). Finally, I plotted the data using PAST *Plot, Landmarks*. In this figure (Fig. 2c, in the Results), however, instead of using the raw coordinates of a specific 'real' individual, I employed the sample mean form calculated as the mean of the superimposed coordinates without size standardization. This choice is optional and does not change the scatter of the points around each landmark, which is the only important information to visualize absolute imprecision. In practice, for the visualization, I copied/pasted the matrix calculated in the spreadsheet into PAST, appended the mean form in the last row and made it red (select the row and click on *Edit, Row color/ symbol*); finally, I selected all data[7] and issued the command *Plot, Landmarks* from the windows menu. Showing the mean form (red circles in Figure 2c) is also optional, but it helps to see if the deviations in one or another landmark tend to happen preferentially in a specific direction (e.g., to the left), thus suggesting a systematic error in digitizing that landmark. Instead of the mean form, one can use, as I mentioned, any real specimen, append its raw coordinates in the data window in PAST and select a colour for this row.

ME ANOVA

The examination of per-landmark absolute imprecision is the only analysis performed using raw coordinates. All other analyses require Procrustes superimposed data to calculate size (CS) and shape. The superimposition can be done in any of the programs I am using (MorphoJ, PAST, TPSRelw, etc.). There might be subtle differences in how it is done (Klingenberg 2020), but they have no impact when there is small variation, as it is the case with the vast majority of biological datasets in GMM (Marcus *et al.* 2000).

I will first informally and briefly explain the principle of the ANOVA, and later exemplify how to do it in MorphoJ after superimposing the data. The aim of a simple one-way ANOVA is to verify if differences in group means (i.e., between group variance) are larger than expected based on the amount of variation normally found among individuals within those same groups (i.e., within group variance). Because usually one cannot measure all individuals in a population, the variation in group means, as well as the variation within the groups, is estimated using samples (more on this in the Appendix A on frequentist statistics). By making a set of assumptions (Greenland 2019), the ANOVA calculates the probability (P value) that the sample data are compatible with the null hypothesis of (typically) identical means. If this probability is very small, the researcher can reasonably assume that mean differences are unlikely

[7] Unlike most other user-friendly software for GMM and/or statistics, the data selection in PAST is done by simply selecting the area of the PAST spreadsheet where the data are and then issuing a command using the window menus. Also, most univariate analyses using groups in PAST require to have the data (e.g., CS) of each group (e.g., species) in columns one next to the other. For instance, in the first column, there could be CS of the Alaskan marmots, in the second CS of the Olympic marmots etc. To compare the groups, one will have to select all the relevant columns (all size columns for the different species, in my case). In contrast to univariate data, but similarly to most other statistical programs, data for multivariate analyses in PAST are organized in a matrix where columns are different variables and observations are in rows, with groups typically arranged one after the other (e.g., the first 16 rows might be Alaskan marmots, the next 14 Olympic marmots etc.). For multivariate comparisons, groups are specified in PAST with different colours, for the selected rows, using the menu command: *Edit, Row color/symbol*.

to have occurred by chance simply because of sampling error (i.e., because of not having measured the entire populations). The probability P depends on how large mean differences are compared to within group variation, but also on the size of the samples, as one has more confidence in the accuracy of estimates if samples are larger rather than smaller. Thus, for instance, an ANOVA can be used to statistically test mean interspecific differences in size or shape. If the resulting P value is very small, differences in species means will be considered *statistically* significant. However, as better discussed in the Appendix, the taxonomist must bear in mind that *statistical* significance, even if strongly supported by a test, is not the same as *biological* significance. For instance, the mean difference could be small and biologically meaningless (e.g., a 1 cm mean difference in adult body height between players of two basketball teams, even if present, has probably, on its own, no impact on performance) or plastic in nature and simply due to different environmental conditions in geographically distinct populations (e.g., nutrient deficiency might stunt growth and, thus, reduce average body size in a population compared to a genetically similar population from a different region with higher food quality, which can be interesting but probably taxonomically uninformative).

Factors, in an ANOVA, can have more than two groups (my six species, for instance) and ANOVAs can use more than one factor (two-way or three-way or, more generally, multi-factorial ANOVAs). With two or more factors, ANOVAs can test not only each factor, but also their interaction, which measures if the effect of one factor is similar across all levels of another one (e.g., in a sexually dimorphic genus with larger males, the interaction might be used to assess whether male size is similarly larger in all species of the genus). The interaction, however, is not considered in the ME ANOVA (see below) and it is, thus, introduced with some more detail later in part B, where it is tested for species and sex. ANOVAs can also be univariate or multivariate (MANOVA). Despite the name, MorphoJ's Procrustes ANOVA for shape is, in fact, a MANOVA. As the number of factors increases, the design of the analysis (be it univariate or multivariate) becomes more complicated and, when sample size varies across groups (unbalanced design), there is a further layer of complexity, including different ways of computing variances and potentially more serious problems with the effect of sampling error. These fairly advanced statistical topics are covered in the ANOVA sections of most statistical textbooks. Among the many written for a readership lacking strong bases in mathematics, I suggest the manuals by Howell (2013) or Moore & MacCabe (2005) for univariate analyses, and the classic introductory book by Hair and colleagues for multivariate methods (Hair *et al.* 2013). Another popular textbook for biologists, although slightly more technical, is 'Biometry' (Sokal & Rohlf 2011), which has a shorter version for those looking for a simpler introduction (Sokal & Rohlf 2009). Old but still very useful is also '*A Survey of Multivariate Methods for Systematics'* (Neff & Marcus 1980), recently made freely available online (https://digitallibrary.amnh.org/handle/2246/6961).

The ANOVA I will be using for the assessment of ME is a three-way analysis testing species, sex and individual variation in the dataset with the duplicates. To simplify the model, following MorphoJ's approach (Klingenberg *et al.* 2002; Klingenberg 2011), interactions among factors are assumed to be negligible[8]. In a study on taxonomic differences, as I argue later in the methods and discussion, the most important level for assessing ME is usually the comparison of inter-individual differences (with individuals represented by the average of the duplicates) to differences between the first and second landmark digitization used to estimate ME. If instead of just two duplicated digitizations, one has more replicates and other sources of ME, the method works the same way. Thus, as outlined at the beginning of this chapter, individual variation should be found to be statistically significant and much larger than the variation due to ME. For testing the significance of individual variation compared to ME, two ANOVAs

[8] As pointed out by a referee, this assumption is generally reasonable. Nonetheless, he/she rightly pointed out, it is unlikely but not impossible that there could be, for example, different measurement problems on the two sexes in different species (e.g., larger ME in females compared to males in one species and the other way round in another species), which would make the species by sex interaction relevant.

are run in parallel in MorphoJ: one for CS and another one, multivariate, for shape. On my data, I run this pair of ME ANOVA multiple times on slightly different data to test the sensitivity of results to the exclusion of low precision landmarks and outliers. Re-testing the same hypothesis many times increases the chances of rejecting it by mistake. This is known as inflation of the type I error rate. However, the issue is secondary here, because I will be focusing more on the relative size of the effects (the magnitude of individual compared to ME variation) rather than on P values. Type I error, the interpretation of P values, and the importance of effect size, are also discussed in the Appendix A on frequentist statistics.

To perform the ANOVA, all 924 observations with their 30 raw coordinates are imported as a 2D dataset in a new MorphoJ project. MorphoJ can directly import TPS files from TPSDig. Alternatively, MorphoJ has its own *.txt ascii format, where data are organized like in a spreadsheet, with observations in rows and variables in columns. Data are separated by commas or 'tab stops'. The first line contains the name of the variables; the first column (whose name must be ID) is a unique specimen identifier; columns after the first are the raw coordinates, for which any column name is acceptable as long as it is a single word with no blanks. To convert TPS to TXT, TPS can be opened in PAST, and then saved as *.txt. When PAST opens a TPS file, the software automatically converts pixels (the units in which raw coordinates are saved in TPSDig) into mm (or other metric units[9]) using the scale factor the researcher has specified in TPSDig. A similar operation can be done in TPSUtil (*Convert tps/nts coordinates file*) using the CSV format, which can be opened in any spreadsheet or text editor. After the appropriate editing, files can be loaded in MorphoJ (*File, Create New Project* or, if part of a bigger project, *File, Create New Dataset*). For the marmot mandible landmark data, using a text editor, I changed the name of the first column to ID and carefully replaced the default identifier created by TPSDig (an integer) with a more descriptive name. V&C provides suggestions on how to name observations, so that they can be more usefully employed in MorphoJ.

The ID column is very important, as it is typically used to match size and shape data with other variables in MorphoJ. In this study, after loading the raw coordinates, I imported *classifiers* (as groups are called in MorphoJ), of which the main ones are species, sex and individual. I also imported some *covariates*, which are any type of continuous numerical variables, but, as in my case, can also be dummy variables coding for groups. A dummy variable can, for instance, binary code sex using 0 for females and 1 for males (or *vice versa*), and can be used to test SDM with a regression (see part B). *Classifiers* and *covariates* are placed in different *.txt files, but they all have the first ID column with the same identifiers as in the raw coordinate file. Once all relevant data are loaded in MorphoJ, a *Procrustes fit* is done from the menu *Preliminaries* after selecting (click on it) the dataset. For the visualization of shape in the same software, I created a wireframe (*Preliminaries, Create or edit wireframe*) by connecting pairs of landmarks with lines that suggest a stylized marmot mandible. The wireframe is optional and there are other possibilities to visualize shape change in MorphoJ or other programs (Klingenberg 2013).

With the data Procrustes superimposed, the ME ANOVA in MorphoJ is performed from the menu *Variation, Procrustes ANOVA*. This opens a window to specify the design. I selected, in the *individual* box, the classifier (called 'indiv' in my dataset) that uniquely identifies each specimen, so that the software knows that the identical labels indicate duplicates. Then, I added species (first) and sex (second) as *additional main effects* (box to the right, in the same window). Main effects are used when the researcher wants to control for additional levels of variation before testing that differences among individuals are larger than ME. By subtracting species and sex mean differences in size and shape, one makes the test for the significance of individual variation compared to ME more conservative (i.e., it is harder to 'claim' that ME is negligible). To put it simply, it is 'as if' the researcher was running the ANOVA using, for instance, only females of the same species (no variability due to SDM or interspecific variation).

[9] Of course, the unit of measure must be the same (e.g., always mm) in all the photographs!

In principle, one could avoid controlling for main factors by running multiple ANOVAs within each species with separate sexes, but that requires many more tests in smaller samples. More importantly, one could argue that, because the level of the analysis I am most interested in is interspecific, the main effect of species should not be removed before examining if individual variation is larger than ME. However, without controlling for species, I cannot be sure that, if I need to run an intraspecific test (for instance, for SDM or allometry), ME is negligible also at this lower taxonomic level. On the other hand, if ME is negligible compared to individual variation regardless of species and sex differences (conservative approach), that implies that ME should be also negligible when I compare species (the most interesting comparison in this taxonomic study). In the Discussion, I will say more on this issue, as well as on the importance of biases (i.e., systematic errors) in relation to ME.

MorphoJ's ANOVA tests all effects first for CS and then, in a separate ANOVA, for shape. The software provides the numerical output in the *Results* window. In the ANOVA, both between species variation and sex differences are tested against individual variation, from which species and sex mean differences have been removed; individual variation, in contrast, is tested against the residual variance, which is the differences between duplicates used to estimate ME. The test of individual variation compared to ME is the main focus, for now, in this preliminary analysis, whereas the tests of species and sex differences will be considered in part B. The ANOVA in MorphoJ is hierarchical and, therefore, the order in which factors are entered in the model is important: specifying species first and sex after is not the same as with sex first and species second [10].

The ANOVA tests in MorphoJ are parametric. Parametric tests assume that data follow a specific distribution (for instance, a normal distribution) characterized by specific parameters (e.g., the mean and variance) and, under this assumption, estimate the P values of the observed test statistic. For CS, the test statistic is the conventional univariate F ratio, explained in any introductory textbook on statistics. For shape, however, there can be two types of multivariate tests (Goodall's F and Pillai's trace). They are based on different assumptions, as explained in detail in Klingenberg *et al.* (2002) and more briefly in the help manual of MorphoJ. Pillai's trace relies on a model which is more realistic, because it does not require isotropic variation, which implies similar uncorrelated circular variation at each landmark (see the power analysis in A3, for more on this). However, Pillai's trace can be computed only when sample size is adequately large in relation to the number of variables: if one does not find Pillai in the *Results*, that indicates that N was too small in relation to number of variables to compute this test statistic. MorphoJ provides P values for all tests, but does not compute the proportion of variance accounted for by each factor in the model, a measure called R square (Rsq). The Rsq is easily calculated in a spreadsheet, after copying and pasting the output of the *Results* window. The output may have to be slightly edited before copying and pasting it (for instance, removing blanks in the column and row names). With data correctly imported in a spreadsheet, the computation is straightforward and works in the same way for CS or shape: sum the numbers in the SS (sum of squares) column to obtain the total SS; take the SS of each effect (species, sex, individual and residuals, which represent ME in this case) and divide it by the total SS. The result is a series of Rsq, which can be multiplied by 100 to be expressed as a percentage of the total variance in CS or shape accounted for by a factor. The expectation for a negligible ME is that the individual Rsq is much larger than the residual Rsq.

[10] Based on my knowledge of marmot biology (see Introduction), I expect species differences to be fairly large and SDM small. This is why I controlled first for the factor (species) for which I expect a larger impact on the comparisons of the factors specified after it (i.e., sex and then individual). With a different taxon, I might have taken a different choice. For instance, with guenon monkeys, I would have run the ME ANOVA separately in females and males, with only species as a main additional effect, because in this group SDM is very large. Besides, SDM in guenons varies in magnitude depending on the species (Cardini & Elton 2008b), which makes it more problematic to accurately control for it using a statistical correction, as in the ANOVA. It is also worth emphasizing that, because the ANOVA tests the main effects against individual variation (for sex, after removing species mean differences), the analysis already provides information on interspecific differences and SDM. However, as mentioned, I generally prefer to use the ANOVA in MorphoJ mainly to assess ME. Later, after having demonstrated that ME is small, I focus on group differences using more specific analyses, which include the potential interaction of species and sex (see part B).

Graphical assessment of replicability

Numerical analyses are typically complemented with a graphical examination of the data. For ME, there are several types of plots that can be used in parallel with the ANOVA. For CS, I plotted the first duplicate versus the second one[11]. The points in this 'CS on CS' scatterplot should be on a diagonal line passing through the origin with a slope of 45°, if replicability is very high and, thus, CS values are virtually identical in the duplicates. With more replicates, one needs to do pairwise plots (e.g., first vs second replicate, first vs third, second vs third etc.). To make the plot, I used PAST *Plot, Graph*, after selecting two columns, one with the CS of the first duplicate and the other with the corresponding CS of the second duplicate. There is a shortcut to rapidly obtain the plot in MorphoJ. One has to split the dataset using a classifier that indicates the first and second duplicate (*Preliminaries, Subdivide dataset by*), link the two datasets (*Preliminaries, Link datasets*) and then run a PLS analysis using CS in both datasets (*Covariation, Partial Least Squares*). The PLS analysis in itself is meaningless in this specific case, and P values should not be computed. It is only used as an expedient to obtain the scatterplot between the CS of the duplicates. If data are further split by group (species and/or sex, in my case) using one or more MorphoJ's classifiers, the 'CS on CS' scatterplot can be repeated within more homogeneous subsamples. This allows a more detailed inspection of the correspondence of CS measurements in the duplicates, but it is time consuming and generally unnecessary unless the 'CS on CS' scatterplot of the entire sample (and/or the ME ANOVA) suggests an evident low precision, which is uncommon for CS.

Replicability in shape data can be graphically explored with ordinations (i.e., methods to draw scatterplots of multivariate data) and cluster analyses. In both approaches, the closer two observations are in the plots, the more similar they are, which for duplicates means that they should form tight pairs if ME is negligible. The simplest ordination is a principal component analysis (PCA). This technique is clearly explained in Neff & Marcus (1980). A PCA produces new variables (PCs) that summarize total sample variance. To this aim, the original variables, which here are the Procrustes shape coordinates, are linearly combined, so that PCs (the linear combinations of the initial variables) maximize variance along statistically uncorrelated directions. Thus, for instance, PC1 has a zero correlation with PC2 and the same holds for any pair of PCs. PCs are useful for dimensionality reduction, as well as for bivariate scatterplots, where distances among observations are proportional to their differences. This means that one might be able to explore the similarity relationships (i.e., who is most similar to whom) among the observations using just a few of the first dimensions, as long as together they account for most of the total variance (e.g., pairwise scatterplots of the first three or four PCs etc.). When a PCA is computed using variance-covariance matrix of the original data, it leaves the distances among the observations in the full data-space of a sample unaltered. Thus, the multivariate distances between any two observations are identical regardless of whether they are calculated using Procrustes shape coordinates or their PCs.

As for the plots of CS, I used PAST also for the PCA. First, I selected the Procrustes shape coordinates imported from MorphoJ (see below) and, then, clicked on *Multivar, Principal components*, with the options *var-covar* and *disregard groups*. The same analysis is easy to do also in MorphoJ and TPSRelw (where PCs are called 'relative warps'), but only in PAST and MorphoJ groups can be visualized using different symbols or colours. In PAST, it is also possible to emphasize groups using *convex hulls* (check the box option), which enclose the observations of a group by drawing a polygon connecting the observations along the boundary of the group scatter. As anticipated, with a small ME and, thus, a high replicability, duplicates in the PCA scatterplots should be in pairs corresponding to each individual, well separated from other individuals. However, to accurately check that duplicates cluster together 'within individuals', a phenogram (next paragraph) is better than a PCA scatterplot, because small distances

[11] Alternatively, one can draw profile plots, with CS on the vertical axis and replicate number (i.e., first or second duplicate in my case) on the horizontal axis. If CS is very precise, lines should be horizontal and parallel; with a systematic error (e.g., if the second replicate tends to underestimate CS), lines will be parallel but diagonal. Profile plots, however, will be difficult to interpret, if the sample is large, because lines overlap.

are better preserved and a single plot is used instead of a series of pairwise scatterplots of several of the first PCs. PCAs, in contrast, can be more helpful to detect systematic ME. With systematic differences between duplicates (i.e., a bias), one expects a degree of separation between the first and second duplicate (plotted as groups using different colours) along one or more PCs. For instance, on PC1, one might find that most individuals in the second duplicate are consistently shifted to the left compared to individuals in the first duplicate (or vice versa). As with the 'CS on CS' scatterplots, PCAs can be done using the total sample, regardless of groups, but also repeated within taxa (and/or sex) to increase the resolution and focus on more homogeneous data.

A hierarchical cluster analysis produces a graphical summary of multivariate data in the form of a tree (Hair *et al.* 2013). It is based on the distances between all possible pairs of observations in a sample. These distances measure the differences among individuals: the larger the distance, the bigger the differences. The tree is called dendrogram or phenogram, and it is built so that the closer the observations are in the tree and the shorter the branches, the higher the similarity between them. Thus, observations that are very similar should group together in the same branch, whose length will be long if the members of the cluster differ markedly from all others. Phenograms should not be confused with cladograms (Felsenstein 2004): the former are purely concerned with similarity (typically, an inaccurate proxy for evolutionary relationships), while the latter is aimed at inferring evolutionary relationships (i.e., phylogeny). Phenograms are built so that the tips of the tree ('leaves') are all equidistant from its root (a property called ultrametricity). Inevitably, phenograms introduce a degree of distortion of the original distances. The amount of distortion can be estimated by the cophenetic index (Sokal & Rohlf 1962; Legendre & Legendre 2012).

There are many types of algorithms to compute a phenogram and one can select different types of distances. An appropriate choice in Procrustean GMM is the Euclidean distance, that corresponds to the length of a straight line between two observations in the multivariate shape space. This distance is generally an excellent approximation of the Procrustes shape distance in the curved space produced by the superimposition (see Introduction). The approximation between the two types of distances can be verified using TPSSmall (Rohlf 2015). V&C provide more information on this methodological aspect, that generally has a negligible impact on biological data (Marcus *et al.* 2000). In terms of the choice of the algorithm, there are just three options in PAST, which is the only software for cluster analyses among the user-friendly programs commonly used in GMM. The unweighted pair group method with arithmetic mean (UPGMA) generally performs well (high cophenetic correlation and, thus, low distortion) compared to other methods [12] (Rohlf 1970). To produce the phenogram in PAST, the following commands are used: *Multivar, Cluster analysis, paired group* option using Euclidean distances. Users must be sure to have imported superimposed data or should do the Procrustes superimposition in PAST. Since PAST uses a slightly different superimposition and requires an additional step to project the shape data in an Euclidean space, I prefer to export the Procrustes shape coordinates from MorphoJ. There is, however, a caveat when TXT files from MorphoJ are opened in PAST: the format is almost perfectly compatible, but PAST does not accept ID as the name of the first column; therefore, ID must be replaced in a text editor with a different single word without blanks (e.g., label or identifier or even just a dot).

The expectation for a negligible ME is that the shape distance between duplicates of the same individual is smaller than the distance of any of them from duplicates of other individuals. This is something that could be directly checked in a matrix of pairwise Procrustes distances, but it is tedious unless done by a macro or by coding in R. In a phenogram, however, one should find, for each individual, duplicates paired in the same cluster as nearest neighbours, to the exclusion of any other individual. For brevity, I refer to this condition as 'sister duplicates' ('sister replicates', if there is more than two) or as clustering

[12] Users may explore other algorithms and, for instance, single-linkage (also called nearest neighbour) might be interesting for outlier detection despite a likely lower cophenetic correlation.

'within individual'. Also, I call 'shape replicability phenogram' the tree with the duplicates (or, more generally, replicates). In this tree, a researcher can count the proportion of individuals with sister duplicates as an estimate of replicability. Because with large samples, and/or more than two replicates, trees become very big, it may be easier to inspect the phenogram in a tree viewer such as Dendroscope (Huson & Scornavacca 2012), that allows to zoom in and out. From PAST, the tree can be saved as nexus file (*.nex) file. NEX is a general file format for trees, but it has slightly different versions depending on the software and they are not always compatible. If the NEX file does not open in a tree viewer, I suggest to open it in a text editor, remove everything except the tree notation in parentheses and its final semicolon, and rename the file extension as *.tree. TREE is a simpler format, also called Newick, which is easily imported in almost all tree viewers.

A shape replicability phenogram is simple and intuitive when, as in my study, there are only duplicates. With more than two replicates, the rationale is the same, but the summary of the information the tree provides on replicability is more articulated. For instance with three replicates, they may be all clustering together 'within individual' or it could be just two of them, with the third replicate being closer to those of other individuals. The researcher has, thus, to make a distinction when counting individuals with replicates in the same cluster and, later, summarize results using two percentages: one for individuals clustering as sisters in triplets (highest precision) and one for those in sister pairs (intermediate precision). For examples of phenograms and summary tables with several replicates per individual, I refer readers to some of our previous papers (Daboul *et al.* 2018; Galimberti *et al.* 2019).

Results (A1)

The results of the ME analyses are in Tables 3–5 and Figures 2–4, following the same order as in the methods: first, the identification of low precision landmarks; second, the ME ANOVA; third, the 'replicability plots'.

The profile plot of Figure 2b indicates that L11 has a much lower precision than any other landmark. Figure 2c suggests that this is because of large uncertainties in its position along the curvature of the mandibular angle. Table 3 confirms that L11 has the lowest precision (~ 8 times the average) and shows that L12 and L13 also have unusually large errors (~ 4 times the average). These three low precision landmarks also display a larger range of differences between the first and second digitization, with their 90th percentiles being almost three to 37 times larger than average. Using arithmetic means and SD, instead of medians and 90th percentiles, does not change the conclusions; in fact, the lower precision of L11-12-13 is even more evident (Fig. 2b). Thus, ME ANOVAs, as well as the cluster analysis of shape duplicates, will be first performed including all 15 landmarks and later repeated without L11-12-13, to assess if their exclusion has an appreciable effect. The 'reduced' 12 landmarks configuration is shown in Figure 3a.

For size (Table 4), the ME ANOVA indicates that, after controlling for the effect of species and sex, individual variation (Rsq = 36%) is hundreds of times larger than digitization error (Rsq ≤ 0.1%) with very little differences in results after excluding low precision landmarks. With shape (Table 5), once variation due to species and sex is removed, individual differences (Rsq = 76%) account for 16 times more variance than digitization error (Rsq = 5%). Excluding L11-12-13 has little effect on individual shape variation (Rsq = 75%), but halves digitization error (Rsq = 2%), so that individual differences become 32 times larger than ME. If the ME ANOVAs of size and shape are repeated after excluding potential outliers (next subsection, A2), results are virtually identical (Tables 4–5), with only a tiny reduction in the Rsq of the individuals.

The scatterplot of CS in the first and second duplicates (Fig. 3b) shows observations which are almost perfectly on a 45° line passing virtually through the origin of the axes (intercept = 0.3 mm). The

Table 3. Per-landmark raw coordinates variance due to digitization error. In this and other tables, the most relevant results for the Discussion are emphasized with a light grey background: for instance, in this table, landmarks with the largest absolute imprecision are shown with a grey background.

Landmark	Times median of medians	Median (mm^2)	Average (mm^2)	90th percentile (mm^2)	SD (mm^2)
L5	0.4	0.02	0.05	0.06	0.20
L14	0.5	0.02	0.03	0.08	0.06
L1	0.8	0.03	0.11	0.25	0.32
L7	0.8	0.03	0.05	0.10	0.19
L6	0.8	0.03	0.08	0.17	0.18
L3	0.9	0.03	0.07	0.15	0.11
L4	1.0	0.04	0.07	0.17	0.17
L2	1.0	0.04	0.12	0.26	0.30
L8	1.0	0.04	0.06	0.13	0.07
L9	1.1	0.04	0.08	0.16	0.14
L10	1.5	0.06	0.11	0.23	0.14
L15	1.5	0.06	0.13	0.25	0.26
L13	4.0	0.16	0.31	0.68	0.56
L12	4.2	0.17	0.42	1.26	0.64
L11	7.8	0.31	1.10	2.23	3.32

correlation between CS in the first and second duplicates is 0.999. Summary statistics for CS in one or the other duplicate are also virtually identical: the CS range is 56.1-92.8 mm in the first duplicate and 56.0-92.5 mm in the second, and means ± SD are respectively 77.2 ± 8.0 mm (first duplicate) and 77.0 ± 8.0 mm (second duplicate).

The shape PCA scatterplots suggest an almost perfect overlap of the two duplicate datasets. In Figure 4a, the PC1-PC2 scatterplot is shown for the 12 landmarks configuration, but the plot is very similar using all 15 landmarks (not shown). Scatterplots of PC3-PC4 and PC5-PC6 for both configurations (not shown) confirm the overlapping distribution of the duplicates. That this overlap depends on the close similarity of the duplicates of each individual is confirmed by the UPGMA phenogram, where the more than 90% of duplicates form sister pairs (see Figure 4b for an example). However, as in the ME ANOVA, the reduction in ME after leaving out the three low precision landmarks is appreciable in the phenogram too: with all individuals included, the percentage of sister duplicates is 92% for the total 15 landmarks configuration, but becomes 98% using the reduced set of 12 landmarks.

Discussion (A1)

General considerations: sources of errors, standardizing photographs and flattening in 2D photographs

The assessment of ME is a preliminary analysis, but it is fundamental. If ME is large and/or biases the group comparisons, analytical results may be invalid. Minimizing ME also makes statistical tests more powerful (Arnqvist & Martensson 1998). Yet, many, if not most, GMM papers do not report anything about ME. This is easy to check by selecting a sample of GMM papers from the literature. For instance,

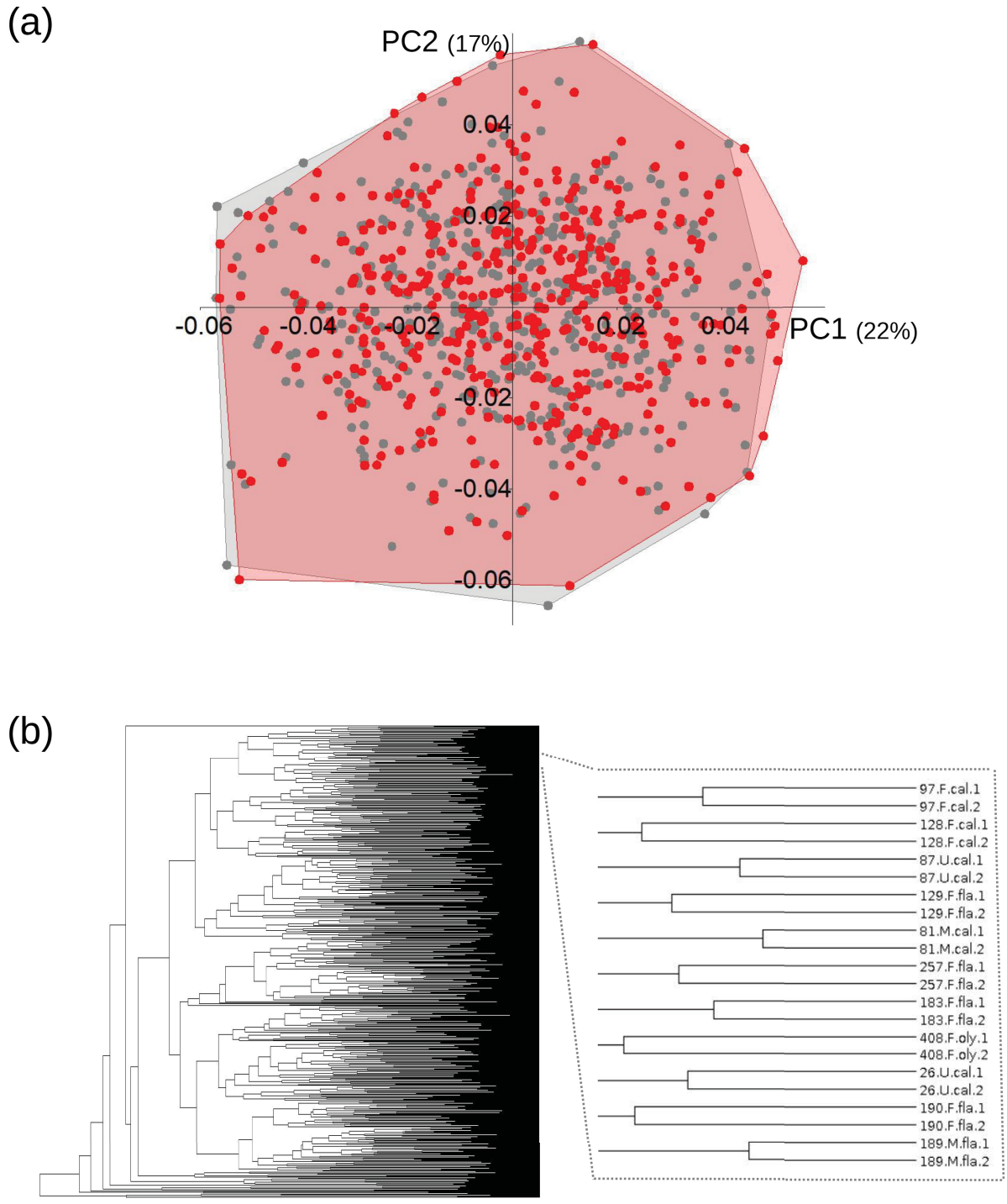


Fig. 4. Graphical examination of replicability in shape using the reduced 12 landmarks configuration. **a**. PCA scatterplot (in parentheses the variance accounted for by each PC) with convex hulls for the first (grey) and second (red) duplicate. **b**. Example of phenogram used to count 'sister duplicates' in the whole sample (the inset zooms in the phenogram to exemplify how duplicates 1 and 2 of each individual, e.g., number 97, a female, or number 81, a male, etc., should cluster in pairs, if ME is smaller than inter-individual differences).

Table 4. ANOVA assessing ME in CS. Abbreviations in this and other tables: SS = sum of squared deviations from the factor group mean; df = degrees of freedom; MS = mean sum of squares, i.e., SS / df; significant P values (P < 0.005) are emphasized in italic. As in Table 3, the most relevant results for the Discussion are shown with a light grey background.

Configuration	Factor	ALL indiv.						No. outliers	
		SS	MS	df	F	P(F)	Rsq	P(F)	Rsq
15L	species	36748.7	7349.7	5	127.7	*<0.0001*	63.0%	*<0.0001*	65.0%
	sex	760.2	760.2	1	13.2	*0.0003*	1.3%	*0.0005*	1.2%
	individual	20727.5	57.6	360	413.4	*<0.0001*	35.6%	*<0.0001*	33.8%
	residual	51.1	0.1	367			0.1%		0.1%
12L	species	30037.7	6007.5	5	125.9	*<0.0001*	62.7%	*<0.0001*	63.7%
	sex	663.6	663.6	1	13.9	*0.0002*	1.4%	*0.0001*	1.5%
	individual	17183.8	47.7	360	759.0	*<0.0001*	35.9%	*<0.0001*	34.7%
	residual	23.1	0.1	367			0.0%		0.0%

searching in google scholar "geometric morphometrics" and "open access" (to quickly find full html papers), only four, out of the first 20 papers I found, reported at least some kind of assessment of ME.

ME, in a broader sense, does not concern only size and shape variables. Researchers must carefully check the accuracy of a priori groups (taxonomic affiliation, sex identification etc.) and covariates (e.g., body mass and other types of traditional morphometric measurements, environmental or genetic variables, geographic coordinates etc.), as well as the presence of individuals unintentionally mislabelled or duplicated. More generally, precision and/or accuracy of any type of data incorporated in a GMM analysis should be examined. This basic necessity runs counter the pressure for obtaining results and publishing rapidly, as well as a variety of other time constraints, but a balanced compromise is feasible.

In this study, I only considered digitization error on photographs. The identification of low precision landmarks is specific to this type of error. All other analyses, in contrast, can be used to assess other sources of ME following the same basic design as for digitizing error. There might be ME because, for instance, of differences in the preparation of the specimens or in the device used to collect the data, the variability in relation to the position of a specimen when photographed or digitized, the error of 2D flattening in 3D structures, and the almost inevitable inter-operator differences, when data are collected by different researchers. In 2D photographs it is clearly fundamental that the orientation of the specimen in relation to the camera is always the same, as slight rotations in one photograph relative to another one are likely to distort and bias the landmark coordinates. One might then wonder whether there is, for instance, an optimal way to position crania in vertebrates. I doubt there is a general rule, like the Frankfurt plane often used in anthropology. The 'best' position varies with the technique for data collection (2D photographs, 3D landmarks measured with a digitizer, photogrammetry or 3D scanners etc.), as well as with the study aim and structure. What specific landmarks one is going to digitize is, of course, also crucial to decide how to collect the data. It is also possible that data on a specific structure might have to be acquired in multiple steps, as it happens sometimes with crania digitized using a Microscribe (https://gomeasure3d.com/microscribe/) because one side or some of its landmarks are not accessible unless the cranium is reoriented. Thus, the operator might first measure, say, the ventral view and later digitize the dorsal side. This is not an issue as long as there are at least three landmarks in common for registering the two datasets by matching the common landmarks as in Frost *et al.* (2003) and Cardini & Elton (2008a). Nonetheless, even if there is no 'optimal orientation' for

Table 5. MANOVA assessing ME in shape. As in Tables 3–4, the most relevant results for the Discussion are emphasized with a light grey background.

Configuration	Factor	ALL indiv.								No. outliers		
		SS	MS	df	F	P(F)	Pillai	P(Pillai)	Rsq	P(F)	P(Pillai)	Rsq
15L	species	0.5048	0.0039	130	17.6	*<0.0001*	2.7	*<0.0001*	18.6%	*<0.0001*	*<0.0001*	20.0%
	sex	0.0159	0.0006	26	2.8	*<0.0001*	0.2	*<0.0001*	0.6%	*<0.0001*	*<0.0001*	0.9%
	individual	2.0680	0.0002	9360	16.5	*<0.0001*	23.7	*<0.0001*	76.1%	*<0.0001*	*<0.0001*	74.4%
	residual	0.1277	0.0000	9542					4.7%			4.7%
12L	species	0.3762	0.0038	100	21.1	*<0.0001*	2.3	*<0.0001*	21.9%	*<0.0001*	*<0.0001*	22.2%
	sex	0.0167	0.0008	20	4.7	*<0.0001*	0.2	*<0.0001*	1.0%	*<0.0001*	*<0.0001*	1.0%
	individual	1.2836	0.0002	7200	33.7	*<0.0001*	18.5	*<0.0001*	74.8%	*<0.0001*	*<0.0001*	74.5%
	residual	0.0389	0.0000	7340					2.3%			2.3%

all uses and purposes, especially when data are acquired by multiple researchers, it is important to standardize the protocol for their collection and consistently use the same position of the structure to minimize biases. A careful description of the protocol, including the position of the specimens, will also help to improve reproducibility and facilitate future studies that might merge different datasets. Merging datasets improves sampling, but it is a risky procedure unless the morphometrician can convincingly demonstrate that differences between datasets are truly negligible and do not introduce errors that might bias results and make them inaccurate and misleading (Fruciano 2016).

Carefully assessing ME using replicates of all relevant steps in the data collection is rarely possible in the entire analysis sample, if the sample is large. However, ME analyses can be done in a subsample, as long as this is representative of the biological variation in the study. In this discussion, I provide a few examples of common sources of ME. There are many others. For a deeper discussion on types of ME and the methods to assess them, I refer the readers to the main reviews in the field of GMM (Arnqvist & Martensson 1998; Fruciano 2016).

Repositioning a specimen before a scan, photograph or the direct digitization of landmarks on a structure can increase ME. The increase is probably negligible for structures such as oak leaves, which are flat and easily placed in the same precise position (Viscosi & Cardini 2011), but can be about as large as digitization error for 3D structures like mammal crania and hemi-mandibles (Cardini, personal observation). With structures with object symmetry, if both sides are landmarked, differences in orientation of a specimen not only will increase variance but may even introduce spurious directional asymmetry (Hulme-Beaman, personal observation, and Gharaibeh 2005). This can happen, if, for instance, most individuals are consistently rotated in the same direction relative to the camera or operator. This type of systematic orientation error can easily happen with photographs of ventral views of mammalian crania, if the cranial vault is not fairly flat and the operator does not check (for example, using a spirit level on the palate) that the specimen lies roughly parallel to the lens of a camera vertically positioned above the cranium on a tripod or copy stand. In fact, a degree of spurious directional asymmetry can originate even when landmarks are digitized in 3D using a Microscribe digitizer, because a right-handed (or left-handed) operator might tend to slightly, but consistently, misplace landmarks. If the bias is similar in magnitude and direction in all individuals, results should be accurate despite the error. However, if the bias varies or the researcher is studying asymmetry, results will be inaccurate and potentially misleading.

With live animals, especially in the wild, standardizing the relative position of the camera and the subject being photographed is especially difficult (see, for instance, the discussions in Galimberti *et al.* (2019) and O'Connell-Rodwell *et al.* (2022)). Even focusing on a specific organ or structure, there might be differences in photographs of a specific individual in relation to its behaviour. Changes in the position of an animal might be accompanied by different patterns of muscle contraction, which in turn change the shape of soft tissues. Thus, ME due to variability in position can be particularly large in field studies of wild animals. Nevertheless, the effect of both position and landmark digitization is fairly easy to estimate. For example, in 2D data, one needs at least two images taken at different times for each specimen and two or more landmark digitizations on each image. With more replicates (multiple images and digitizations), the assessment of ME will be more accurate. Averaging many replicates can also increase precision in the analysis sample. Galimberti *et al.* (2019) provide an example of ME analysis using photographs of the head in elephant seal males studied in the field. In their research, photographing and landmarking were likely to be the main sources of ME, because there was minimal variation in the device (same camera, although with a variable zoom); all data were collected by the same operator; and 2D flattening was minimal, as landmarks were digitized on the nasal midplane outline. As expected in live animals, errors caused by the difficulty in standardizing photographs were so large that they accounted for up to almost five times the amount of variance due to mere digitization error in repeated measurements of the same photograph.

In 2D studies of 3D structures, the problem of flattening the third dimension is generally overlooked (Cardini 2014). Marmot hemi-mandibles are relatively flat (~ < 1 cm in-depth compared to an average length of ~ 6 cm). They have been shown to be generally appropriate for 2D analyses of fairly small differences (Cardini 2014). The configuration I am using consists mainly of landmarks lying almost on the same plane, with the majority of the anatomical points approximately on the mandible outline in side view. As briefly reviewed by Cardini *et al.* (2022, and references therein), using quasi co-planar landmarks in mammalian taxonomy, 2D data on crania and mandibles often produce results in good agreement with those of the corresponding 3D analyses. Yet, the morphology of highly three-dimensional structures is inevitably distorted to a degree by the flattening of the third dimension in a 2D image and the number of studies comparing 2D and 3D results is still too limited to attempt any generalization. Besides, the quality and settings of the device used to obtain the images can further aggravate the distortion. For instance, a low-quality camera lens, especially when kept close to the study structure, typically introduces barrel-shaped distortions, as the distance increases from the center to the margins of the photograph. I have previously made suggestions on how to minimize some of these problems (Cardini & Tongiorgi 2003; Viscosi & Cardini 2011; Cardini *et al.* 2022). Nonetheless, before embarking in a large 2D data collection on 3D structures, it is strongly advisable to carefully investigate the problem in a small but representative sample. Ideally, this preliminary work should use the complete landmark configuration, but, as a compromise, one might select a subset of anatomical points that covers the depth of a structure in the third dimension (Cardini 2014; Cardini *et al.* 2022). Low-cost 3D data can be obtained relatively easily for a small landmark configuration using linear measurements and the truss method (Cardini & Chiapelli 2020) or 3D photogrammetry (Olsen & Westneat 2015). A researcher might also be able to borrow a 3D scanner or digitizer for a pilot study, as these instruments are increasingly common in major natural history museums.

The design of the data collection of the replicates is crucial to be sure that no main source of ME is missed. In this respect, the effect of time lags in the data collection seems to have received little attention in GMM. As in my example using marmot mandibles, duplicates are usually collected after randomizing the order of the individuals and waiting a few days between the first and second duplicate. This time interval is relatively short, but simulates fairly well differences one might expect when data are collected over serval days in a row. Yet, taxonomists who collect data over a long period of time need to be particularly wary of longer time lags, if they merge older and newer data in the analysis. Even when

all data are collected by the same operator with the same instrument, there is the possibility that, over time, a researcher unintentionally changes, slightly but consistently, where landmarks are digitized. Low precision landmarks will be the easiest to misplace, which, if done systematically, can introduce a time-related bias. Thus, a sample collected in a specific year and month may look different in size or shape from one collected months (or years) later, simply because of systematic error. However, by comparing replicates of the same individuals remeasured after months or years, as appropriate to simulate the time lag between older and newer data, one should be able to detect this type of systematic error. As usual, the individuals used to re-assess ME over time should, at least as a reasonable approximation, be representative of the groups being studied over the entire period. If newer data are collected by different operators or using different instruments, biases are even more likely and the effect of operator and/or instrument must be included in the ME analysis (Fruciano 2016).

Examples of the effect of biases and random error

Sometimes, even without replicates, one might be able to spot biases due to a consistent misplacement of landmarks. Unfortunately, if there are no replicates, one cannot be sure of the causes and cannot try to correct the error. For instance, Figure 5a shows summary scatterplots for 3D cranial shapes of African men. There are 32 different populations, according to geographical origin and ancestry. However, the ordinations suggest that the main separation is between data collected for a first study (in green) and data collected years later for a follow up (in blue). The separation is inconsistent with geography/ancestry. If all 'blue' populations were, say, close relatives from north-west Africa and all green ones were a largely separate lineage from the south of the continent, the separation might make some sense, but this is not the case. In other words, green and blue clusters do not correspond to geographic or genetic groups. 'Green-blue' differences account for a relatively small amount of total variance (4%). Yet, these differences dominate the pattern of shape variation. They are small, but clear on PC2, and suggest an almost perfect separation on DF1, the main axis of group separation using shape in a DA. A DA is an ordination method that tries to maximize the separation among a priori groups in a multivariate space, as better explained in part B. The between group variance accounted for by DF1 (17%) is three or more times larger than those of any other axis (DF2, DF3 etc., each accounting for just 5% to 2% of between group differences). The striking aspect of the almost complete separation between 'green' and 'blue' on DF1 is that the groups, whose shape differences were maximized, are the 32 real ancestry groups (not the 'green' and 'blue' clusters, corresponding to the first and second round of data collection). This suggests that the 'green-blue' pattern of variation is strong enough to distort any real difference among the 32 populations. This is a clear indication of a systematic error caused by unintentional, small, directional changes in how data were collected originally (green) and years later (blue).

What was the cause of this bias? We would need replicates of, at least, a subsample of the green group (the first to be measured) to explore what happened in the second round of data collection (blue cluster), but these are not available as no 'green' individual was re-measured at the time of the 'blue' data collection. Data were acquired by the same operator using a 3D Microscribe digitizer. The digitizer was not the same identical model as the one used in the first set of measurements. However, unless one of the two machines was damaged, the accuracy of the slightly different types of Microscribe is high and similar (~ 0.1-0.3 mm - http://microscribe.ghost3d.com/gt_microscribe.htm).

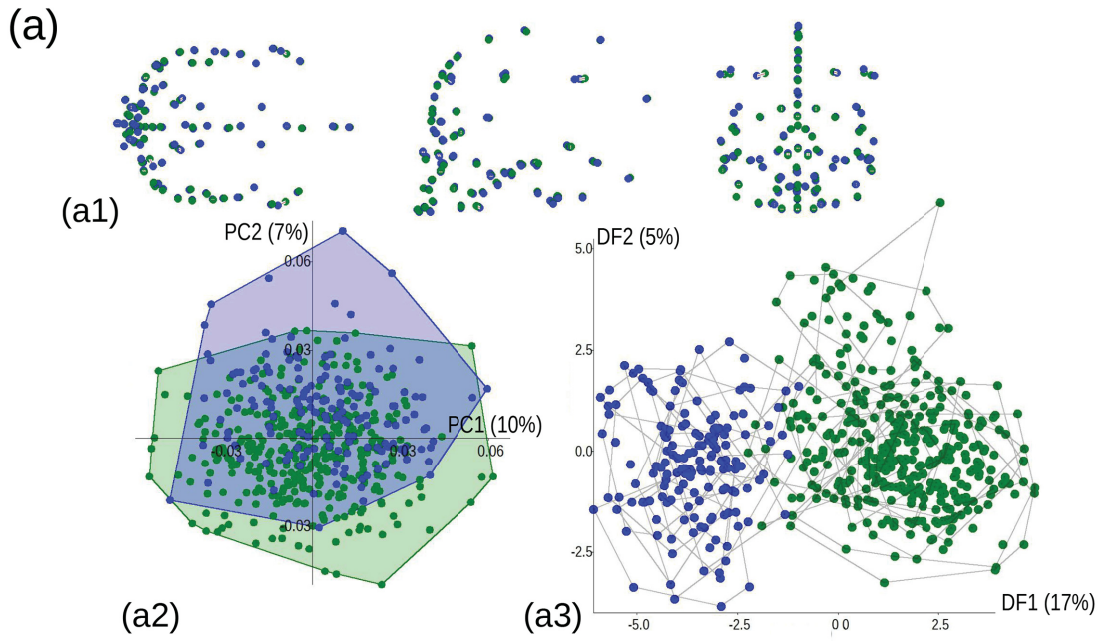
Some of the more than 90 landmarks in the configuration used for this study are harder to locate precisely. Among these, three pairs of bilateral cranial vault landmarks seem to vary remarkably, when green to blue changes in mean shape are visualized. If these six landmarks are removed, the variance accounted for by the 'green-blue' groups drops from 4% to 2%. Likely, these few, low precision, landmarks have been misplaced in a consistent way in the second round of data collection. If correct, this observation suggests another reason to exclude highly imprecise landmarks. Yet, after removing them, DF1 still remains completely dominated by the 'green-blue' differences (not shown). The same happens even

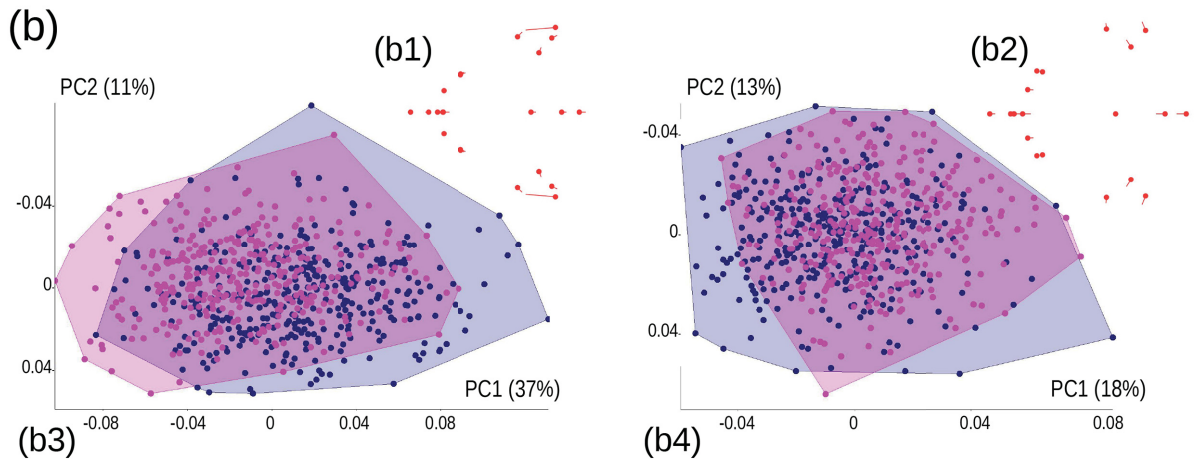
when only a subset with the most precise landmarks, which is just ¼ of the landmarks in the original configuration, is left in the analysis (not shown). Thus, the bias is stronger for some landmarks, but is found even in the most precise subset of anatomical points.

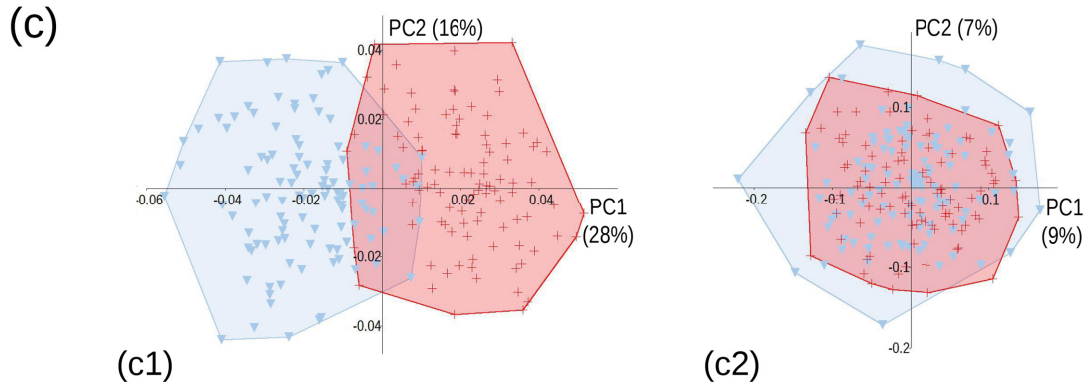
The operator possibly consistently measured the crania in a very slightly different way, but a problem with the Microscribe cannot be excluded and both issues might have occurred. Without a clear understanding of its causes, the likely error cannot be accurately quantified, which also prevents any possibility of potentially correcting the bias and, thus, renders the data in the combined sample unusable for a study. To reduce the risk of this type of problems, the advice is simple: it is always better to reassess ME in a sample, already measured at the beginning of a study, before going on collecting data on new specimens after a relatively long time. By doing this check, one can spot and correct a bias. Sometimes, the operator might need to re-train her/his ability to precisely locate the landmarks in a configuration already used before. In other cases, it might be that a researcher has overlooked part of the original protocol for data collection. A small difference in the settings (light conditions, the orientation of the specimen, a different camera etc.) may be enough to introduce a bias, that creates spurious directional variation and might lead to inaccurate results.

Figure 5b shows another example of likely impact of ME on the results of shape analysis. In this case, because replicates were available, the main source of imprecision could be explored and its effect partly controlled for. Data are craniofacial 3D shapes from a European sample of several hundreds men and women (recently published, using a reduced configuration (Daboul *et al.* 2023)). In this example, the focus is on SDM instead of population differences. Sex variation in Figure 5b is explored in a PCA of the symmetric component of Procrustes shape using two almost identical sets of landmarks. The only

Fig. 5 (next page). Examples of ME. **a**. Suspected bias due to a time lag in the data collection. The scatterplots (a2-a3) are ordinations of 3D cranial shapes in a large sample of African adult men; above the ordinations (a1), the mean shape differences between green (first round of data collection, N = 377) and blue (second round, N = 161) groups are shown (magnified five times) in dorsal, side and frontal views. In the PCA (a2) there is a small amount of separation between 'green' and 'blue' on PC2 (variance explained in parentheses). In the DA (a3), the 'green-blue' separation on the horizontal axis (DF1, with, in parentheses, the between group variance explained) is almost perfect despite the fact that groups, whose differences are being maximized in this plot, are in fact the real 32 geographic populations (shown using convex hulls - see main text). **b**. PC1-2 scatterplots of adult 3D craniofacial shapes in a European sample of adult women (N = 351, shown in pink) and men (N = 380, in blue). The configuration is smaller and different from the one in (a). Above the scatterplots, the shape corresponding to the positive extreme of PC1 (variance explained in parentheses) is shown, in dorsal view, using displacement vectors (b1-b2). The scatterplot to the left (b3) is the full landmark configuration, whereas the one to the right (b4) has euryon removed. Euryon dominates PC1 differences (b3) in the full configuration (b1). After removing it (b2), not only there is no landmark that dominates variation on PC1 (b4), but also the separation of females and males disappears and (see main text) the shape variance accounted for by sex drops from 6% to 3%. **c**. PC1-2 scatterplot (variance explained in parentheses) of marmot mandibular shape using the reduced landmark configuration in real (c1) and simulated (c2) samples of hoary marmots and woodchucks. The simulated data are obtained by adding to the original shape coordinates large random gaussian noise (SD = 0.05) before Procrustes re-superimposing the data. Random noise (c2) completely obliterates the real differences (c1) and brings the F ratio and Rsq from F = 38 and Rsq = 20%, in the real data, to F = 2 and Rsq = 0.8% in the simulated ones. In both datasets, the tangent space approximation (assessed in TPSSmall) in the real data (c1) produces a correlation of one between shape distances; however, whereas the slope of the least square regression (tangent space Euclidean shape distances onto Procrustes shape distances) is one in the real data, it is 0.97 in the simulated ones, which suggests an almost problematically large amount of shape variance in this second dataset.


(a)
(a1)
PC2 (7%)
PC1 (10%)
(a2)
DF2 (5%)
DF1 (17%)
(a3)


(b)
(b1)
PC2 (11%)
PC1 (37%)
(b3)
(b2)
PC2 (13%)
PC1 (18%)
(b4)


(c)
PC2 (16%)
PC1
(28%)
(c1)
PC2 (7%)
PC1
(9%)
(c2)

difference between the two is that, from an original configuration of 22 landmarks, a couple of cranial vault bilateral landmarks called euryon has been removed in one dataset. Despite the minimal difference in the configuration, when euryon is present (left scatterplot, in Fig. 5b), its position in relation to other landmarks seems to dominate the pattern of variation on PC1. This is suggested by the long displacement vectors [13] of euryon in the visualization of the positive extreme of PC1 (Fig. 5b1). The vectors of all other landmarks, in contrast, are almost invisible in the shape diagram. This sharp dishomogeneity in the lengths of the displacement vectors disappears after excluding euryon (Fig. 5b2). The profound influence of this pair of landmarks on the results is also inferred by comparing the dispersion of the observations in the PCA scatterplots. With euryon, the scatter is more elliptical (Fig. 5b3), but becomes almost circular after its exclusion (Fig. 5b4). Indeed, PC1, with euryon, accounts for three times more variance than PC2, but it is only 40% larger without it. Worryingly, euryon on its own (Fig. 5b3) seems to produce a small degree of separation between women (pink) and men (blue), that completely overlap without it (Fig. 5b4). If SDM in shape is tested, Rsq is 6% with euryon included, but drops to 3% when this pair of points is left out.

Is the better separation of sexes caused by euryon true or an artefact of ME? Because, unlike the African dataset, in this example, duplicates are available for a subsample of individuals, this question can be answered. Euryon corresponds to the points on the sides of the vault at which the greatest cranial width is measured. It is notoriously difficult to locate with precision. Its low precision is confirmed using the same type of analysis, as in the first subsection of the ME methods, which indicates a median absolute imprecision approximately twice the average of other landmarks (unpublished observations). In the ANOVA of shape using the duplicates, the error in digitizing euryon turns out to be relatively large. This is deduced by observing that its exclusion (i.e., the exclusion of just two landmarks out of 22) is enough to reduce the ME Rsq from 19% to 15%. Unlike in marmot mandibles, where individual shape variation was more than 30 times larger than ME, in this sample of human craniofacial data, individuals account for just four to five times (respectively including or excluding euryon) more variance than ME. The sharp effect of euryon on the patterns of shape variation, the estimate of SDM and the magnitude of ME, all strongly suggest that euryon is highly imprecise. However, there is not yet a demonstration that euryon is biasing (inflating, in this case) SDM. For assessing whether this is the case, one might add a couple of steps to the ME protocol. The extra steps (below) allow to assess the magnitude and direction of a systematic error in relation to a factor being investigated, which in this example is SDM.

The magnitude of a bias due to ME can be estimated by a multivariate regression of shape onto a dummy variable for the duplicates. I use duplicates as an example, because it is simpler, but a similar approach can be extended to three or more replicates. What is a dummy variable? A dummy variable is, as already briefly said, simply an expedient to code groups with an integer (e..g, first duplicate = 0 and second duplicate = 1). This use of a regression to compare group means is the same analysis I will later (part B) employ for testing SDM in marmot mandibles, as well as pairwise species differences. In the context of ME, the regression estimates the mean difference between the duplicates relative to the variance of the sample: if errors are random, the difference will be close to zero; if, however, at least some landmarks are consistently misplaced (say, they tend to be digitized in duplicate 2 a few mm forward compare to their position in duplicate 1), there will be a mean difference between duplicates. This difference is the systematic error or bias.

[13] Displacement vectors (or lollipops (Klingenberg 2011, 2013)) are small arrows showing the position of the landmarks in a specimen, compared to another one (often, the sample mean shape). As mentioned in this paper and its twin (part B), and extensively discussed by Cardini & Verderame (2022), per-landmark variation in Procrustes shape data is inaccurate and potentially misleading, which also means that displacement vectors cannot be interpreted one at a time or in subgroups. If this is done, as with euryon in this example, the vectors can only provide a very approximate clue to local variation, that must be interpreted with the greatest caution.

A systematic error that varies in females and males might lead to an overestimate (inflation) or underestimate (deflation) of the true differences between sexes. For instance, in the second digitization, an operator might accidentally tend to place euryon more laterally in females than in males. Because, in humans, crania are on average more dolichocephalic (i.e., narrow) in males than females (Milella *et al.* 2021), and references therein), the digitization bias make the cranial vault of females look even wider, in relative terms, and, thus, inflate this aspect of SDM. The Rsq of the regression, which is the sample variance (all duplicates) accounted for by mean differences (first duplicate versus second one), captures the total magnitude of the bias. The vector of the regression coefficients of the Procrustes shape coordinates, instead, measures the direction (i.e., the pattern) of mean group differences. If there is a component of the average shape change between the two digitizations (here, likely the more laterally displaced euryon in females in the second digitization) that is partly collinear with shape SDM (the dolichocephalic trend in male crania), the vectors of the two regressions, one for the bias and one for SDM, will be correlated or, which is the same, will form a small angle. The smaller the vector angle, the stronger the similarity in the shape changes captured by the two regressions and, thus, the stronger the effect of the bias on the real pattern of, in this case, sex differences. Thus, although a bias always introduces inaccuracies, when the bias is approximately in the same direction of the group differences one is interested in, the vectors will be positively correlated and the error will inflate those differences (or deflate them, if the digitization and group differences are negatively correlated). Angles between multivariate vectors can be easily computed in MorphoJ, as explained in part B. To calculate the vector of shape SDM (see also B1), a research needs to regress individual shapes (i.e., the averages of the duplicates) onto a dummy variable for sex. In my example, because duplicates are available only for some of the hundreds of individuals shown in Figure 5b, the two regressions, one for the bias and the other for SDM, are done respectively in a subsample of the complete dataset and in the total sample. In general, when vector angles of high dimensional data are estimated in relatively small samples, inaccuracies can be large (Cardini & Elton 2007). The bias is, therefore, less accurately estimated than SDM and the analysis of their relationships must be seen as preliminary and interpreted with caution.

I ran all multivariate shape regressions in MorphoJ and also computed the vector correlations in the same software (with commands briefly explained in B1). For the duplicates, users should be aware not to compute the test of significance in MorphoJ because the regression is not designed to take into account that the data are non-independent paired observations. Besides, in this specific context, exploratory and in a subsample, P values are less interesting. The regression shows that, using the human crania subsample with duplicated digitizations, Rsq for the average difference between duplicates (bias) accounts for 6% of shape variance including all landmarks. However, when euryon is excluded, the Rsq of the bias drops to 2%. For SDM, the regression Rsq is, as anticipated, is 6% with euryon and just 3% without it. When I tested the angle between the vector of mean duplicate differences and the SDM vector, the angle was large[14] (62°), but significantly smaller than expected by chance for vectors in random direction (P = 0.0025), if euryon was included. Without euryon, however, the angle was even larger (86°) and the test was no longer significant (P = 0.362). The conclusion seems to be that digitization error in this dataset introduces a bias, that the bias is largely due to euryon and that including euryon tends to increases the estimate of sex differences probably at least in part because of the bias.

[14] For tests of angles between dummy variable-regression vectors, there are a few caveats. If duplicate one is coded 0 and duplicate two is coded 1 in the dummy variable, the regression vector is not the same as using 1 for duplicate one and 0 for duplicate two: the Rsq and absolute values of the coefficients are identical, but the sign of the coefficients is reversed. The same holds for the regression of shape on sex. Because of the arbitrariness in coding the dummy variable, I suggest to use the smallest of the angles one obtains depending on whether duplicates were coded as 0 1 or 1 0; this is equivalent to consider the absolute vector correlation. For instance, in my example, depending on the coding of the duplicate dummy variable, the angle between the vector for the bias and the SDM vector can be 62° (r = 0.47) or 118° (r = -0.47), which is simply 180°-62°. I used 62° to infer the effect of the bias on SDM and argued that the bias spuriously increases SDM because the SDM Rsq drops after excluding euryon, the landmarks mainly responsible for the bias. Also, I visualized in MorphoJ using lollipops (displacement vectors for landmarks), the average differences between duplicates and the average differences between men and women and did notice that euryon tends to be laterally displaced in the second duplicate, thus biasing SDM.

Overall, in this example, the evidence shows that euryon is very imprecise and can have an important impact on results. Thus, it seems wise to remove it (as we did in Daboul *et al.* 2023). Generally, when, in the ME ANOVA, differences at the level of variation one is interested in (individual SDM, taxonomic differences etc.) are much larger (not just significantly larger!) than ME, it is unlikely that a systematic error, if present, may strongly influence results. It will introduce a degree of inaccuracy, but that should not be large enough to change the conclusions of the study. As I mentioned, ME in the human crania example was ¼ (1/5 without euryon) of individual variation, which is moderate but not tiny. However, in that study, after excluding euryon, a negligible digitization error was confirmed by the observation that 96% of the time duplicates of an individual clustered together as nearest neighbours in the phenogram based on Procrustes shape distances. In the marmot mandible dataset, in contrast ME was just 1/30 of individual differences and, thus, accounted for a minuscule amount of shape variation. More importantly, for a study mainly focusing on taxonomic variation, with ME being ~ 2% and average marmot species differences ~ 20%, in terms of Rsq (Table 5), it seems improbable that a potential bias may appreciably impact taxonomic differences that are 10 times larger than total ME. Nonetheless, if a researcher is in doubt or has reasons to suspect a strong bias, he/she can add the steps I exemplified with the human crania to explore the magnitude and direction of systematic errors. If this is done in the marmot mandible shape dataset, consistently with the graphical examination (Figure 4), which did not suggest any appreciable systematic error, the Rsq of the regression for the mean differences between duplicates ranges from 0.4% (15 landmarks configuration) to 0.1% (reduced 12 landmarks configuration). Such a small Rsq confirms the absence of any strong bias and makes unnecessary any further exploration using vector angles.

Systematic errors, especially if collinear with the aspect of biological variation under investigation, are an important source of inaccuracy. Systematic ME might, in fact, be more important than random ME, as recently argued by some contributors in morphmet, the email discussion list of morphometricians (see https://www.mail-archive.com/morphmet2@googlegroups.com/ and search the thread entitled "Measurement error in geometric morphometrics"). My view is that both random and systematic ME are relevant. For instance, random ME, despite not biasing results, can reduce statistical power and potentially lead to inaccurate conclusions. This is shown, using a mix of real and simulated data, in Figure 5c. The data are a subset of the marmot mandibles used in this paper. The subset includes only woodchucks and hoary marmots using the 12 landmarks configuration. In the PCA scatterplot to the left (Fig. 5c1), I analysed the real data. The separation between the two species is evident. Even if a conventional PCA is not aimed at maximizing group differences, here they are so large that dominate total variance and suggest a minimal overlap between the two species on PC1 (Fig. 5c1). In the PCA scatterplot to the right (Fig. 5c2), I used simulated data created by adding a huge amount of random (non-directional) noise to the Procrustes shape coordinates of the two species. This simulates the variability one might find if there is a very large random imprecision in the digitization of each landmark. The simulated data are Procrustes re-superimposed and analysed. Now, in the PCA scatterplot (left, Fig. 5c2), the two species overlap almost completely, even if the Procrustes distance between their means is virtually identical to the distance in the real samples (respectively, 0.039 and 0.041). The differences are still there, but are masked by the vast random noise in the data. If differences are tested, Rsq is 20% (highly statistically significant) for the real data, but only 0.8% (non statistically significant) for the simulated ones. Statistical power dropped and the random error has completely obliterated the biological differences by inflating within group variance. This example is, of course, an exaggeration, as it is unlikely that one might have such a large ME. Nonetheless, it is a proof of principle that helps stressing how random error is no less relevant than a systematic bias: they are both important, as they can alter results by introducing inaccuracies. However, with a ME much smaller than the biological variation being investigated (within and between group differences, in my marmot study), it is unlikely that any of these two sources of inaccuracy may strongly impact results.

Imprecise landmarks, data dimensionality and assumptions

After discussing examples of common sources of ME and suggesting additional analytical steps for exploring systematic errors, let me go back to landmark precision, which, like ME biases, had not been considered by V&C.

Finding low precision landmarks in Procrustean GMM is not straightforward. This type of analysis requires the use of methods that summarize digitization error (and no other source of imprecision in landmark coordinates) before specimens are superimposed (Cardini & Tongiorgi 2003; von Cramon-Taubadel *et al.* 2007). After the superimposition, one might still get some clues on low precision landmarks, by simply examining the amount of scatter of each landmark around their mean, but this type of inspection can be very misleading, because the superimposition tends to spread differences across the entire configuration (Cardini & Verderame 2022).

The identification of imprecise landmarks is optional, but can be informative and helpful, as shown with euryon, to potentially reduce sources of both random and systematic error. It may happen that specific landmarks spuriously make a major contribution to shape variation, because they are difficult to locate with a good replicability. Sometimes, biologists know in advance which landmarks could be problematic in this respect. With marmot mandibles, I knew that L12 is easily misplaced (Cardini & Tongiorgi 2003), but I also had doubts on the precision of L11, L13 and L15. For L11, L12 and L13 (but not for L15), the analysis of absolute per-landmark precision confirmed the large errors in their digitization. The error was largely random in this case, but on its own enough to double the effect of ME compared to individual variation in shape, as shown in the Results and briefly discussed below. This observation, as well as the larger than average imprecision of these three landmarks, led to the decision to exclude them in all subsequent analyses.

By comparing the ME ANOVAs with and without the highly imprecise landmarks, one can gain further insight on their effect. For CS, the exclusion of L11, L12 and L13 made no appreciable difference in the results. This is unsurprising, because individual variation in CS was hundreds of times larger than digitizing error. In contrast, the exclusion of L11, L12 and L13 had a small, but measurable effect on shape. Leaving them out, the variance accounted for by ME was almost halved and, thus, the ratio of individual to ME Rsq, which measure the relative magnitude of biological variation compared to variation due to ME, went from 16 to 32. The percentage of sister duplicates in the phenogram of shape also raised from little more than 90% to almost 100%. In this dataset, it seems that the information potentially added by the three low precision landmarks was very 'noisy', even if individual variation was much larger than ME despite the additional noise. L11, L12 and L13 map on anatomically important regions, especially in relation to the insertion of masticatory muscles. Without them, the morphology of the lower part of the horizontal ramus is poorly described. However, there is a trade-off between quantity (more measurements) and quality (higher precision and potentially better accuracy). The morphometrician has to decide whether no information is better than imprecise measurements, that, in the case of my marmot dataset, double the impact of ME on shape.

For anatomical regions, such as the marmot lower part of the horizontal ramus or, for instance, the cranial vault of most mammals, that largely lack landmarks with a clear anatomical correspondence, a researcher might want to consider the use of semilandmarks, a series of generally closely spaced, arbitrary points used to approximate curves and surfaces. Yet, as explained in the last section of Appendix A, semilandmarks are not the same of landmarks, regardless of how mathematically treated. They can provide useful information, especially for individual identification and the reconstruction of fragmentary fossils (Hublin *et al.* 2009), but they also increase dimensionality and may add noise rather than increase the signal relevant to a specific study hypothesis (Cardini 2020a, 2020b).

A final consideration on ME concerns assumptions. As in all other analyses, results of ME analyses can be sensitive to violations of statistical assumptions. Non-independence of the observations, and the possibility of autocorrelation in the data, are considered in Cardini (2020a) and briefly discussed in relation to outlier detection and in the Appendix A. Most of the methods used in this paper use linear models, which assume a linear relationship among continuous variables (Hair *et al.* 2013). The ME ANOVA makes further assumptions: homoscedasticity (i.e., similarity of variances and covariances), negligible interactions among factors and, for the Goodall's F, isotropic variation in shape. Homoscedasticity is also discussed in the Appendix A, whereas the reason for the assumption of negligible interactions has already been concisely explained. There are tests for all these assumptions, but they go beyond the scope of this paper and, especially for multivariate data, they may require a degree of expertise with R or other statistical programming languages.

Adequate sample size is also important for both for accuracy and statistical power. This is true for all tests, including the assessment of ME. Power is discussed later. On sample size, I anticipate a general consideration. In parametric tests, a first warning that samples might be too small is an error message in the statistical software or a test missing a specific statistic. This happens, for instance, in MorphoJ's ME ANOVA, with Pillai's trace that cannot be computed unless N is large enough relative to p (Klingenberg *et al.* 2002). When N is too small, MorphoJ reports the P value of the Goodall's F, without the multivariate Pillai's test. Because Goodall's F assumes isotropic variation around landmarks, the F test may be inaccurate when deviations from the isotropic model are present.

If the absence of a parametric test is a clear sign of problems, one cannot, on the other hand, assume that, when the test is present, sample size is appropriate. Resampling statistics generally is less restrictive in terms of sample size-related computational constraints, but that does not mean that estimates of the test statistics and other sample parameters are accurate. When differences are small, as in most taxonomic comparisons, one typically needs large samples (Cardini *et al.* 2021). How large relative to the number of shape coordinates? There is no simple rule of thumb for a desirable p/N in a multivariate analysis, but there is an increasing amount of evidence that large p/N ratios (i.e., many more variables than individuals in a sample) may create problems even in simple exploratory analyses (Bookstein 2017; Cardini *et al.* 2019, and references therein). Dimensionality reduction using a PCA can help in these cases, but it is not a panacea and requires a demonstration, often far from trivial and sometimes easily misleading, that all the variance relevant to the study aim has been preserved.

When sample size varies across groups, there might be further issues, as one drifts away from an ideal perfectly balanced design, with an identical N in all groups. For heterogeneous sample sizes, a researcher can start exploring the impact on results by double checking findings after excluding the smallest samples. This is what I will do in the majority of the analyses in part B. In the ME ANOVA, I did not do it, however, because differences were large (e.g., the individual Rsq compared to the ME Rsq) and results were strongly supported by graphical analyses that are usually less strongly affected by the presence of small samples. Yet, a replicate analysis that includes only larger samples is easily done in MorphoJ. This requires subsetting the main dataset (*Preliminaries, Include or Exclude Observations* or *Preliminaries, Subdivide Dataset by*) using an ad-hoc classifier to select the largest samples. If I had done it, so that only *M. caligata*, *M. flaviventris* and *M. monax* were included, using the reduced 12 landmarks configuration, Rsq and P values would have lead to the same conclusion: individual variation would be almost 800 times larger than ME for CS and 34 times larger for shape compared to ~ 750 and 32 times, for respectively CS and shape, including all species.

A2) Search for potential outliers

Methods (A2)

In general, an outlier can be defined as "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1980: 1, cited in

Cook *et al.* 2021). 'Outliers' in morphometrics are cases, whose size or shape set them apart from others in the same sample. Outliers can be single individuals or a small cluster of unusual observations. For simplicity, I use a single term, outlier, for all types of unusual observations regardless of whether they bias the results ('influential' cases) or not.

I mostly look for outliers in data plots. Graphical approaches are just one option and there is a variety of alternatives. Outlier detection is in itself a subdiscipline of statistics (Smiti 2020). An inevitable limitation, with all methods, is that spotting outliers in GMM is only really feasible if they are strikingly different and/or the reference sample is large. As sample size becomes smaller, finding outliers is increasingly harder or impossible. Yet, this is another fundamental preliminary step, but one that, in the GMM literature, is mentioned even more rarely than the assessment of ME. Sometimes an outlier represents a case of true, but rare, phenotypic variation. However, often, as I later discuss, outliers are simply the consequence of mistakes during the data collection. Outliers can also be misclassified individuals or specimens who lived in captivity or had a pathology. They must be found and corrected, when due to errors, or potentially excluded.

I looked for outliers twice, as I anticipated. First, I did it relatively quickly and focusing on extreme outliers, using all the data, including the duplicates. Later, I removed low precision landmarks, averaged duplicates and checked again, more carefully, the potential outliers. The main reason for repeating the search for outliers after assessing ME is that the landmark configuration has been reduced and, thus, shape relationships may have changed. Searching for outliers after averaging duplicates/replicates has also the advantage that outlier detection is not confounded by replicates of the same individual and the data are precisely those used in the main analysis (part B of this study). A researcher should look for outliers also when a pilot study is performed in a smaller sample, for instance to assess ME or explore statistical power, before embarking in a larger project. Later, he/she will have to redo the outlier detection in the main study sample including all groups and individuals.

I used the same approach (explained in detail in the next subsections) in the first and second search for outliers. Although I may look for outliers in the total sample, outliers are more difficult to spot if the data mix within group variation with between group differences. Within a taxon, one should also take age and sex differences into account, if these types of variation are potentially important in a study group. As typical in science, the basic idea of this and other analyses is to control for confounders while focusing on the specific factor one is interested in, which in this case is 'unusual' specimens in a sample. Thus, I plot data using the total sample, but mainly inspect them carefully within each taxon. Age variation in my marmot dataset is unlikely, as all specimens are adults and ontogenetic change is minimal or absent in sexually mature, fully developed, individuals. In contrast, because I am analysing both sexes, I may have to look for outliers within species in separate samples of females and males. I did not split species by sex for detecting outliers, however, since, in marmots, males tend to have larger body mass (Schulte-Hostedde 2008), but SDM is small in foot and cranial length (Matějů & Kratochvíl 2013) and usually negligible in marmot mandibles (Cardini 2003; Nagorsen & Cardini 2009). Nonetheless, I preliminary inspected graphically that large differences between females and males were not evident. The almost complete overlap of females and males within each species supported the absence of large differences and, therefore, the likely appropriateness of pooling sexes in the search for outliers. In contrast, as with ME (see above), if the analysis was in a taxon, such as, for instance, guenons with their large SDM (Cardini & Elton 2008b), I would use separate sex analyses also for the outlier detection. The same reasoning applies to ontogenetic studies: potential outliers should be examined both using all individuals in a taxon, as well as within more homogeneous samples of specimens of similar age.

Outliers: univariate size

For the detection of CS outliers, I used within-species box and jitter-plots. In the box-plot, the 'box' marks the boundaries of the quartiles right above or below the median (usually shown by a straight line within the box), while the whiskers outside the box define the range of the data (minimum to maximum). In fact, the box-plot in PAST (*Plot, Barchart/Boxplot* and then select *box-plot*) has an option to modify the whiskers so that they emphasize outliers, if present. The option is commonly used in statistics and well explained in the pdf manual of the software: "If the '*Outliers*' box is ticked... The whiskers are drawn from the top of the box up to the largest data point less than 1.5 times the box height from the box (the 'upper inner fence'), and similarly below the box. Values outside the inner fences are shown as circles, values further than 3 times the box height from the box (the 'outer fences') are shown as stars" (p. 28, in the manual of PAST 2.17c).

The box-plot is intuitive, but I also explore jitter-plots. Jitter-plots are univariate scatterplots of the observations in one or more groups. Jitter refers to the possibility of slightly shifting the points, randomly to the left or right in a vertical plot, such as the one I am using, so that individual observations are easier to see in case they overlap. The jitter-plot provides details on the distribution of the individual observations (their density, possible clusters etc.) and may help to see unusual observations that were overlooked in the box-plot. Sometimes, with large samples, it may be useful also to inspect normality within groups (e.g., species or species with separate sexes) using a histogram[15]. In all these graphical analyses, a very isolated individual is a potential outlier.

Outliers: multivariate shape

Multivariate outlier detection for shape data is less simple than for univariate data. MorphoJ offers a method (*Preliminaries, Find Outliers*) based on the shape distance of each individual from the mean of its sample. In taxonomic studies, *Find Outliers* should be used within homogeneous taxa (each marmot species, in my case). The distribution of observed shape distances is compared with the expectations for multivariate normally distributed data. In the output window, specimens are sorted in order of decreasing distance to the mean. Therefore, the 'top' specimens in the list are those to carefully inspect. The approach seems to work well with large samples, but is less useful in small ones, such as Alaskan and Olympic marmots. A main advantage of MorphoJ is that the potential outlier is visualized, relative to the sample mean shape, using displacement vectors. With the caveat that per-landmark interpretations of shape variation can be, as already mentioned, misleading (Cardini & Verderame 2022), a very long vector provides a clue to detect a potentially misplaced landmark. When there is a pair of unusually long vectors, maybe involving two consecutive landmarks with vectors pointing towards each other, they might indicate that their order has been swapped by mistake during the digitization (e.g., L2 digitized before L1).

MorphoJ's *Find Outliers* needs to be complemented with other methods. I explored multivariate summary plots using PCAs (in MorphoJ or PAST) and UPGMA phenograms (in PAST). In the scatterplots, isolated specimens are potential outliers. Researchers often explore only PC1vs PC2, but one should almost always inspect scatterplots of several pairs of the first PCs. In phenograms, outliers are generally lone specimens, or small groups of individuals, isolated on one or more basal (i.e., close to the root) branches. The path between their branch and the next cluster should be relatively long, thus indicating a clear separation. PCAs and phenograms, as well as MorphoJ's *Find Outliers*, typically produce congruent results for strong outliers.

[15] I am not exemplifying this method, but it is easy to use in PAST: select the column with the CS of one group, click on *Plot, Histograms*, and then check the box *fit normality* and optimize the number of bins in relation to sample size. Normality is not an issue if one is using resampling methods for testing differences. Nonetheless, a histogram is useful to explore the data distribution (more or less symmetric, uni- or multimodal etc.), as well as helpful as complimentary approach to spot isolated observations. With univariate measures such as CS, sometimes log-transformed data improve normality, which may be important for parametric tests.

I created PCA scatterplots and UPGMA phenograms within each species after quickly inspecting the total sample using the same methods. To perform the analyses rapidly, one can superimpose the whole dataset (all species and specimens) in MorphoJ, export groups and Procrustes shape coordinates as text files, and do ordinations and other analyses in PAST. When analyses are run within a group, however, it may be more accurate to re-superimpose the data. This is because, when data are subsampled, there can be a tiny difference in the shape distances measured in the tangent space. PAST can do the re-superimposition: one has to select the rows with the individuals of the group of interest (one of the marmot species, in my case) and, then, click on *Transform, Procrustes (2D/3D)* [16], selecting the appropriate options for 2D or 3D landmarks. Yet, for taxonomic studies of small differences, the re-superimposition within a group might not make any practical difference. This means that the risk of inaccuracy by re-using Procrustes shape coordinates in a subsample seems minimal. Thus, at least for preliminary analyses such as the detection of outliers, one can run them in PAST by selecting all Procrustes shape coordinates for the total sample, but just the rows corresponding to a specific group when examining shape data within samples [17].

Finally, what happens if a potential outlier is spotted using, for instance, a PCA scatterplot? First, I suggest to double check the specimen in the phenogram and in the list made using MorphoJ's *Find Outliers* (and vice versa, when one spots the outlier in the phenogram or in MorphoJ). If it is a very strong outlier, all different methods should flag the specimen as 'unusual'. As for CS outliers, however, I also recommend to inspect the raw data of that specimen (image, landmark positions and order, group affiliation) to detect potential errors (photographic distortions, misplaced or mislabelled landmarks, wrong group etc.). Although this is not related to the outlier detection, once a phenogram is computed, one can verify that no cluster has specimens with zero shape distances. If this happens, 'leaves' (i.e., the terminal branches) are connected by a line perpendicular to the branch. Because a zero shape distance indicates perfect identity, it can be used as a tip to spot individuals duplicated by mistake. More generally, sorting specimen identifiers, as well as museum catalogue numbers (for museum specimens), in ascending (or descending) order in a spreadsheet and checking that no consecutive entries are identical also helps to find unintentionally duplicated observations.

Results (A2)

Overall, I detected 17 potential outliers. None of them looks extreme, as they are moderately isolated but close to the range of variation in their species samples. Of these 17 individuals, three were potential outliers for CS, eight for shape, one was a young-looking individual that was also an outlier for shape, and five did not look unusual in the statistical analyses, but they were either strongly damaged or suggested a pathology. The exclusion of these 17 specimens left 445 individuals (96% of the original sample) for the main analyses.

---

[16] Linux users, who are for now likely to be bound to use the old 2.17c version of PAST, must bear in mind that the data are not automatically projected into the tangent space (*Transform, Project to tangent space* if you want to do it). Also, in this version of PAST, there is a bug in the 3D superimposition which, in fact, does not standardize CS. For this and other reasons, I advise to superimpose the data in MorphoJ (or TPSRelw) and then to import them in PAST.

[17] Alternatively, there is a 'trick' in MorphoJ to simultaneously perform a specific analysis within all subsamples (all species, in this example). This works with some analyses, such as the PCA, but not with others, such as *Find Outliers*. With the PCA, it may be faster than the approach I suggested for doing within-species PCA in PAST. First, one splits the total sample by species using *Preliminaries, Subdivide Dataset By* with the species classifier. Then, he/she selects all species (hold the shift key while clicking on the species subsets). Finally, the two MorphoJ's commands for the PCA (*Preliminaries, Generate Covariance Matrix* and *Variation, Principal Component Analysis*) are issued using the *together* option in the pop-up window. As a result, in my marmot dataset, all six within-species PCAs are done at the same time. Clicking on the appropriate window in the *Graphics* sub-windows, the researcher will, then, be able to inspect the scatterplots of one or the other species. However, the within-species phenograms will still have to be done in PAST because MorphoJ does not do cluster analyses.

Figure 6 shows how potential outliers in size (marked in red) are detected using box and jitter-plots in PAST. In this software, the two types of plots are different options, available in the same command window, but, unlike in R (e.g., Figure 2 of Cardini & Chiapelli 2020), they cannot be merged into a single graphic. A researcher can plot them separately or, as I did, overlay them in a photo-editing software, after having saved each, one at a time, as an image file. The whiskers in the box-plot suggest only one potential outlier. This individual is a hoary marmot (cal) with an unusually small mandible. The specimen has a complete permanent dentition, with a degree of tooth-wear, and does not look young despite its small size. The other two individuals with a somewhat unusual mandibular size belong one to a yellow-bellied marmot (fla) and the other to a woodchuck (mon). They are very large specimens relative to the range of CS in their respective species samples. None is above the threshold for outliers in PAST, but they are isolated by a visible gap separating them from the main cluster of their conspecifics. In the other three species, the range of size differences is smaller and there are no apparent outliers or isolated individuals.

For shape, the outlier detection is exemplified in Figure 7 using yellow-bellied marmots. The UPGMA phenogram does not suggest very strong outliers. However, there are three individuals (identified by the numbers 193, 254 and 278, and marked in red) on relatively isolated branches at or close to the tree root (Fig. 7a). Although these individuals are not isolated in PCA scatterplots (Fig. 7b1-2), all of them have scores close or equal to the maximum on several of the first PCs. Two (193 and 254) of the three individuals are shown to have the 2nd and 4th largest shape distance from the species mean shape using MorphoJ's *Find Outliers*. Figure 7c illustrates, as an example, the difference between the yellow-bellied marmot sample mean shape (black landmarks) and individual 193 (red displacement vectors). The diagram is modified from the display of MorphoJ by adding a wireframe to aid the visualization. By inspecting shape change in relation to the photograph of 193 (Fig. 7d), this specimen seems somewhat unusual because of a rather long coronoid, as well as for the vertical bone loss in the alveolar region of the toothrow, which makes the roots of the teeth clearly visible above the lowered alveolar margin. Specimen 193 does not look diseased, but, as a consequence of the alveolar bone loss, landmarks on

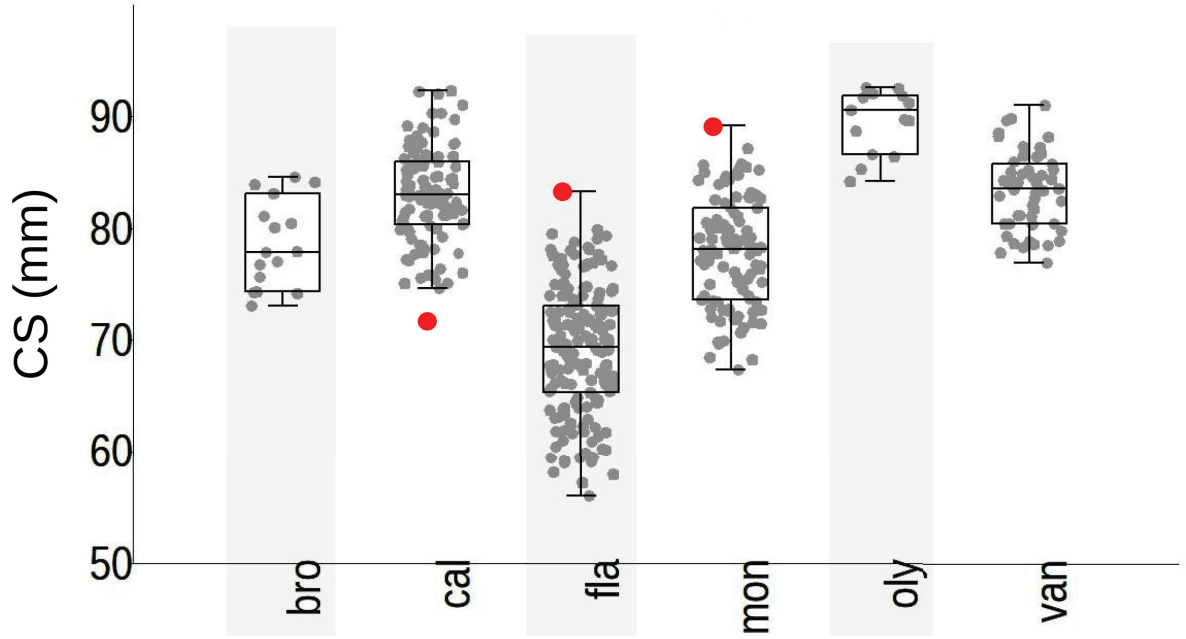


Fig. 6. Box and jitter plot of species mandible CS. Red circles mark potential outliers. Species names are abbreviated as shown in Table 2.

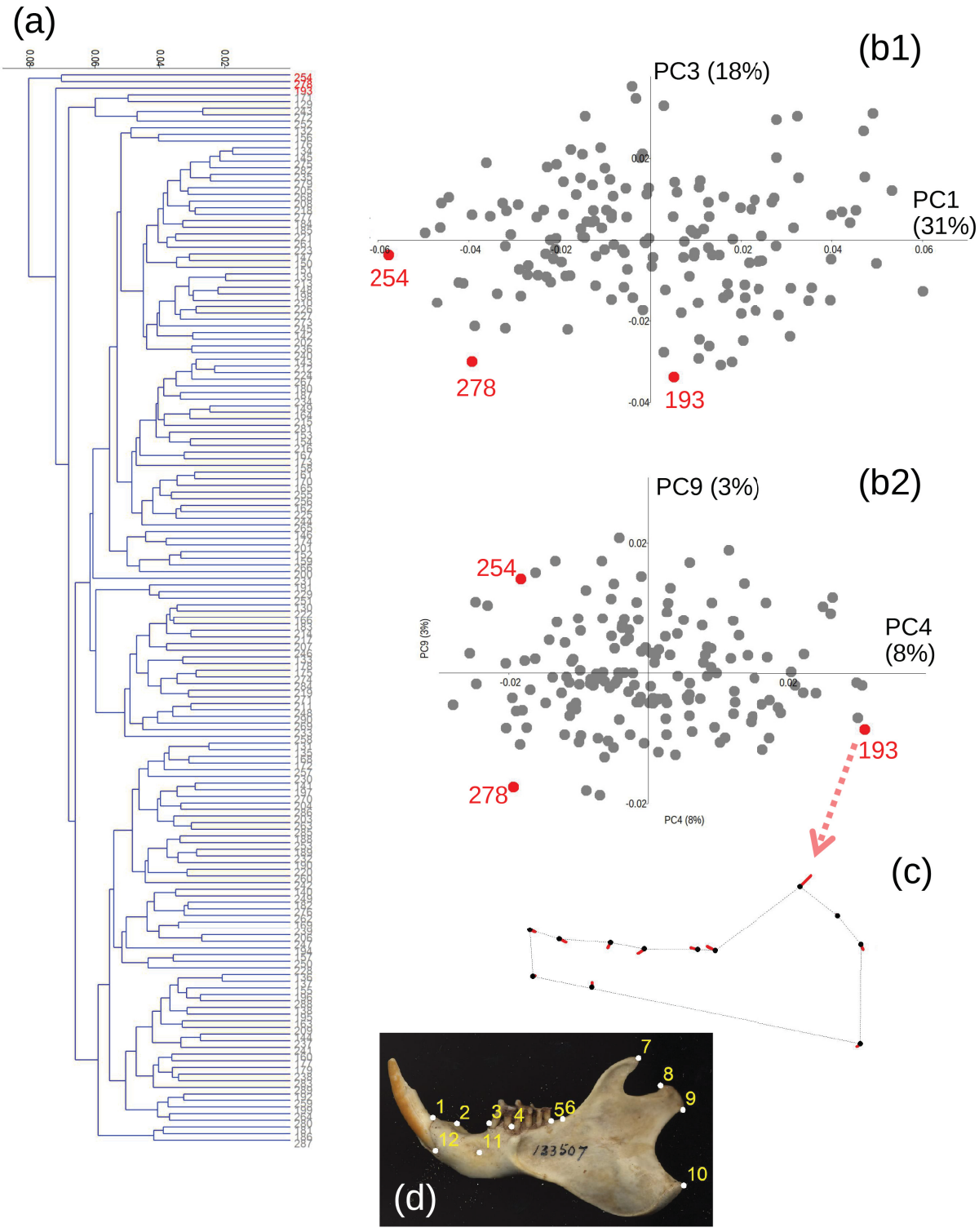


Fig. 7. Example of search for shape outliers (emphasized in red) in yellow-bellied marmots: (a) phenogram of shape; (b1-2) scatterplots of PC1 vs PC3 and PC4 vs PC9 (percentages of variance accounted for by each PC shown in parentheses); (c-d) visualization of individual 193 using displacement vectors for this specimen relative to the sample mean shape (c), as well as its original photograph (d).

its margin may be mapping a subtle pathological aspect of anatomy specific to this individual. Thus, even if it is not extreme as a potential outlier, I decided to exclude 193 from the main analysis (part B). I similarly excluded the other suspected size or shape outliers. The decision might be overly cautious and its pros and cons are considered in the Discussion.

Discussion (A2)

Outlier detection almost always implies a degree of arbitrariness. The choice of the approach is arbitrary, as well as the decision on what to do with potential outliers, if detected. In fact, some argue that outliers should not be excluded from analyses or should be considered as a potentially important source of information (Cook *et al.* 2021). Although I agree that outliers, if not caused by mistakes during the data collection, can be interesting, their impact on statistical results cannot be overlooked. The sensitivity to the occurrence of outliers also varies depending on the method, with some being more robust than others (Daszykowski *et al.* 2007). Regardless of the method, the decision to include or exclude potential outliers should not be guided by the desire to obtain results that match the expectations of a researcher: either they are excluded prior to the main analyses or analyses are performed with and without potential outliers and results, if different, are reported and discussed for both cases.

I look for outliers mainly graphically, using univariate and multivariate plots to find unusual observations. How much, however, should an individual be isolated to be inspected as a potential outlier? Isolation is, always, a relative concept, that must be scaled to the amount of variation in the specific sample, which can be a population, subspecies or species, depending on the level of a taxonomic analysis. Unless isolation is extreme, however, the decision about how far from others a specimen should be for flagging it as a potential outlier is fairly subjective. With multivariate data, also, the approach I am using rely on methods that summarize data graphically (ordinations, phenograms) or using distances (Procrustes or Mahalanobis, typically). Multivariate summaries, like all summaries, do not convey the full information in the data and might, thus, mislead or miss important aspects of variation.

There are alternative, more rigorous, approaches for outlier detection. Well-defined 'rules', or thresholds, have the advantage of reproducibility, but they are not free of arbitrariness. As I discuss later, one needs to select the method, check its assumptions, and decide a threshold. Regardless of the choice of a specific method, outlier detection is problematic or almost impossible with very few specimens. As in other analyses, reliably searching for outliers requires large and representative samples. How large does a 'reasonably large sample' have to be? The answer varies in relation to the study material and specific study question. Research in mammals (Cardini *et al.* 2021, and references therein) provides a very crude approximation of a minimum number of individuals one might need when dealing with the amount of variation common in interspecific studies of differences among closely related mammalian species or subspecies. Using adults, we suggested an adequate minimum sample size, on average, in the order of at least 25–40 specimens per homogeneous group (i.e., within taxon or, if SDM is large, within taxon and sex). This number is mostly congruent with previous estimates in morphometric studies of other terrestrial vertebrates (Cope & Lacy 1992; Cope 1993; Polly 2005; Cardini & Elton 2007; Cardini *et al.* 2015; Brown & Vavrek 2015; Kryštufek *et al.* 2016; Schlis-Elias 2020). I stress, here, something discussed better later: large samples are not just a matter of statistical power (i.e., being able to demonstrate differences), but also of accuracy in the description of patterns (means and their similarities or dissimilarities among species, the amount of variation in a species etc.). I also emphasize that estimates for desirable sample sizes cannot be easily generalized. Indeed, in Cardini *et al.* (2021) and most other studies, the minimum desirable N varied widely depending on the taxon, structure and type of analysis. Also, as already mentioned, with large numbers of landmarks and semilandmarks, unfavourable p/N ratios become almost inevitable and they can have an impact on analytical results (Björklund 2019; Cardini 2019; Cardini *et al.* 2019). Thus, taking into consideration p is another important parameter to decide N.

Specimens may be unusual in many ways and for many reasons. A common problem is the occurrence of misidentified specimens. In museum collections, for instance, outdated or erroneous classifications may not be rare. An expert might be able to reassess the classification, for instance, in mammals, by checking diagnostic characters on crania, but also looking at skins and geographic localities. However, if a mistake in the taxonomic identification goes unnoticed, the misclassified individual may become an outlier whose size or shape differs markedly from others in its putative group. I mostly work on museum specimens and, because I am not a group specialist, I rely on classifications made by others. There could be errors or outdated classifications. This is one reason why I am particularly cautious with potential outliers in size and shape. Thus, in large samples, I often prefer to leave putative outliers out, even when they are not extreme ones. This very cautious approach might lead to overlook genuine elements of biological variation. However, the effect of excluding a few individuals is generally a minor one when samples are big. As already discussed, when there are doubts, a researcher may assess the sensitivity of results by redoing at least the most important analyses with or without suspected outliers. Of course, repeating analyses requires a good amount of additional work and, although uncommon in my experience, one might find differences. If this happens, results must be flagged, differences discussed and uncertainty acknowledged.

Outliers in taxonomic comparisons may occur also because of subtle age differences. In this respect, often, biological age can be more important than chronological (or absolute) age. The latter refers to the number of days, months or years of life, while the former can be informally defined as how old one seems to be. In ectotherms, depending on the environmental conditions, the difference between biological and chronological age can be pronounced and biological age is in general more relevant. For instance, adult size and sexual maturity can be reached faster in ectotherms when the temperature is higher or food more abundant. In mammals and other endotherms, this distinction tends to be less crucial. However, there might be cases when, in a study on adult mammals of unknown age, one or more specimens have, for example, fully erupted teeth and adult size, and, yet, incompletely fused sutures and absence of tooth wear. When there is conflicting evidence on age, I graphically check size and shape. If, despite a 'youngish' appearance, the individual is within the normal range of adult variation in its group, I keep it in the analysis.

Damaged or imperfectly preserved specimens can be problematic. Some landmarks could be hard to locate accurately because of, for instance, missing teeth or broken bones, as not uncommon with the rim of the lacrimal foramen or the fragile coronoid process in mammals. For specimen 193, as exemplified in Figure 7c, bone loss in the alveolar region of the premolar and first molar might have contributed to its somewhat unusual shape. I normally include slightly damaged individuals in the main analysis, if they cluster close to the others in the same taxon. Sometimes, like when only a small piece of a bone is missing in a small number of individuals, it may be easy to guess where the landmark should be. This is often the case with the tip of the coronoid process in rodents, if only its very end is missing. One has, of course, be very cautious and guessing a missing landmark should happen only very exceptionally.

With missing landmarks, a more rigorous option is to estimate them mathematically. This type of estimate is generally unproblematic, if just one or a few points are missing in a minority of individuals (e.g., Cardini & Elton 2008a), and there is a variety of methods to estimate missing landmarks (Arbour & Brown 2014). Few of them, however, are implemented in user-friendly software. The 'old' Morpheus *et al.* (Slice 1999) offers three different approaches (use the command *list pmiss impute options* to get more information). One of them simply replaces missing Procrustes shape coordinates with the sample mean (e.g., within species or subspecies), an operation that is easily done also in a spreadsheet. The 'mean substitution' approach, however, is potentially less accurate than some of the alternative methods (Gunz *et al.* 2009). The inaccuracy is unlikely to make a difference in a large and homogeneous sample with very few missing landmarks, but it can be important in other cases. To select the method and

explore its accuracy, missing data can also be simulated. A proper extensive simulation will be more accurate, but a small experiment is easy to do empirically. If a sample is large, one can erase landmarks in a random subsample, so that their positions, as well as the proportion of specimens lacking one or the other landmark, is a realistic approximation of the total sample. Estimates of missing landmarks are later made using one or more approaches and compared for accuracy with the original values (Cardini & Elton 2008a; Arbour & Brown 2014). To explore the sensitivity of results, a researcher can also replicate the analyses using different types of estimates of missing landmarks and compare their results to assess precision.

When outliers are discovered, they must be carefully inspected using the raw data. This might mean checking the digital images and the landmarks digitized on them, but also the accuracy of the classifiers. For example, a specimen could look unusually small or large because the researcher forgot, misplaced or misused the scale factor in a photograph. It is also possible that data were, inadvertently, converted into different units of measures, so that one specimen is in, say, cm and another in mm or inches. If, in a photograph, a ruler or another type of scale factor used to convert pixels into units of the decimal metric system, is inconsistently placed below or above the height at which it is positioned in most specimens, it might lead to over- or under-estimates of CS for that individual. Sometimes, errors in the scaling of the coordinates happen because of different international settings in the operating system (OS), with dots instead of commas as decimal separator (or vice versa) or intermediate csv files using semicolon (as commonly in Windows OS for continental Europe) instead of commas (as in UK-US OS).

It is also not rare, in large data collections, to digitize one or a few landmarks in the wrong order. For example, two contiguous landmarks might be swapped by mistake. This type of error usually has a negligible impact on CS, but strongly affects shape. Swapping landmarks is particularly easy if one restarts the data collection after a relatively long time. Long interruptions in a data collection are never advisable, as discussed in the subsection on ME. When just a few more specimens are digitized by the same operator, after a long time, the problem of a possible bias is easy to overlook. Thus, a researcher may find that the new data form a small, isolated cluster which he/she cannot say if it is genuine or spurious, unless landmark replicability has been re-assessed, as already discussed.

The statistical literature offers a large number of methods for outlier detection (Smiti 2020). For univariate data, the type of box-plot implemented in PAST is probably the most common technique. Unlike the jitter plot, it is based on setting one or more specific threshold, as explained in the Methods. Numerous threshold-based techniques have been developed also for multivariate outliers. A traditional approach, common in morphometrics and implemented in R (*Morpho* package (Schlager 2017)), but not available in any of the free user-friendly programs I know, employs typicality probabilities in a DA (Albrecht 1992; Kovarovic *et al.* 2011). Typicality probabilities assume that the data are multivariate normally distributed; then they predict the probability of an observation having a certain Mahalanobis distance[18] from the mean of its sample using the multi-normal density model; finally they assign the observation to an unknown group if the probability is lower than a threshold (e.g., 0.01). Bootstrapping the data to estimate a, for instance, 99th percentile for the Procrustes shape distances of the individuals from their sample mean shape is analogous to the use of typicality probabilities, but avoids the need of standardizing the data to compute Mahalanobis distances and does not assume multivariate normality. As with typicality, a simple bootstrap is relatively easy to code in an R script.

Threshold-based models using univariate or multivariate distances are rigorous and reproducible, but the cut-off threshold is arbitrary and assumptions, such as normality, may be difficult to verify.

[18] The Mahalanobis distance is a multivariate distance in a transformed standardized statistical space (see Discussion B3).

Taxonomic data often include small samples and sample size may vary considerable across groups. As with most statistical analyses, small and heterogeneous sample size makes rigid threshold-based methods for outlier detection potentially problematic. Also, if the model includes the outliers when built (as almost unavoidable, as one does not know in advance the outliers), the model itself might be influenced by the outliers (Zimek & Filzmoser 2018). A more general issue, affecting all outlier detection methods, including purely graphical and exploratory ones, as in this study, is that individuals in taxonomic samples are rarely a random representative sample of independent observations (Cardini 2020a). For instance, specimens collected in the same locality and year are unlikely to be statistically independent; they may be a type of pseudo-replicate (Colegrave & Ruxton 2018); and they certainly do not represent the full range of variation in a larger population unless, perhaps, when the population is an endemism with a very narrow distribution range. More generally, data collected opportunistically, based on what is available, as it happens with museum specimens, have limitations, which affect all analyses, including the search for outliers. These limitations are hard to avoid, but can be, somewhat, mitigated by a well-designed data collection. This means, for instance, avoiding gaps in the distribution range, and increasing sample size while limiting pseudo-replications by privileging, if possible, multiple localities over multiple individuals from the same locality[19]. It also means trying to control for temporal variation by looking for data collected within a reasonable time span. The time span might change in relation to generation time in a species and, thus, the expected evolutionary change in a population. Thus, for example, a 100 years time span in a sample of elephants might represent just five or six generations, whereas the same time span corresponds to hundreds of generations in a small mouse, whose likely much larger population may have changed significantly over the decades.

Despite all uncertainties and difficulties, checking data for outliers remains fundamental. One might have to specifically tailor the approach to the data and research question. I have not been able to find a review on outlier detection specific to morphometrics, but there is a large number of statistical articles that introduce methods and summarize the main approaches (e.g., Zimek & Filzmoser 2018; Smiti 2020). Consulting this literature allows a researcher to go beyond a simplistic approach, such as the one I described on marmots, and find the most appropriate method for her/his specific study.

Finally, if potential outliers are found, one has to decide what to do next. The decision to include them or not in the main study adds another layer of arbitrariness. That is true even for rigid threshold-based methods, as both the specific algorithm and the threshold are set by users, who select one of many options. The decision can be simpler if samples are large and representative, and outliers are observations separated by a wide gap from all others. No such 'extreme' outliers were present in my marmot mandible dataset. However, the species samples with a few potential outliers were so large, that these moderately unusual individuals could be left out without appreciably altering the total sample size. One could, of course, argue that, for this same reason, including them in the study was unlikely to change the results. However, this argument should be backed by evidence, which, as mentioned, may require to redo the main analyses with and without potential outliers. This large additional work does not seem worth in a dataset with large samples and a very small proportion of suspected outliers. For this reason, as well as for the uncertainties in the accuracy of museum classifications of marmots, I opted for the conservative decision of excluding all suspected outliers from further analyses.

[19] If one has a strong suspicion that individuals from the same locality are likely pseudo-replicates (e.g., the locality is very precise, the collector and year of collection the same etc.), but he/she, nevertheless, wants to measure them, those multiple individuals could be averaged and their average be treated as a single observation in the sample of that taxon.

A3) Statistical power: an example using TPSPower

Methods (A3)

Statistical power is related to the probability of detecting an effect if the effect is real (Krzywinski & Altman 2013). It is also described as the probability of correctly rejecting the null hypothesis (no differences, in my case) when the alternative is true (differences are real). As with P values (see Appendix A on frequentist statistics), power is a frequency probability over very many repetitions of the same study design. However, it is not, and should not be confused with, the chance of being correct about the alternative hypothesis (Greenland *et al.* 2016; Greenland 2019). Power is never about the compatibility of the data with the alternative hypothesis. It is only related to the specific null hypothesis used in the model. Statistical power increases with effect and sample size, and is inversely related to within group variance (Quinn & Keough 2002). Therefore, in taxonomic analyses, power increases with the magnitude of mean differences relative to within group variation, and with the size of the samples being compared. More exactly, power increases with the square root of N. Therefore, to double power one needs a four-fold increase in N.

Power analyses can be done prospectively (a priori), to estimate the number of individuals required for a specific analysis to have adequate chances of finding differences of a certain magnitude. However, power analyses can also be done retrospectively (post hoc), to assess the observed power given the specific samples and their differences. Both are in theory possible using TPSPower (Rohlf 2015). Specifically, with this software, researchers can explore power in GMM using a simple design for pairwise tests of group mean shape differences. Put it simply, TPSPower simulates variation around mean shapes, tests the differences between simulated samples, repeat this many times, and finally counts how frequently tests are significant (given those mean differences and amount of simulated variance). If that happens most of the time, power is adequate (conventionally, the threshold is 0.8 or 80% significant tests); otherwise (> 20% of non-significant tests), power is low. Because power analyses are very rare in GMM, I will dedicate some more space to explain how to use TPSPower, a program which seems to have been cited only in a few orthodontic applications (Singh *et al.* 2005, 2007). In the Discussion, using the results of the power analysis in marmots as an example, I will go back to the theory and mainly focus on the distinction between types of power analysis, their pros and cons, and the type of information one might obtain from a power simulation. As the subject is vast and I only discuss a simple, specific, case of power analysis, readers interested to learn more should explore the extensive literature on statistical power. For a concise introduction, I suggest Krzywinski & Altman (2013) with multivariate examples in Hair *et al.* (2013).

TPSPower requires inputting the Procrustes mean shapes of the two groups to be compared. The computation of a mean shape can be done manually in a spreadsheet, where Procrustes shape coordinates of each group have been imported, but it is easily done also in other programs such as MorphoJ (*Preliminaries, Average Observations By*, in my case, species or, within species, sex) or TPSRelw (load data one group at a time, click *Consensus* for the superimposition, and save the mean from the menu *File, Save, Consensus*). TPSPower is limited to a single factor (e.g., sex) and only two groups at a time, and is specific for shape data. Thus, TPSPower cannot do power analyses for CS. For CS, readers can explore G*Power 3.1 (Faul *et al.* 2009). G*Power is user-friendly and provides a large variety of types of power analyses for univariate data, including pairwise t-tests for differences in group means, analogous to the tests of mean shape differences between two groups in TPSPower.

In the power analysis, I will focus on the three largest species samples, the hoary and yellow-bellied marmots, and the woodchuck. First, within each species, I estimate statistical power for the tests of shape SDM. Then, I do the same for pairwise comparisons of species mean shapes regardless of sex. In practice, if, for instance, I first want to compare *M. caligata* with *M. flaviventris*, I have to load their

mean shapes in TPSPower using two separate NTS files. For *M. caligata*, the text file with extension *.nts will be:
1 1 24 0 dim = 2
-0.35652 0.04791 -0.28426 0.02270 -0.16435 0.02322 -0.09131 0.01213 0.02506 0.01040 0.05126 0.00781 0.25116 0.14660 0.34010 0.07676 0.39027 0.00343 0.40680 -0.20559 -0.20691 -0.08417 -0.36133 -0.06119

The format, described in the software help, is also briefly exemplified in B2 using a larger rectangular data matrix. Briefly, the four numbers in the first line refer to: the type of matrix (one for a vector or a rectangular matrix); sample size (one, here, because it is a single mean shape); the number of variables (24 Procrustes shape coordinates); the absence of missing data (coded as zero). I added dim=2 to indicate that it is 2D data. Dim=2 (or dim=3 for 3D landmarks) is optional, but avoids the need to later define the dimensionality with a specific command in TPSPower. In the second line, I pasted the mean Procrustes shape coordinates of the hoary marmot exported as TXT file from Morphoj. Similarly, I created an NTS file with the mean shape of the yellow-bellied marmot.

Once two mean shapes are loaded in the program, the researcher sets the significance threshold (alpha), the number of iterations (i.e., how many times the test for mean differences will be repeated), and the size of the samples (N). For alpha, instead of the usual 0.05, I employed a more conservative 0.005 threshold, which is also used in all main tests (see part B and also the Appendix A on frequentist statistics). For the number of iterations, after experimenting a little with different values (e.g., 100, 1000, 5000), 1000 seemed a good compromise between computational time and precision. For N, there is, for now, a limitation in TPSPower, because sample size must be the same in both groups. Because taxonomic data rarely have such perfectly balanced samples, I employed as an approximation the average of the Ns of the pair of groups being compared. Thus, for instance, for hoary (N = 108) and yellow-bellied (N = 156) marmots, I set N = (108+156) / 2 = 132.

Using the observed sample size is appropriate for a retrospective power analysis (see Discussion). However, since I was also interested to indirectly explore power in the smallest samples, I repeated all power analyses using again the mean shapes of the woodchuck, hoary and yellow-bellied marmots, but now setting N to a lower value. The value I chose is N = 10, because it is close to the sample size of Alaskan and Olympic marmots, but also of VAN for within-species tests of SDM. This means that, in the N = 10 power tests, the largest samples are used as a proxy for the smallest ones. I could have used directly the Procrustes shape coordinates of VAN, Alaskan and Olympic marmots, but did not do it, because mean differences tend to be inflated in small samples (Cardini *et al.* 2021, and Discussion) and that leads to overestimate power.

The penultimate step, before running the simulation in TPSPower, is to tell the program how much within sample variance is simulated using an isotropic model. The isotropic model is a requirement for the Goodall's F test, used in TPSPower as well as in the ME ANOVA, where it was briefly mentioned. Because a random isotropic model requires the same amount of random variation in all directions ('circular variation') around each landmark, the model implies uncorrelated Procrustes shape coordinates. A complete lack of correlation in shape coordinates, however, is an unrealistic assumption for biological data, but, for now, this is the only option available in TPSPower. The variance is set by pressing the *Var/covar* button, where one manually writes a single standard deviation (SD) for the simulated variation of any of the, in my case, 24 Procrustes shape coordinates. SD should be realistic and mirror the expectation in a population. As an estimate of SD, I computed the average SD of the coordinates within each species and, then, averaged it across species samples, which produced an SD of 0.008 units of Procrustes shape distance. Estimates of SD were similar (not shown) if computed with separate sexes. Therefore, I employed the same SD in the power analysis of both within-species SDM

and interspecific mean differences. However, I computed also a second summary estimate of the SDs using the average of the 90th percentiles of within-species SDs, which produced a value of 0.012 units of Procrustes shape distance. Thus, the TPSPower simulation was done using either SD = 0.008 or SD = 0.012, with the lower value approximating the main trend in the data and the higher value trying to correct for a potential underestimate in a sample originating from a much larger population. Because the larger 0.012 SD produces more overlap between samples, rejecting the null hypothesis is harder and, therefore, this SD produces lower estimates of statistical power for a given N. In general, one can experiment with several different SDs to assess the sensitivity of the power analysis to the amount of within group variation. It is also easy to check that the SD used in TPSPower simulates approximately the same variation as in the real data. Using the function *=rnorm(0, SD)* in a spreadsheet, with SD = 0.008, for instance, a user can create isotropic normally distributed variance for N specimens and repeat the operation in as many columns as the number of shape coordinates (24, in my main marmot dataset). In the spreadsheet, these random normally distributed numbers can be added to one or the other of the two mean shapes used in TPSPower. This will be like a single run of simulated variation in TPSPower for one of the two groups in the comparison. The resulting matrix can be pasted in PAST and used for a PCA. Whether data are superimposed or not might make a very small difference, unless SD is large. One can try both and compare results, if in doubt. The PCA scatterplots should show an amount of differences among individuals about as large as in the real data for that species, even if the scatter will be more circular. Also, the total variance, estimated by the sum of the eigenvalues, which are the variances of the individual PCs, should be approximately the same as in the real data.

Finally, to perform the simulation in TPSPower, one has to check the box for the Goodall's F test. This test is analogous, although not exactly identical, to a test using Rsq, which is the one I chose in pairwise tests of mean shape differences (B1 and B3). When the analysis is run (press *Compute*), TPSPower simulates isotropic variance around each of the two mean shapes and tests the significance of their differences using the Goodall's F. This is repeated, in my analysis, 1000 times. When the simulation is over, the software reports the proportion of tests (out of 1000 simulations), in which the null hypothesis of no differences was rejected at the specified alpha (0.005, in my analysis). Because the two mean shapes belong to different samples and, thus, we know they are different, this proportion corresponds to the estimate of statistical power for those data.

Results (A3)

The results of the analysis of statistical power in pairwise tests of mean shape differences are shown in Table 6. I summarize first the interspecific analyses, whose outcome is simpler, as power is consistently high ($\geq$ 0.8) even when samples are small (N=10). The only minor exception is the comparison of yellow-bellied marmot and woodchuck mean shapes with N = 10 (small samples) and SD = 0.012 (large within-species variance), but even in this case power is very close to 0.8. Bearing in mind the limitations of TPSPower (see Methods and Discussion), the conclusion for the tests of the differences between two species mean shapes is that N is adequate virtually all the time, including comparisons of the species with the smallest samples (N $\approx$ 10) and despite the conservative alpha = 0.005.

The picture is more complex in the within-species power analysis for the tests of shape SDM. When samples are large (N > 40), power is about one with SD = 0.008 and between ~ 0.7 and 0.9 when SD = 0.012. Thus, even with larger within-species variance, power should be adequate for testing small sex differences in mean mandibular shape in yellow-bellied marmots (per sex N $\approx$ 70). Power is very high ($\geq$ 0.99) also in the smaller samples of ~ 40-45 individuals of the hoary marmots and woodchucks, when a moderate amount of within-species variation (SD = 0.008) is simulated. However, with more variation (SD = 0012), there is a modest but appreciable reduction in power in hoary marmots and woodchucks (power $\approx$ 0.7). Unsurprisingly, when the simulated sample size is even smaller (N = 10), power drops (range = 0.04-0.3 ca.) and it is, thus, too low for a meaningful test of shape SDM.

Table 6. Retrospective power analysis in the largest available samples. The following parameters are used in TPSPower: 1,000 iterations of Goodall's F test for mean shape differences (using the observed values) between balanced samples (group1 vs group2*) with simulated N = observed average of N1 and N2 or N = 10, as an approximation for the smallest samples; alpha = 0.005; variance simulated using an isotropic model with two summary estimates (see main text) of SD in the shape coordinates. Low power (< 0.8) is emphasized using a grey background.

Level	Group1 vs group2		Observed	Observed		Simulated	SD = 0.008	SD = 0.012	Simulated	SD = 0.008	SD = 0.012
			Shape distance	N1	N2	N	power	power	N	power	power
within	calF	calM	0.01353	41	40	41	0.996	0.669	10	0.274	0.073
species	flaF	flaM	0.01147	73	72	73	1.000	0.856	10	0.154	0.042
	monF	monM	0.01198	51	38	45	0.987	0.660	10	0.185	0.048
interspecific	cal	fla	0.03397	108	156	132	1.000	1.000	10	1.000	0.932
	cal	mon	0.04143	108	101	105	1.000	1.000	10	1.000	0.997
	fla	mon	0.02923	156	101	129	1.000	1.000	10	1.000	0.773

* If group2 was the same as group1, the estimated 'power' (proportion of rejections in 1000) would become the empirical estimate of the rate of type I errors, which should thus be approximately equal to the chosen alpha = 0.005. This is indeed the case (results not shown), as it would range from 0.004 to 0.008.

Discussion (A3)

Sensitivity to sample size and randomized subsampling experiments

Statistical power does not depend exclusively on sample size. However, sample size is the main parameter a taxonomist can try to influence to detect (or reject) group differences. On the other hand, as anticipated in the first paragraphs of the Discussion on outliers, the question of a desirable minimum sample size does not concern only statistical power, but also and more generally the accuracy of all results. For instance, one could have high power and significant results, but nevertheless over- or under-estimate the true amount of differences between taxa, because samples are not representative of the corresponding populations. I mentioned that several morphometric studies (Cardini *et al.* 2021, and references therein) suggest that, for relatively small amounts of interspecific variation in mammals, such as those typically found within a genus, several dozens of individuals might be necessary, in each group, in order to achieve a fairly precise quantitative description of means, variances and covariances of a morphospecies, as well as for taxonomic identification and testing differences with other species. However, there is no universal answer about adequate sample size and, as with the choice of the landmarks and ME, whenever possible, it is useful to explore potential problems directly in the context of one's own specific research settings. Because power and accuracy of the parameters estimated in a sample are both impacted by N, I use the discussion on power to also touch on the issue of the effect of sampling error on group comparisons.

Sensitivity analyses is one approach to explore the influence of sample size on parameter estimates. Sensitivity to N can be assessed in different ways. One can tackle the issue from a theoretical perspective and/or using simulations. Statistical modelling is usually rigorous, but rarely covers all cases. Thus, it is better than point estimates from empirical studies, but still leaves open the question of generalizability (i.e., going from internal to external validity). Also, statistical models make assumptions and findings are accurate only as long as the assumptions are realistic and verified. In practice, most taxonomists may feel uneasy with theoretical studies and simulations unless supported by a statistician. A potentially less

rigorous, but simpler and often complimentary approach is to empirically explore sampling error, as in Cardini *et al.* (2021) and most of the studies they cite. These authors performed randomized resampling experiments by extracting progressively smaller random samples from the groups with the largest sample size. Then, they used the resulting subsamples to replicate a number of estimates and analyses (from simple group means, variances and covariances to species separation and identification). Finally, they assessed the congruence of findings from random subsamples with those of the total sample in order to infer minimum sample sizes required for precisely replicating results. This type of empirical analysis assumes that results in largest samples are accurate, and their conclusions are applicable to other groups. Because we do not know if this is the case, randomized resampling experiments have limitations and mainly concern precision. They require cautious interpretations, but have the advantage of simplicity and allow to obtain useful clues on sampling error, instead of, as most common, completely ignoring the problem. In this study, I provide two brief examples of this empirical approach. One, in the Appendix A, explores the sensitivity of the estimate of mean female to male Procrustes shape distance in yellow-bellied marmots and its implications for interpreting SDM in the small samples of females and males of VIM, Olympic and Alaskan marmots. The other, in part B5, preliminarily investigates the robustness, in relation to sample size, of mean similarity relationships among marmot species.

Power analyses in taxonomy using GMM: statistical errors, relevant parameters and types of power analysis

Compared to the randomized subsampling experiments I briefly outlined in the previous subsection, a power analysis answers a different question. Its aim is not to approximate the minimum N for adequate precision in the sample estimates of means, variances and covariances, or other parameters. A power analysis explores how large samples must be to have good chances of detecting differences (i.e., rejecting the null hypothesis) in relation to their magnitude (e.g., the distance between two means) and the amount of variance in the samples. As explained in the Methods, a threshold which is arbitrary, but often considered adequate for statistical power, is 0.8. This means that there is an 80% chance of rejecting the null hypothesis, if differences are real (i.e., not just due to sampling error).

The importance of power analyses is acknowledged in all statistical manuals (e.g., Quinn & Keough 2002; Moore & McCabe 2005; Hair *et al.* 2013; Howell 2013) and statistical power is the subject of numerous introductory papers (e.g., Hoenig & Heisey 2001; Krzywinski & Altman 2013; Uttley 2019). However, power analyses remain uncommon in taxonomy and almost absent in GMM. This rarity depends on a propensity to overlook issues with sample size and power, but also on a relative paucity of software for estimating power. The complexity of power analyses, especially in multivariate statistics, contributes to the problem. Besides, power analyses typically require some a priori knowledge on the size of the effect being tested, but this type of information is rarely available in biodiversity studies of taxonomic and evolutionary differences.

A powerful statistical test, in morphometrics applied to taxonomic analysis, is one that finds true population differences using estimates from samples. There will virtually always be differences between two samples. Using a statistical test, we estimate if these differences are unlikely to simply represent a 'by-product' of sampling error. More precisely, the statistical test helps to understand probabilistically if the data in the samples are highly incompatible with what is expected if there are no differences. If P, the probability of the data assuming no group differences, is very low, one is reasonably confident that the null hypothesis is a poor model for the data and can, therefore, be rejected. The risk of an incorrect conclusion, however, is not zero.

In statistics, claiming differences which are not there is a type I error. In contrast, failing to find true differences is a type II error. In biomedicine, the first type of error is also referred to as a false positive (a patient diagnosed with a disease he/she does not have) and the second as a false negative (an undiagnosed

disease). A powerful test has a small chance of producing false negatives. More precisely, power equals one minus the rate of type II errors (i.e., false negatives). In taxonomy, a false positive might lead to an incorrect separation of groups based on spurious differences and might, thus, contribute to taxonomic inflation (Zachos 2018). A false negative, in contrast, reduces the chances of detecting unique components of biological diversity, with their potential relevance to conservation priorities and the preservation of the evolutionary potential of a lineage (Minelli 2019). A study without adequate power is also often a useless waste of time and money. Thus, ideally, a researcher wants a test that minimizes both type I and type II errors. Unfortunately, as clarified using an example in the next paragraph, there is an inverse relationship between the two types of error.

Statistical power increases with sample size, but also with the magnitude of between to within group differences and alpha, the significance threshold. It is intuitive that larger samples provide more confidence in a statistical result and also that it might be easier to reject the null hypothesis if differences are bigger. The relationship with alpha seems less obvious. It becomes clear if one considers that alpha (often 0.05 but, in this study, 0.005) is the risk, expressed as probability, one is ready to take of rejecting the null hypothesis by mistake (i.e., the "cost of rejections" or the "maximum tolerable type I error rate", see Greenland 2019: 108). For instance, if a test for mean differences has P = 0.03, that means that the probability of sampling error producing differences as large or larger than observed, if the null hypothesis was true, is no more than 3%. If we have chosen an alpha of 0.05, we are within the range of probabilities we consider acceptable in terms of the risk of claiming differences that are, in fact, an artefact of sampling error. A lower alpha reduces the risk of spurious claims, but also reduces power, because it makes it harder to reject the hypothesis of no differences. Thus, lowering alpha makes a test more conservative, whereas a higher alpha makes it more liberal. The choice of alpha may vary depending on the scientific question (Greenland *et al.* 2016; Benjamin *et al.* 2018). In taxonomy using GMM, like in most descriptive research on biological variation, one cannot change the magnitude of between and within group differences. In contrast, the scientist can decide to be on the conservative or liberal side of statistical testing, and thus use a lower or higher alpha to alter power. I chose a conservative threshold that makes tests less powerful, but might be appropriate using museum data, typically heterogeneous and often autocorrelated (Cardini 2020a). Although this is a separate problem, a lower alpha also offers some protection from the possibility that multiple tests in a study might increase the rate of type I errors (for a brief discussion on when this issue becomes particularly important see Armstrong (2014)).

Once the appropriate alpha has been decided, the most important parameter to play with, in order to increase statistical power in taxonomic comparisons, is sample size. Before embarking in a big project, and a large and expensive data collection, it is useful to have at least an approximate idea about how much data one needs in order to find group differences, if present. Prospective power analyses have the precise aim of helping researchers to better design a study by optimizing data collection. This type of power analysis, however, requires a good amount of knowledge on the expected group differences in relation to within group variation. In taxonomy, this knowledge is rarely available for some of the most interesting groups, which are likely to be lesser known lineages, for which specimens are poorly represented in museum collections or difficult (almost impossible for protected species) to collect in the field.

In morphometric studies of biodiversity, it is also hard, if not impossible, to anticipate how large differences have to be for being considered taxonomically relevant. Cryptic species, for instance, might show no differences or tiny ones. Yet, in these species, a very small amount of morphometric variation might be the clue that leads to the discovery of a deeper molecular divergence in significant evolutionary units (Tobias *et al.* 2010). Besides, the information on the magnitude of expected differences not only is inevitably specific to the taxonomic group, but also varies in GMM depending on the anatomical structure and the landmark configuration. Because Procrustes shape data cannot be compared unless

the same identical set of landmark is used on the same structure, one cannot easily use results from the literature (for instance, mean shape distances or group variances) to design a power analysis. Pilot studies, focusing on few groups for which data are already available or easier to collect, might help to explore power. Thus, a researcher could obtain estimates of mean differences and variances in close relatives of the taxa he/she wishes to study. Assuming that parameters are similar in closely related lineages, and maybe also trying a range of estimates in the vicinity of the values available for the pilot study, one can have useful indications on the size of the samples required to achieve adequate power.

Retrospective power analyses are easier than prospective ones, because they employ parameters (mainly, means and variances, for tests of group differences), which are estimated from data already collected. However, they are less useful, since the study has already been done, and are, thus, heavily criticized (Hoenig & Heisey 2001; Lenth 2001). Critics argue that, when a null hypothesis is rejected (i.e., the data suggest differences), a higher retrospective power cannot be interpreted as stronger evidence for true differences. This is because, if the hypothesis of no differences is rejected, it is a truism that power must be adequate. Critics also add that, in the opposite scenario, when one fails to reject the null hypothesis despite high power, the high statistical power does not provide conclusive evidence that differences are not there. More importantly, they maintain that, because there is an inverse relationship between the observed P value and power, high P values inevitably imply low power (Hoenig & Heisey 2001). Thus, as they are not independent, one cannot use the latter (power) to strengthen the evidence for the former (a significant or a non-significant P). Instead, Hoenig & Heisey (2001) and Lenth (2001) suggest to either perform prospective power analyses or use, retrospectively, confidence intervals for the test statistics (in our case, the observed mean difference). Confidence intervals do not concern power, but provide information on the magnitude of the uncertainties around estimates of parameters, such as mean differences.

A general lack of a priori knowledge unfortunately limits the use of prospective power analyses in GMM. For confidence intervals, currently available user-friendly GMM software do not offer many options for multivariate data, although the now discontinued IMP Series calculated confidence intervals as part of the output of various analyses (Zelditch *et al.* 2004). For these reasons, but also to exemplify the use of TPSPower, I opted for a power analysis which is inevitably retrospective, since data have already been collected and analysed. This is a limitation. Also, power is explored only in pairwise tests of group mean shape differences using an isotropic model in balanced samples. These are all real issues, but they are somewhat secondary in a didactic analysis. Indeed, the main aim here is to introduce a neglected topic using TPSPower and Procrustes shape data. With the difference that parameters (mean differences and SD) will have originated from previous research or a pilot study, the 'mechanics' will be similar in a much more useful prospective analysis.

In fact, the first part of the power analysis I performed in the largest samples, using approximately the observed N, is purely retrospective, but the second part, exploring the reduction of power if those samples were much smaller, has a 'prospective side'. This is discussed in the next subsection. The isotropic model and balanced pairwise design, in contrast, are for now unavoidable constraints that might reduce the predictive accuracy of power analyses in TPSPower. This limitation might be removed in future versions. Despite the current limitations, however, a careful interpretation of the results obtained in TPSPower can still be insightful.

Interpretation of power in the comparison of marmot mandible mean shapes

The retrospective power analysis in woodchucks, hoary and yellow-bellied marmots, using their average observed sample size, confirmed adequate power in all tests of shape variation (Table 6). Indeed, power was high not only in pairwise comparisons of interspecific differences, but also in within-species tests of SDM. The first observation is trivial, because interspecific differences are large (see part B) and, in

agreement with Hoenig & Heisey (2001), very low P values (< 0.0001 in all tests) must be associated with very high power. SDM, in contrast, is much smaller than between species variation and largely negligible in tests (A1 and B1-2), which leads to expectations of low power in apparent contradiction with the generally adequate power estimated by the simulations. Thus, the results of the power analysis for SDM require a closer examination.

The Rsq of shape SDM is, on average, ~ 1/6 of between species Rsq and P values are always well above the 0.005 alpha threshold (P approximately ≥ 0.02, see Table 1 of B1). Yet, estimated power in observed samples of the species with larger N (Table 6) is close to one, using the intermediate estimate of within group variation (SD = 0.008), and between 0.7 and 0.9, when larger variation is simulated (SD = 0.012). These findings seem, as anticipated, to contradict Hoenig & Heisey (2001). These authors demonstrate that power is 0.5 with P = 0.05, which implies that power must be > 0.5 if P < 0.05 but < 0.5 if P > 0.05. In woodchucks, hoary and yellow-bellied marmots, the average P value of shape SDM is 0.04 (B1, Table 2). 0.04 is very close to 0.05 and, therefore, according to Hoenig & Heisey, should lead to a statistical power just above 0.5, which is inadequate (< 0.8). Yet, for SDM tests of shape in these three species, predicted power using the simulations in TPSPower is between ~ 0.7 and one, which indicates almost adequate (close to 0.8) to very high power. This unexpected finding is even more surprising once we consider that alpha was 0.005 in both the tests and power simulations. A lower alpha should further reduce power by making rejections more difficult. Put it the other way round, because on average power is ~ 0.9 in the simulations, when within sex N > 40, almost all statistical tests should be significant at the 0.005 threshold. In contrast, none of the tests has P < 0.005, which suggests that power is overestimated in TPSPower. If so, why does this happen?

The most likely explanation is that the power analysis is not using the same precise test as in the within-species assessment of shape SDM. TPSPower uses Goodall's F, which is analogous to a ratio of unexplained to explained variance and should, thus, be largely equivalent to the Rsq used in MorphoJ as a test statistics for shape SDM (Table 1, B1). However, Goodall's F assumes isotropic variation, unlike the Rsq that takes the observed variance-covariance structure into account. Consistently with the assumption of Goodall's F, the data generated by TPSPower have isotropic variation, instead of using the variance-covariance structure of the real data, as in the shape SDM regressions in MorphoJ. Also, Goodall's F test in TPSPower is tested parametrically, whereas MorphoJ uses permutations for the Rsq. Generally, when their assumptions are met, parametric tests tend to be more powerful than resampling statistics. Indeed, if shape SDM tests are repeated in TPSRegr, which provides both parametric and permutational P values for the Goodall's F, the reason for the incongruence between results of the TPSPower analysis and the shape SDM tests in MorphoJ becomes clear. Permutational Ps in TPSRegr using Goodall's F are similar for all three species to those obtained in MorphoJ using Rsq, which supports the equivalence of the test statistics when tested using resampling methods. In contrast, when the test is parametric, TPSRegr Goodall's Ps are lower (hoary marmot P = 0.0011; yellow-bellied marmot P = 0.0003; woodchuck P = 0.0189). This means that, when the test is parametric in all analyses (power simulations and SDM tests), the discrepancy disappears almost completely: P becomes significant (< 0.005) in 2/3 of the SDM tests of shape, and this corresponds fairly well to the predictions of high power in TPSPower.

This example is an important reminder that power should be estimated using exactly the same model as in the test of group differences. Ideally, I should have estimated power by simulating Rsq using permutational tests, but this is not possible in TPSPower. However, if one bears in mind that P values should be interpreted as a continuum of probabilities indicating the compatibility of the data with the null hypothesis (Appendix A on frequentist statistics), the results of parametric and permutational tests for shape SDM lead to a broadly similar conclusion: SDM is small (Rsq < 3%), but detectable and close to significance (or significant using more liberal alpha thresholds) in large samples (N > 40), a result which is in fair agreement with TPSPower prediction of adequate statistical power. As I show in B1-2,

the marginally significant, small, effect size of shape SDM suggests subtle differences between females and males. Although testing groups was not the main aim of the ME ANOVA, its results (A1, Table 5) also suggested that shape SDM is present but negligible in species comparisons, whose Rsq is about 20 times larger than that of SDM.

Up to this point, the power analysis I have discussed is purely retrospective, as it is based on observed N and mean shape differences. The 'prospective side' employs the same estimates of interspecific mean differences and within group variance to predict power when sample size is much smaller (N = 10). This means simulating tests of SDM and interspecific mean shape differences in small samples of woodchucks, hoary and yellow-bellied marmots, as a proxy for the other three species, with their smaller samples. Thus, the aim is to check if group differences in mean shapes have adequate power in Alaskan, Olympic marmots, and, for SDM analyses, also VAN. In these comparisons, the average N per group is nine, although N ranges, depending on the test, from as few as three to five individuals (SDM in Alaskan marmots) to ~ 15 (interspecific comparison of Alaskan and Olympic marmots). Thus, exploring power with N = 10 is useful to better understand the potential limitations of tests in small samples of North American marmots in this study, but also in future studies of within and between species variation in Eurasiatic species. Such small N are almost inevitably problematic (Cardini *et al.* 2021), but not uncommon in GMM (Cardini *et al.* 2015).

The simulations with N = 10 indicated that power may, indeed, be adequate (range ~ 0.8-1.0) in interspecific tests of mandibular mean shape differences. In contrast, however, TPSPower unequivocally suggested that N = 10 is too small to detect shape SDM (power < 0.3). In agreement with these predictions, in B3, I show that interspecific comparisons of mandibular shape involving Olympic and Alaskan marmots are always significant despite their small samples, whereas, in these two species and also in VAN, tests of mean shape differences between females and males never reach significance (B1). That power is too low to test shape SDM in small samples is unsurprising, because SDM in marmots is very modest, as anticipated in this paper and further discussed in part B. In contrast, it is reassuring that power is adequate for interspecific comparisons even when N = 10 and that, therefore, significant shape differences between, for instance, Alaskan and Olympic marmots (observed N ≈ 10, see part B) are unlikely to be a false positive, caused by inflated distances between inaccurate estimates of means in small samples.

The tendency to overestimate differences between group means when N is small should be always borne in mind. Why does it happen? I answer the question in this paragraph, but I will use randomized subsampling experiments to provide, in Appendix A, an example that makes the effect of N on estimates of means more tangible. In the marmot dataset, the propensity to inflated mean differences in small samples is easier to understand by focusing on shape SDM, because it is where sample sizes are smallest and the effect more evident. Indeed, if carefully examined, a total lack of significant SDM ($P > 0.3$ both using permutations or parametric Goodall's F tests, see B1) in VAN, Olympic and Alaskan marmots is almost counter-intuitive: tests are not significant in these three species (Table 1, in B1), but this occurs despite their Rsq being on average four times larger than in woodchucks, hoary and yellow-bellied marmots, whose tests are, nonetheless, marginally significant ($0.1 > P > 0.01$). It, thus, seems that non-significance is associated to large mean SDM in VAN, Olympic and Alaskan marmots, and the opposite happens in woodchucks, hoary and yellow-bellied marmots, with smaller SDM but quasi-significant tests. The apparent incongruence is expected, in fact, because Rsq is simple to compute and interpret, but it is a positively biased estimator in small samples (Cramer 1987). Thus, Rsq tends to be overestimated when N is small. The overestimate occurs because, even if sample means are unbiased, they have larger uncertainties in smaller samples (Wainer 2007). With the larger uncertainty, it becomes more likely that the mean of a species with a small sample happens to be further apart not only from its true population mean, but also from the mean of other species. The distance between the two means, thus, tends to

increase, on average, as N decreases. To put it simply, if a mean is based on just a few individuals, it is more likely for the mean to 'pick up' some of the unusual features of those individuals, thus inflating mean differences. In contrast, in large samples, the small differences that make each individual in a sample unique are averaged out, which tends to reduce the distance to the means of other groups. It is precisely to avoid using overestimated differences in the power analysis that I did not use the observed mean shapes of VAN, Olympic and Alaskan marmots.

Results would have, indeed, been different with a 'naive' application of a purely retrospective power analysis in VAN, Alaskan and Olympic marmots. For shape SDM, for instance, using the observed mean differences of the small samples of the Alaskan, Olympic and Vancouver Island marmots, estimates of SDM are likely inflated. Their mean shape distances are, on average, 70% larger than in woodchucks, hoary and yellow-bellied marmots, despite the literature typically reporting similar levels of SDM across marmot species (see B1-2 discussions). In a simulation, using larger average differences, assuming a similar amount of within group variation, will increase power. However, the increase is spurious if simply caused by inaccurate estimates of mean shapes which lead to overestimate differences. As an extreme case, in the Alaskan marmot, which has the smallest samples of all species, the statistical power of the SDM test in a retrospective analysis using its observed female and male means in TPSPower would have been (respectively with SD = 0.008 and SD = 0.012): 1.000 and 0.938 using N = 10, and 0.804 and 0.249 with N = 4, which is the average number of females and males in the samples of this species. These estimates of power are about four to 16 times larger than those obtained using the means of the three largest samples for N = 10 (Table 6). Thus, we might have concluded that N = 4 is too small for testing sex in Alaskan marmots (in fair agreement with the non-significant SDM test in part B), but also that N = 10 would be appropriate, which is misleading. This is because the power analysis using observed, and likely inflated, mean differences in very small samples of Alaskan marmots is inaccurate. In contrast, simulating small sample size prospectively, using parameters from species with large samples, suggests that a small effect, such as shape SDM in marmot mandibles, cannot be accurately tested with N = 10 and generally requires several dozens of individuals in each sex to be detected. This conclusion is more accurate and likely to be valid in other marmot species, whose SDM in structural morphological characters tends to be similar (Cardini 2003; Matějů & Kratochvíl 2013).

Conclusions

Preliminary analyses are easily overlooked, but the assessment of ME (measurement error) and the effect of potential outliers are essential steps in any analysis. GMM is no exception. Despite their importance, however, ME is not reported in a large number of GMM studies and outliers are mentioned even less often. Statistical power is also neglected, with only a handful of studies including multivariate power analyses of Procrustes shape data (e.g., Gharaibeh 2005; Singh *et al.* 2005, 2007). Yet, can we trust results without any idea of the magnitude of ME and the presence of outliers? A large ME reduces statistical power, introduces inaccuracies and may bias results. Outliers also can make findings inaccurate and biased. Besides, errors and outliers might be present not only in CS and shape, but also in grouping factors or covariates, whose accuracy must be checked.

I suggested a series of steps to estimate the impact of ME and detect potential outliers. For reducing ME, highly imprecise landmarks can be identified and removed. Using replicates, the magnitude of ME can be estimated and compared to the magnitude of individual variation at the specific level one is interested in. Although I exemplified this type of ME analysis considering only landmark digitization error, I stressed more than once that the same approach can be used for other sources of error. In fact, replicates should be carefully designed so that they incorporate all, or at least the most important, factors that can introduce ME (Arnqvist & Martensson 1998). Long-time lags during the data collection, for instance, may reduce precision and easily introduce systematic errors. Even if the methods I suggested do not assess directionality in ME, I have explained in the Discussion how to extend the analyses to

explore biases in relation to a specific study question. In general, if the magnitude of ME is very small, both random and systematic error are unlikely to have an impact. Nevertheless, the effect of ME must be related to the research questions and errors, which are negligible for a specific study aim, may not be negligible in a different context. For taxonomic studies of gonochoric species, the most important level of biological variation is group (populations, subspecies or species) differences. However, assessing ME in relation to variation within groups (e.g., controlling for both species and sex differences, as in the current study), even if overcautious, increases the confidence that ME is truly small and, thus, negligible in tests of between group differences, as well as in a range of within group analyses one may need to perform (e.g., sexual dimorphism or allometry).

I have also emphasized and exemplified how graphical analyses are as useful as numerical ones. The importance of plotting the data should be borne in mind at every step of a study. It is helpful for assessing ME and central to the methods I used to detect outliers. For outlier detection, there is a large number of alternatives to the relatively simple and mostly graphical analysis I have adopted. All methods, however, incorporate a degree of arbitrariness and, in small samples, there are inevitable limitations, such that outlier detection may be hard or impossible. More generally, small samples, heterogeneous sample size and biased sampling (in space or time, because of autocorrelations etc.) can be a serious problem in taxonomic studies.

If problems with sampling cannot be mitigated, they must be acknowledged and, when possible, their effect on results explored. To this aim, sensitivity analyses excluding the smallest samples and randomized subsampling experiments offer a simple tool for obtaining at least some clues on inaccuracies in small samples. In multivariate analyses, when samples are small and the number of variables large, as common in GMM especially with semilandmarks, analytical problems usually become more serious (see Rohlf 2021, and references therein). In studies of group differences, highly unfavourable p / N ratios might even create artefacts (Bookstein 2017; Björklund 2019; Cardini 2019; Cardini *et al.* 2019; Rohlf 2021). Power analyses, using parameters from larger samples of related taxa, help to predict an adequate sample size for future studies, and also contribute to understand the limitations of analyses in small samples. Even if, for now, user-friendly programs have few options for resampling experiments and power analyses, I have shown that there is the possibility of preliminarily examining some of the consequences of small sample size on results. Despite the constraints, user-friendly software allows a fairly deep investigation of potential issues with ME, outliers and sample size. For those with programming experience, it is easy to implement, improve and expand all the analyses I suggested using a statistical environment such as R.

The protocol I have described is one among many alternatives. Regardless of how they are done, I hope to have persuaded users that preliminary analyses (ME, outliers and, possibly also power) are not optional. They must be done to improve robustness and accuracy. If results of taxonomic comparisons using GMM are flawed, because of ME, outliers and/or inadequate sampling were overlooked, they might provide misleading conclusions, contribute to taxonomic inflation or fail to detect potentially important information to explore the evolutionary significance of poorly studied natural populations.

Acknowledgements

The twin papers (part A and B) are dedicated to the memory of my father Alberto (1939-2022) and of my friend and colleague Luigi Sala (1954–2022). With my dad, I first saw marmots, as a young child, in the Alps: I owe him much, including my love for the mountains and their fauna and flora. With Luigi, together with Paolo Tongiorgi (1936–2018), I began studying marmots in the Apennines: it was a short but great time that started a long friendship, which taught me more than Luigi was aware of. I miss my father's exclamation of wonder at the beauty of nature and Luigi's explosive laugh at my rude jokes. I am also in debt to many people who, over the years, have helped with data, suggestions, discussions (and, sometimes, disagreement!) on specific topics, as well as with advice on references. Among them

(and in alphabetical order), a special thanks to: Göran Arnqvist, Amro Daboul, Joe Felsenstein, Dan Franklin, Øyvind Hammer, Ardern Hulme-Beaman, Paula Jenkins (and all museum curators I met working on marmots and other mammals), Philipp Mitteroecker, Dave Nagorsen, David Polly, Riccardo Poloni, Jim Rohlf, Tim Smith and Davide Tamagnini. I am very much in debt also with Frank Zachos, Kristiaan Hoedemakers and Fabio Cianferoni, for their support and outstanding work as editors. And I thank again Frank Zachos, as well as Vida Jojić and two anonymous referees for reviewing so carefully the papers, despite their length. It was a huge task and I sincerely appreciate the work they did and the many good suggestions that improved the articles. The study was supported by the Fondo Ateneo di Ricerca (project TAXON), a grant of the Italian Ministero dell'Università e della Ricerca (PRIN Project 2022MAM9ZB), and stimulated by countless productive discussions with colleagues during many morphometric workshops, as well as by the always fruitful interactions with museum curators and the amazing team of SYNTHESYS (Synthesys of Systematic Resources: https://www.synthesys.info/ and https://www.dissco.eu/synthesys/).

References

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 (4): 393–399. https://doi.org/10.1111/2041-210X.12035

Adams D.C., Rohlf F.J. & Slice D.E. 2004. Geometric morphometrics: ten years of progress following the 'revolution'. *Italian Journal of Zoology* 71 (1): 5–16.

Adams D.C., Rohlf F.J. & Slice D.E. 2013. A field comes of age: geometric morphometrics in the 21st century. *Hystrix* 24 (1): 7–14. https://doi.org/10.4404/hystrix-24.1-6283

Albrecht G. 1992. Assessing the affinities of fossils using canonical variates and generalized distances. *Human Evolution* 7 (4): 49–69.

Arbour J.H. & Brown C.M. 2014. Incomplete specimens in geometric morphometric analyses. *Methods in Ecology and Evolution* 5 (1): 16–26. https://doi.org/10.1111/2041-210X.12128

Armendáriz-Toledano F., López-Posadas M.A., Utrera-Vélez Y., Nápoles J.R. & Castro-Valderrama U. 2023. More than 80 years without new taxa: analysis of morphological variation among members of Mexican *Aeneolamia* Fennah (Hemiptera, Cercopidae) support a new species in the genus. *ZooKeys* 1139: 71–106. https://doi.org/10.3897/zookeys.1139.85270

Armitage K.B. 2000. The evolution, ecology, and systematics of marmots. *Oecologia Montana* 9 (1–2): 1–18.

Armitage K.B. 2014. *Marmot Biology: Sociality, Individual Fitness, and Population Dynamics*. Cambridge University Press, Cambridge UK.

Armstrong R.A. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics* 34 (5): 502–508. https://doi.org/10.1111/opo.12131

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 (1–2): 73–96.

Bastir M., O'Higgins P. & Rosas A. 2007. Facial ontogeny in Neanderthals and modern humans. *Proceedings of the Royal Society B: Biological Sciences* 274 (1614): 1125–1132. https://doi.org/10.1098/rspb.2006.0448

Benjamin D.J., Berger J.O., Johannesson M., Nosek B.A., Wagenmakers E.-J., Berk R., Bollen K.A., Brembs B., Brown L., Camerer C., Cesarini D., Chambers C.D., Clyde M., Cook T.D., De Boeck P., Dienes Z., Dreber A., Easwaran K., Efferson C., Fehr E., Fidler F., Field A.P., Forster M., George E.I.,

Gonzalez R., Goodman S., Green E., Green D.P., Greenwald A.G., Hadfield J.D., Hedges L.V., Held L., Hua Ho T., Hoijtink H., Hruschka D.J., Imai K., Imbens G., Ioannidis J.P.A., Jeon M., Jones J.H., Kirchler M., Laibson D., List J., Little R., Lupia A., Machery E., Maxwell S.E., McCarthy M., Moore D.A., Morgan S.L., Munafó M., Nakagawa S., Nyhan B., Parker T.H., Pericchi L., Perugini M., Rouder J., Rousseau J., Savalei V., Schönbrodt F.D., Sellke T., Sinclair B., Tingley D., Van Zandt T., Vazire S., Watts D.J., Winship C., Wolpert R.L., Xie Y., Young C., Zinman J. & Johnson V.E. 2018. Redefine statistical significance. *Nature Human Behaviour* 2 (1): 6–10. https://doi.org/10.1038/s41562-017-0189-z

Björklund M. 2019. Be careful with your principal components. *Evolution* 73 (10): 2151–2158. https://doi.org/10.1111/evo.13835

Blumstein D. 1999. Alarm calling in three species of marmots. *Behaviour* 136 (6): 731–757. https://doi.org/10.1163/156853999501540

Bonhomme V., Picq S., Gaucherel C. & Claude J. 2014. Momocs: Outline Analysis Using R. Available from https://cran.r-project.org/web/packages/Momocs/index.html [accessed 5 Apr. 2024].

Bookstein F.L. 2017. A newly noticed formula enforces fundamental limits on geometric morphometric analyses. *Evolutionary Biology* 44 (4): 522–541. https://doi.org/10.1007/s11692-017-9424-9

Brown C.M. & Vavrek M.J. 2015. Small sample sizes in the study of ontogenetic allometry; implications for palaeobiology. *PeerJ* 3: e818. https://doi.org/10.7717/peerj.818

Cardini A. 2003. The geometry of the marmot (Rodentia: Sciuridae) mandible: phylogeny and patterns of morphological evolution. *Systematic Biology* 52 (2): 186–205. https://doi.org/10.1080/10635150390192807

Cardini A. 2013. *Geometric Morphometrics.* Biological Science Fundamental and Systematics. UNESCO, Encyclopedia of Life Support Systems (EOLSS), Oxford, UK.

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, the Italian Journal of Mammalogy* 25 (2): 73–81. https://doi.org/10.4404/hystrix-25.2-10993

Cardini A. 2017. Left, right or both? Estimating and improving accuracy of one-side-only geometric morphometric analyses of cranial variation. *Journal of Zoological Systematics and Evolutionary Research* 55 (1): 1–10. https://doi.org/10.1111/jzs.12144

Cardini A. 2019. Integration and modularity in Procrustes shape data: is there a risk of spurious results? *Evolutionary Biology* (46): 90–105. https://doi.org/10.1007/s11692-018-9463-x

Cardini A. 2020a. Modern morphometrics and the study of population differences: Good data behind clever analyses and cool pictures? *The Anatomical Record* 303 (11): 2747–2765. https://doi.org/10.1002/ar.24397

Cardini A. 2020b. Less tautology, more biology? A comment on "high-density" morphometrics. *Zoomorphology* 139 (4): 513–529. https://doi.org/10.1007/s00435-020-00499-w

Cardini A. & Chiapelli 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. & Elton S. 2007. Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology* 126 (2): 121–134. https://doi.org/10.1007/s00435-007-0036-2

Cardini A. & Elton S. 2008a. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. *Biological Journal of the Linnean Society* 93 (4): 813–834. https://doi.org/10.1111/j.1095-8312.2008.01011.x

Cardini A. & Elton S. 2008b. Variation in guenon skulls (II): sexual dimorphism. *Journal of Human Evolution* 54 (5): 638–647.

Cardini A. & O'Higgins P. 2005. Post-natal ontogeny of the mandible and ventral cranium in *Marmota* species (Rodentia, Sciuridae): allometry and phylogeny. *Zoomorphology* 124 (4): 189–203. https://doi.org/10.1007/s00435-005-0008-3

Cardini A. & Tongiorgi P. 2003. Yellow-bellied marmots (*Marmota flaviventris*)'in the shape space'(Rodentia, Sciuridae): sexual dimorphism, growth and allometry of the mandible. *Zoomorphology* 122 (1): 11–23.

Cardini A. & Verderame M. 2022. Procrustes shape cannot be analyzed, interpreted or visualized one landmark at a time. *Evolutionary Biology* 49 (2): 239–254. https://doi.org/10.1007/s11692-022-09565-1

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 (3): 258–268. https://doi.org/10.1111/j.1439-0469.2005.00316.x

Cardini A., Thorington R.W. & 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 (5): 1833–1846. https://doi.org/10.1111/j.1420-9101.2007.01398.x

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 & Evolution* 21 (3): 209–223. https://doi.org/10.1080/08927014.2009.9522476

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 (2): 149–163. https://doi.org/10.1007/s00435-015-0253-z

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 (4): 303–316. https://doi.org/10.1007/s11692-019-09487-5

Cardini A., Elton S., Kovarovic K., Strand Viđarsdóttir U. & Polly P.D. 2021. On the misidentification of species: sampling error in primates and other mammals using geometric morphometrics in more than 4000 individuals. *Evolutionary Biology* 48 (2): 190–220. https://doi.org/10.1007/s11692-021-09531-3

Cardini A., de Jong Y.A. & Butynski T.M. 2022. Can morphotaxa be assessed with photographs? Estimating the accuracy of two-dimensional cranial geometric morphometrics for the study of threatened populations of African monkeys. *The Anatomical Record* 305 (6): 1402–1434. https://doi.org/10.1002/ar.24787

Caumul R. & Polly P.D. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). *Evolution* 59 (11): 2460–2472. https://doi.org/10.1111/j.0014-3820.2005.tb00955.x

Claude J. 2008. *Morphometrics with R*. Springer Verlag, New York, USA.

Colegrave N. & Ruxton G.D. 2018. Using biological insight and pragmatism when thinking about pseudoreplication. *Trends in Ecology & Evolution* 33 (1): 28–35. https://doi.org/10.1016/j.tree.2017.10.007

Cook C.N., Freeman A.R., Liao J.C. & Mangiamele L.A. 2021. The philosophy of outliers: reintegrating rare events into biological science. *Integrative and Comparative Biology* 61 (6): 2191–2198. https://doi.org/10.1093/icb/icab166

Cooke S.B. & Terhune C.E. 2015. Form, function, and geometric morphometrics. *The Anatomical Record* 298 (1): 5–28. https://doi.org/10.1002/ar.23065

Cope D.A. 1993. Measures of dental variation as indicators of multiple taxa in samples of sympatric *Cercopithecus* species. *In*: Kimbel W.H. & Martin L.B. (eds) *Species, Species Concepts and Primate Evolution*: 211–237. Springer US, Boston, MA.

Cope D.A. & Lacy M.G. 1992. Falsification of a single species hypothesis using the coefficient of variation: a simulation approach. *American Journal of Physical Anthropology* 89 (3): 359–378. https://doi.org/10.1002/ajpa.1330890309

Corti M. 1993. Geometric morphometrics: An extension of the revolution. *Trends in Ecology & Evolution* 8 (8): 302–303. https://doi.org/10.1016/0169-5347(93)90261-M

Corti M., Fadda C., Simson S. & Nevo E. 1996. Size and shape variation in the mandible of the fossorial rodent *Spalax ehrenbergi*. *In*: Marcus L.F., Corti M., Loy A., Naylor G.J.P. & Slice D.E. (eds) *Advances in Morphometrics*: 303–320. Springer US, Boston, MA.

Craig J.M., Crampton W.G. & Albert J.S. 2017. Revision of the polytypic electric fish *Gymnotus carapo* (Gymnotiformes, Teleostei), with descriptions of seven subspecies. *Zootaxa* 4318 (3): 401–438. https://doi.org/10.11646/zootaxa.4318.3.1

Cramer J.S. 1987. Mean and variance of R2 in small and moderate samples. *Journal of Econometrics* 35 (2): 253–266. https://doi.org/10.1016/0304-4076(87)90027-3

Daboul A., Ivanovska T., Bülow R., Biffar R. & Cardini A. 2018. Procrustes-based geometric morphometrics on MRI images: An example of inter-operator bias in 3D landmarks and its impact on big datasets. *PLoS One* 13 (5): e0197675. https://doi.org/10.1371/journal.pone.0197675

Daboul A., Krüger M., Ivanovska T., Obst A., Ewert R., Stubbe B., Fietze I., Penzel T., Hosten N., Biffar R. & Cardini A. 2023. Do brachycephaly and nose size predict the severity of obstructive sleep apnea (OSA)? A sample-based geometric morphometric analysis of craniofacial variation in relation to OSA syndrome and the role of confounding factors. *Journal of Sleep Research* 32 (3): e13801. https://doi.org/10.1111/jsr.13801

Daszykowski M., Kaczmarek K., Vander Heyden Y. & Walczak B. 2007. Robust statistics in data analysis — A review: basic concepts. *Chemometrics and Intelligent Laboratory Systems* 85 (2): 203–219. https://doi.org/10.1016/j.chemolab.2006.06.016

Dayrat B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* 85 (3): 407–417. https://doi.org/10.1111/j.1095-8312.2005.00503.x

Dehon M., Engel M.S., Gérard M., Aytekin A.M., Ghisbain G., Williams P.H., Rasmont P. & Michez D. 2019. Morphometric analysis of fossil bumble bees (Hymenoptera, Apidae, Bombini) reveals their taxonomic affinities. *ZooKeys* 891: 71–118. https://doi.org/10.3897/zookeys.891.36027

Duarte L.C., Monteiro L.R., von Zuben F.J. & Dos Reis S.F. 2000. Variation in mandible shape in *Thrichomys apereoides* (Mammalia: Rodentia): geometric analysis of a complex morphological structure. *Systematic Biology* 49 (3): 563–578. https://doi.org/10.1080/10635159950127394

Evans A.R. 2013. Shape descriptors as ecometrics in dental ecology. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 133-140. https://doi.org/10.4404/hystrix-24.1-6363

Evenhuis N.L. 2007. Helping solve the "other" taxonomic impediment: completing the eight steps to total enlightenment and taxonomic nirvana. *Zootaxa* 1407 (3–12): 67–68.

Faul F., Erdfelder E., Buchner A. & Lang A.-G. 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods* 41 (4): 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Felsenstein J. 2004. *Inferring Phylogenies*. Sinauer Associates, Incorporated, Sunderland, Massachusetts.

Fox J. & Weisberg S. 2019. *An R Companion to Applied Regression*. Third. Sage, Thousand Oaks, CA.

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 Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology* 275A (2): 1048–1072. https://doi.org/10.1002/ar.a.10112

Fruciano C. 2016. Measurement error in geometric morphometrics. *Development Genes and Evolution* 226 (3): 139–158. https://doi.org/10.1007/s00427-016-0537-4

Galimberti F., Sanvito S., Vinesi M.C. & Cardini A. 2019. "Nose-metrics" of wild southern elephant seal (*Mirounga leonina*) males using image analysis and geometric morphometrics. *Journal of Zoological Systematics and Evolutionary Research* 57 (3): 710–720. https://doi.org/10.1111/jzs.12276

Gharaibeh W. 2005. Correcting for the effect of orientation in geometric morphometric studies of side-view images of human heads. *In*: Slice D.E. (ed.) *Modern Morphometrics in Physical Anthropology*: 117–143. Springer US, Boston, MA.

Giangrande A. 2003. Biodiversity, conservation, and the 'Taxonomic impediment'. *Aquatic Conservation: Marine and Freshwater Ecosystems* 13 (5): 451–459. https://doi.org/10.1002/aqc.584

Goodall C. 1991. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)* 53 (2): 285–321. https://doi.org/10.1111/j.2517-6161.1991.tb01825.x

Gower J.C. 1975. Generalized Procrustes analysis. *Psychometrika* 40 (1): 33–51. https://doi.org/10.1007/BF02291478

Greenland S. 2019. Valid P-values behave exactly as they should: some misleading criticisms of p-values and their resolution with S-values. *The American Statistician* 73 (sup1): 106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland S., Senn S.J., Rothman K.J., Carlin J.B., Poole C., Goodman S.N. & Altman D.G. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31 (4): 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grossnickle D.M. 2017. The evolutionary origin of jaw yaw in mammals. *Scientific Reports* 7 (1): 45094. https://doi.org/10.1038/srep45094

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 (1): 48–62. https://doi.org/10.1016/j.jhevol.2009.04.004

Gutierrez B.L., MacLeod N. & Edgecombe G. 2011. Detecting taxonomic signal in an under-utilised character system: geometric morphometrics of the forcipular coxae of Scutigeromorpha (Chilopoda). *ZooKeys* 156: 49–66. https://doi.org/10.3897/zookeys.156.1997

Hair J.F., Black W.C., Babin B.J. & Anderson R.E. 2013. *Multivariate Data Analysis*. Pearson Education Limited.

Hammer O., Harper D. & Ryan P. 2001. PAST: Paleontological statistics software package for education and data analysis. *Paleontologica Electron* 4 (1): 1–9.

Hawkins D.M. 1980. *Identification of Outliers*. Springer Netherlands, Dordrecht.

Hendrichs J., Vera M.T., De Meyer M. & Clarke A.R. 2015. Resolving cryptic species complexes of major tephritid pests. *ZooKeys* (540): 5–39. https://doi.org/10.3897/zookeys.540.9656

Herron M.D., Castoe T.A. & Parkinson C.L. 2004. Sciurid phylogeny and the paraphyly of Holarctic ground squirrels (*Spermophilus*). *Molecular Phylogenetics and Evolution* 31 (3): 1015–1030. https://doi.org/10.1016/j.ympev.2003.09.015

Hoenig J.M. & Heisey D.M. 2001. The abuse of power. *The American Statistician* 55 (1): 19–24. https://doi.org/10.1198/000313001300339897

Howell D.C. 2013. *Statistical Methods for Psychology (Eight Edition)*. Wadsworth Cengage Learning, Wadsworth, USA.

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 (6): 777–785. https://doi.org/10.1016/j.jhevol.2009.09.001

Hulme-Beaman A., Claude J., Chaval Y., Evin A., Morand S., Vigne J.D., Dobney K. & Cucchi T. 2019. Dental shape variation and phylogenetic signal in the Rattini tribe species of mainland Southeast Asia. *Journal of Mammalian Evolution* 26 (3): 435–446. https://doi.org/10.1007/s10914-017-9423-8

Huson D.H. & Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* 61 (6): 1061–1067. https://doi.org/10.1093/sysbio/sys062

Jojić V., Bugarski-Stanojević V., Blagojević J. & Vujošević M. 2014. Discrimination of the sibling species *Apodemus flavicollis* and *A. sylvaticus* (Rodentia, Muridae). *Zoologischer Anzeiger - A Journal of Comparative Zoology* 253 (4): 261–269. https://doi.org/10.1016/j.jcz.2014.02.002

Kangas A.T., Evans A.R., Thesleff I. & Jernvall J. 2004. Nonindependence of mammalian dental characters. *Nature* 432 (7014): 211–214. https://doi.org/10.1038/nature02927

Karagic N., Meyer A. & Hulsey C.D. 2020. Phenotypic plasticity in vertebrate dentitions. *Integrative and Comparative Biology* 60 (3): 608–618. https://doi.org/10.1093/icb/icaa077

Kelt D.A. & Patton J.L. 2020. *A Manual of the Mammalia: An Homage to Lawlor's "Handbook to the Orders and Families of Living Mammals"*. University of Chicago Press, Chicago, USA.

Kendall D.G. 1989. A Survey of the Statistical Theory of Shape. *Statistical Science* 4 (2): 87–99.

Kenyon-Flatt B., Conaway M.A., Lycett S.J. & von Cramon-Taubadel N. 2020. The relative efficacy of the cranium and os coxa for taxonomic assessment in macaques. *American Journal of Physical Anthropology* 173 (2): 350–367. https://doi.org/10.1002/ajpa.24100

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 (4): 810–826. https://doi.org/10.1093/jmammal/gyv089

Klingenberg C.P. 2008. Novelty and "homology-free" morphometrics: What's in a Name? *Evolutionary Biology* 35 (3): 186–190. https://doi.org/10.1007/s11692-008-9029-4

Klingenberg C.P. 2011. MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources* 11 (2): 353–357. https://doi.org/10.1111/j.1755-0998.2010.02924.x

Klingenberg C.P. 2013. Visualizations in geometric morphometrics: how to read and how to make graphs showing shape changes. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 15–24. https://doi.org/10.4404/hystrix-24.1-7691

Klingenberg C.P. 2016. Size, shape, and form: concepts of allometry in geometric morphometrics. *Development Genes and Evolution* 226 (3): 113–137. https://doi.org/10.1007/s00427-016-0539-2

Klingenberg C.P. 2020. Walking on Kendall's shape space: understanding shape spaces and their coordinate systems. *Evolutionary Biology* 47 (4): 334–352. https://doi.org/10.1007/s11692-020-09513-x

Klingenberg C.P. 2022. Methods for studying allometry in geometric morphometrics: a comparison of performance. *Evolutionary Ecology* 36 (4): 439–470. https://doi.org/10.1007/s10682-022-10170-z

Klingenberg C.P. & Leamy L.J. 2001. Quantitative genetics of geometric shape in the mouse mandible. *Evolution* 55 (11): 2342–2352.

Klingenberg C.P., Barluenga M. & Meyer A. 2002. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. *Evolution* 56 (10): 1909–1920. https://doi.org/10.1111/j.0014-3820.2002.tb00117.x

Kotov A.A. & Gololobova M.A. 2016. Traditional taxonomy: quo vadis? *Integrative Zoology* 11 (6): 500–505. https://doi.org/10.1111/1749-4877.12215

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 (11): 3006–3018. https://doi.org/10.1016/j.jas.2011.06.028

Kruckenhauser L., Pinsker W., Haring E. & Arnold W. 1999. Marmot phylogeny revisited: molecular evidence for a diphyletic origin of sociality. *Journal of Zoological Systematics and Evolutionary Research* 37 (1): 49–56. https://doi.org/10.1046/j.1439-0469.1999.95100.x

Kryštufek B., Janžekovič F., Hutterer R. & Klenovšek T. 2016. Morphological evolution of the skull in closely related bandicoot rats: a comparative study using geometric morphometrics. *Hystrix* 27 (2): 1–7. https://doi.org/10.4404/hystrix-27.2-11639

Krzywinski M. & Altman N. 2013. Power and sample size. *Nature Methods* 10 (12): 1139–1140. https://doi.org/10.1038/nmeth.2738

Kuzminsky S.C. & Gardiner M.S. 2012. Three-dimensional laser scanning: potential uses for museum conservation and scientific research. *Journal of Archaeological Science* 39 (8): 2744–2751. https://doi.org/10.1016/j.jas.2012.04.020

Legendre P. & Legendre L. 2012. *Numerical Ecology*. Elsevier, Oxford, UK.

Lenth R.V. 2001. Some practical guidelines for effective sample size determination. *The American Statistician* 55 (3): 187–193. https://doi.org/10.1198/000313001317098149

Marcus L.F. 1990. *Traditional morphometrics*. *In*: Rohlf F.J. & Bookstein F.L. (eds) *Proceedings of the Michigan Morphometrics Workshop - Special Publication Number 2*: 77–122. University of Michigan Museum of Zoology, Ann Arbor.

Marcus L.F., Hingst-Zaher E. & Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. *Hystrix, the Italian Journal of Mammalogy* 11 (1): 24–47. https://doi.org/10.4404/hystrix-11.1-4135

Marín A.G., Pérez C.H.F., Minoli I., Morando M. & Avila L.J. 2016. A new lizard species of the *Phymaturus patagonicus* group (Squamata: Liolaemini) from northern Patagonia, Neuquén, Argentina. *Zootaxa* 4121 (4): 412–430. https://doi.org/10.11646/zootaxa.4121.4.3

Marugán-Lobón J., Chiappe L.M. & Farke A.A. 2013. The variability of inner ear orientation in saurischian dinosaurs: testing the use of semicircular canals as a reference system for comparative anatomy. *PeerJ* 1: e124. https://doi.org/10.7717/peerj.124

Matějů J. & Kratochvíl L. 2013. Sexual size dimorphism in ground squirrels (Rodentia: Sciuridae: Marmotini) does not correlate with body size and sociality. *Frontiers in Zoology* 10 (1): 27. https://doi.org/10.1186/1742-9994-10-27

May R.M. 1990. Taxonomy as destiny. *Nature* 347 (6289): 129–130. https://doi.org/10.1038/347129a0

Milella M., Franklin D., Belcastro M.G. & Cardini A. 2021. Sexual differences in human cranial morphology: is one sex more variable or one region more dimorphic? *The Anatomical Record* 304: 2789-2810. https://doi.org/10.1002/ar.24626

Miller J.P., Delicado D., García-Guerrero F., Khalloufi N. & Ramos M.A. 2023. Morphology and taxonomic assessment of eight genetic clades of *Mercuria* Boeters, 1971 (Caenogastropoda, Hydrobiidae), with the description of five new species. *European Journal of Taxonomy* 866: 1–63. https://doi.org/10.5852/ejt.2023.866.2107

Millien V. 2006. Morphological evolution is accelerated among island mammals. *PLoS Biology* 4 (10): e321. https://doi.org/10.1371/journal.pbio.0040321

Mills K.K., Everson K.M., Hildebrandt K.B.P., Brandler O.V., Steppan S.J. & Olson L.E. 2023. Ultraconserved elements improve resolution of marmot phylogeny and offer insights into biogeographic history. *Molecular Phylogenetics and Evolution* 184: 107785. https://doi.org/10.1016/j.ympev.2023.107785

Minelli A. 2019. Biodiversity, disparity and evolvability. *In*: Casetta E., Marques da Silva J. & Vecchi D. (eds) *From Assessing to Conserving Biodiversity.* History, Philosophy and Theory of the Life Sciences, Vol. 24. Springer, Cham. https://doi.org/10.1007/978-3-030-10991-2_11

Moore D.S. & McCabe G.P. 2005. *Introduction to the Practice of Statistics*. WH Freeman & Co., New York, USA.

Moyers R.E. & Bookstein F.L. 1979. The inappropriateness of conventional cephalometrics. *American Journal of Orthodontics* 75 (6): 599–617.

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 (3): 258–267. https://doi.org/10.1111/j.1439-0469.2008.00503.x

Neff N.A. & Marcus L.F. 1980. *A Survey of Multivariate Methods for Systematics*. American Museum of Natural History, New York, USA.

O'Connell-Rodwell C.E., Freeman P.T., Kinzley C., Sandri M.N., Berezin J.L., Wiśniewska M., Jessup K. & Rodwell T.C. 2022. A novel technique for aging male African elephants (*Loxodonta africana*) using craniofacial photogrammetry and geometric morphometrics. *Mammalian Biology* 102 (3): 591–613. https://doi.org/10.1007/s42991-022-00238-2

O'Higgins P. 1997. Methodological issues in the description of forms. *In*: P. Lestrel (ed.) *Fourier Descriptors and Their Applications in Biology* (pp. 74-105). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511529870.005

O'Higgins P. 2000. The study of morphological variation in the hominid fossil record: biology, landmarks and geometry. *Journal of Anatomy* 197 (1): 103–120. https://doi.org/10.1046/j.1469-7580.2000.19710103.x

Oksanen J., Simpson G.L., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Solymos P., Stevens M.H.H., Szoecs E., Wagner H., Barbour M., Bedward M., Bolker B., Borcard D., Carvalho G., Chirico M., De Caceres M., Durand S., Antoniazi Evangelista H.B., Fitzjohn R., Friendly M., Furneaux B., Hannigan G., Hill M.O., Lahti L., McGlinn D., Ouellette M.-H., Ribeiro Cunha E., Smith T., Stier A., Ter Braak C.J.F. & Weedon J. 2022. *vegan: Community Ecology Package*. Available from https://cran.r-project.org/web/packages/vegan/index.html [accessed 9 Apr. 2024].

Okumura M. & Araujo A.G.M. 2019. Archaeology, biology, and borrowing: a critical examination of geometric morphometrics in archaeology. *Journal of Archaeological Science* 101: 149–158. https://doi.org/10.1016/j.jas.2017.09.015

Olsen A.M. & Westneat M.W. 2015. StereoMorph: an R package for the collection of 3D landmarks and curves using a stereo camera set-up. *Methods in Ecology and Evolution* 6 (3): 351–356. https://doi.org/10.1111/2041-210X.12326

Oxnard C. & O'Higgins P. 2009. Biology clearly needs morphometrics. does morphometrics need biology? *Biological Theory* 4 (1): 84–97. https://doi.org/10.1162/biot.2009.4.1.84

Perez K.E., Cruz M.A.M., Steury B.W. & Barker G.M. 2021. A fresh start in ambersnail (Gastropoda: Succineidae) taxonomy: finding a foothold using a widespread species of *Oxyloma*. *European Journal of Taxonomy* 757: 102–126. https://doi.org/10.5852/ejt.2021.757.1419

Polly P.D. 2005. Development and phenotypic correlations: the evolution of tooth shape in *Sorex araneus*. *Evolution & Development* 7 (1): 29–41. https://doi.org/10.1111/j.1525-142X.2005.05004.x

Polly P.D. & Motz G.J. 2016. Patterns and processes in morphospace: geometric morphometrics of three-dimensional objects. *The Paleontological Society Papers* 22: 71–99. https://doi.org/10.1017/scs.2017.9

Qubaiová J., Růžička J. & Šípková H. 2015. Taxonomic revision of genus *Ablattaria* Reitter (Coleoptera, Silphidae) using geometric morphometrics. *ZooKeys* 477: 79–142. https://doi.org/10.3897/zookeys.477.8446

Quinn G.P. & Keough M.J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press. https://doi.org/10.1017/CBO9780511806384

R Core Team 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rankin A.M., Schwartz R.S., Floyd C.H. & Galbreath K.E. 2019. Contrasting consequences of historical climate change for marmots at northern and temperate latitudes. *Journal of Mammalogy* 100 (2): 328–344. https://doi.org/10.1093/jmammal/gyz025

Reyment R.A. 2010. Morphometrics: an historical essay. *In*: Elewa A.M.T. (ed.) *Morphometrics for Nonmorphometricians*: 9–24. Springer Berlin Heidelberg, Berlin, Heidelberg.

Richtsmeier J.T., Deleon V.B. & Lele S. 2002. The promise of geometric morphometrics. *American Journal of Physical Anthropology* 119 (S35): 63–91. https://doi.org/10.1002/ajpa.10174

Roach N. 2017. *Marmota vancouverensis*. The IUCN Red List of Threatened Species 2017: eT12828A22259184. https://doi.org/10.2305/IUCN.UK.2017-2.RLTS.T12828A22259184.en

Rohlf F.J. 1970. Adaptive hierarchical clustering schemes. *Systematic Zoology* 19 (1): 58–82. https://doi.org/10.2307/2412027

Rohlf F.J. 1990. Morphometrics. *Annual Review of Ecology and Systematics* 21 (1): 299–316. https://doi.org/10.1146/annurev.es.21.110190.001503

Rohlf F.J. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology* 47 (1): 147–158.

Rohlf F.J. 2015. The tps series of software. *Hystrix, the Italian Journal of Mammalogy* 26 (1): 9–12. https://doi.org/10.4404/hystrix-26.1-11264

Rohlf F.J. 2021. Why clusters and other patterns can seem to be found in analyses of high-dimensional data. *Evolutionary Biology* 48 (1): 1–16. https://doi.org/10.1007/s11692-020-09518-6

Rohlf F.J. & Marcus L.F. 1993. A revolution morphometrics. *Trends in Ecology & Evolution* 8 (4): 129–132. https://doi.org/10.1016/0169-5347(93)90024-J

Rohlf F.J. & Slice D. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology* 39 (1): 40–59. https://doi.org/10.2307/2992207

Rohlf F.J., Loy A. & Corti M. 1996. Morphometric analysis of Old World Talpidae (Mammalia, Insectivora) using partial-warp scores. *Systematic Biology* 45 (3): 344–362. https://doi.org/10.1093/sysbio/45.3.344

Sargis E.J. 2002. A multivariate analysis of the postcranium of tree shrews (Scandentia, Tupaiidae) and its taxonomic implications. *Mammalia* 66 (4): 579–598. https://doi.org/10.1515/mamm.2002.66.4.579

Sargis E.J., Terranova C.J. & Gebo D.L. 2008. Evolutionary Morphology of the Guenon Postcranium and Its Taxonomic Implications. *In*: Sargis E.J. & Dagosto M. (eds) *Mammalian Evolutionary Morphology: A Tribute to Frederick S. Szalay*: 361–372. Springer Netherlands, Dordrecht.

Sargis E.J., Woodman N., Morningstar N.C., Bell T.N. & Olson L.E. 2017. Skeletal variation and taxonomic boundaries among mainland and island populations of the common treeshrew (Mammalia: Scandentia: Tupaiidae). *Biological Journal of the Linnean Society* 120 (2): 286–312. https://doi.org/10.1111/bij.12876

Sasakawa K. 2016. Two new species of the ground beetle subgenus *Sadonebria Ledoux* Roux, 2005 (Coleoptera, Carabidae, *Nebria*) from Japan and first description of larvae of the subgenus. *ZooKeys* 578: 97–113. https://doi.org/10.3897/zookeys.578.7424

Schlager S. 2017. Morpho and Rvcg – shape analysis in R. *In*: Zheng G., Li S. & Szekely G. (eds) *Statistical Shape and Deformation Analysis*: 217–256. Academic Press.

Schlis-Elias M.C. 2020. Ecological release and allometry explain insular gigantism and shape variation in a widespread North American rodent. MSc Thesis.
Available from https://aspire.apsu.edu/handle/20.500.11989/6700 [accessed 5 Apr. 2024].

Schulte-Hostedde A.I. 2008. Sexual Size Dimorphism in Rodents. *In*: Wolff J.O. & Sherman P.W. (eds) 2007. *Rodent Societies: An Ecological and Evolutionary Perspective*: 115–128. University of Chicago Press, Chicago, IL.

Schwarzfeld M.D. & Sperling F.A.H. 2014. Species delimitation using morphology, morphometrics, and molecules: definition of the *Ophion scutellaris* Thomson species group, with descriptions of six new species (Hymenoptera, Ichneumonidae). *ZooKeys* (462): 59–114. https://doi.org/10.3897/zookeys.462.8229

Singh G.D., Levy-Bercowski D. & Santiago P.E. 2005. Three-dimensional nasal changes following nasoalveolar molding in patients with unilateral cleft lip and palate: geometric morphometrics. *The Cleft Palate Craniofacial Journal* 42 (4): 403–409. https://doi.org/10.1597/04-063.1

Singh G.D., Levy-Bercowski D., Yáñez M. & Santiago P. 2007. Three-dimensional facial morphology following surgical repair of unilateral cleft lip and palate in patients after nasoalveolar molding. *Orthodontics & Craniofacial Research* 10 (3): 161–166. https://doi.org/10.1111/j.1601-6343.2007.00390.x

Slice D.E. 1999. *Morpheus et al. Ecology and Evolution. State University of New York, Stony Brook*. Available from https://sbmorphometrics.org/morphmet/morpheus_vienna_2006.zip [accessed 5 Apr. 2024].

Slice D.E. 2001. Landmark Coordinates Aligned by Procrustes Analysis Do Not Lie in Kendall's Shape Space. *Systematic Biology* 50 (1): 141–149. https://doi.org/10.1080/10635150119110

Smith G.R. 1990. Homology in morphometrics and phylogenetics. *In*: Rohlf F.J. & Bookstein F.L. (eds) *Proceedings of the Michigan Morphometrics Workshop.* Museum of Zoology, University of Michigan. Special Publication 2.

Smiti A. 2020. A critical overview of outlier detection methods. *Computer Science Review* 38: 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Sneath P.H.A. 1967. Trend-surface analysis of transformation grids. *Journal of Zoology* 151 (1): 65–122. https://doi.org/10.1111/j.1469-7998.1967.tb02866.x

Sokal R.R. & Rohlf F.J. 1962. The comparison of dendrograms by objective methods. *Taxon* 11 (2): 33–40. https://doi.org/10.2307/1217208

Sokal R.R. & Rohlf F.J. 2009. Introduction to biostatistics second edition. *Dover Publications Inc, NY* 1081: 1–590.

Sokal R.R. & Rohlf F.J. 2011. *Biometry*. W.H. Freeman and Company, New York, US.

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 (4): 715–734. https://doi.org/10.1080/106351599259988

Steppan S.J., Kenagy G.J., Zawadzki C., Robles R., Lyapunova E.A. & Hoffmann R.S. 2011. Molecular data resolve placement of the Olympic marmot and estimate dates of trans-Beringian interchange. *Journal of Mammalogy* 92 (5): 1028–1037. https://doi.org/10.1644/10-MAMM-A-272.1

Su J., Guan K., Wang J. & Yang Y. 2015. Significance of hind wing morphology in distinguishing genera and species of cantharid beetles with a geometric morphometric analysis. *ZooKeys* 502: 11–25. https://doi.org/10.3897/zookeys.502.9191

Sugasawa S., Klump B.C., St Clair J.J.H. & Rutz C. 2017. Causes and consequences of tool shape variation in New Caledonian crows. *Current Biology* 27 (24): 3885-3890.e4. https://doi.org/10.1016/j.cub.2017.11.028

Taylor R.W. 1983. *Descriptive Taxonomy: Past, Present, and Future*. Canberra, CSIRO.

Tobias J.A., Seddon N., Spottiswoode C.N., Pilgrim J.D., Fishpool L.D.C. & Collar N.J. 2010. Quantitative criteria for species delimitation. *Ibis* 152 (4): 724–746. https://doi.org/10.1111/j.1474-919X.2010.01051.x

Uttley J. 2019. Power analysis, sample size, and assessment of statistical assumptions—improving the evidential value of lighting research. *LEUKOS* 15 (2–3): 143–162. https://doi.org/10.1080/15502724.2018.1533851

Valdez-Mondragón A., Navarro-Rodríguez C.I., Solís-Catalán K.P., Cortez-Roldán M.R. & Juárez-Sánchez A.R. 2019. Under an integrative taxonomic approach: the description of a new species of the genus *Loxosceles* (Araneae, Sicariidae) from Mexico City. *ZooKeys* 892: 93–133. https://doi.org/10.3897/zookeys.892.39558

Viscosi V. & Cardini A. 2011. Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. *PLoS One* 6 (10): e25630. https://doi.org/10.1371/journal.pone.0025630

von Cramon-Taubadel N., Frazier B.C. & Lahr M.M. 2007. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. *American Journal of Physical Anthropology* 134 (1): 24–35. https://doi.org/10.1002/ajpa.20616

Wainer H. 2007. The most dangerous equation. *American Scientist* 95 (3): 249. https://doi.org/10.1511/2007.65.249

Wheeler Q. 2014. Are reports of the death of taxonomy an exaggeration? *New Phytologist* 201 (2): 370–371. https://doi.org/10.1111/nph.12612

Whelan N.V., Strong E.E., Gladstone N.S. & Mays J.W. 2023. Using genomics, morphometrics, and environmental niche modeling to test the validity of a narrow-range endemic snail, *Patera nantahala* (Gastropoda, Polygyridae). *ZooKeys* 1158: 91–120. https://doi.org/10.3897/zookeys.1158.94152

Wilson E.O. 2002. *The Future of Life*. Knopf Doubleday Publishing Group, NY (US).

Yazdi F.T., Adriaens D. & Darvish J. 2014. Cranial phenotypic variation in *Meriones crassus* and *M. libycus* (Rodentia, Gerbillinae), and a morphological divergence in *M. crassus* from the Iranian Plateau and Mesopotamia (Western Zagros Mountains). *European Journal of Taxonomy* 88: 1–28. https://doi.org/10.5852/ejt.2014.88

Yezerinac S.M., Lougheed S.C. & Handford P. 1992. Measurement error and morphometric studies: statistical power and observer experience. *Systematic Biology* 41 (4): 471–482. https://doi.org/10.2307/2992588

Zachos F.E. 2018. Mammals and meaningful taxonomic units: the debate about species concepts and conservation. *Mammal Review* 48 (3): 153–159. https://doi.org/10.1111/mam.12121

Zelditch M., Swiderski D., Sheets D. & Fink W. 2004. *Geometric Morphometrics for Biologists: A Primer*. Elsevier Academic Press. Waltham, MA (US).

Zimek A. & Filzmoser P. 2018. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs Data Mining and Knowledge Discovery* 8 (6): e1280. https://doi.org/10.1002/widm.1280

Manuscript received: 30 July 2023
Manuscript accepted: 9 February 2024
Published on: 15 May 2024
Topic editor: Tony Robillard
Section editor: Frank Zachos
Desk editor: Kristiaan Hoedemakers

Supplementary files

File descriptions for both parts A and B

Formats: mj.txt = MorphoJ (with 'id', the specimen label*, in the first column); past.txt = PAST (with its own label in the first column, which for multivariate analyses also contains the group colour code); nts = TPS Series

*(bro_mle_MVZ_8360 is a specimen that should be renamed as cal_mle_MVZ_8360, because it is a hoary marmot, as correctly reported in the species classifiers. In the label, which was not used for any analysis, I kept the wrong abbreviation (bro) used in the original jpg image name. However, in general, it is better to have accurate, descriptive labels, as discussed in V&C - see main text).

Supp. file 1. ALL_RAW_15L_N462by2.mj.txt: this is the main data file with the raw coordinates of all 15 landmarks and 462 individuals (including possible outliers), each with its two digitizations. It is the main morphometric dataset, from which all others can be obtained. It should be used in MorphoJ for assessing ME, but also, once low precision landmarks and outliers are removed, it can be used for all main analyses with averaged individuals (*Preliminaries, Average observations by ...* using the classifier 'indiv'). https://doi.org/10.5852/ejt.2024.934.2527.11347

Supp. file 2. ALL_CLASSIFIERS.mj.txt: the file contains the following variables: subgenus, species, modern_paleo (which is relevant only for *M. vancounverensis* in order to distinguish recent and subfossil specimens), sex, indiv (an integer used as a simple individual identifier useful to recognize duplicates), side (of the mandible), OUTLIER (marks the 17 potential outliers, which were excluded from the main analyses), collection (where specimens originated), catalogue_number (in the corresponding museum), year_coll (year when the specimen was captured), Country, Province_State, Locality (with these last three variables containing the information, if available, on geographical origin of a specimen). https://doi.org/10.5852/ejt.2024.934.2527.11349

Supp. file 3. ALL_COVARIATES.mj.txt: this file can be created from the previous one, as it simply recodes a few variables using integer numbers (species, sex - with 0 for females and 1 for males - and year_coll, which was already numeric). For instance, it can be useful to test sex using regressions on dummy variables after removing outliers and unsexed individuals (in MorphoJ: *Preliminaries, include or exclude observations*) and splitting data by species (in MorphoJ: *Preliminaries, Subdivide dataset by ...*). HOWEVER, the species covariate cannot be used for similar purposes without being modified. For ANOVAs/MANOVAs using the regression approach, one needs a design matrix. For pairwise tests of species differences using regressions, one has first to subset the data (e.g., select VAN and woodchucks by first splitting by species and then combining these two species in MorphoJ) and probably replace the species code with that conventionally employed for dummy variables (say, VAN = 1 and woodchuck = -1). https://doi.org/10.5852/ejt.2024.934.2527.11351

Supp. file 4. TESTING_SDM_CS_12L_N356.past.txt: this is to run the species by sex ANOVA of CS in PAST (B1). It only includes the 356 specimens of known sex. The main variables are sp.n (coding species with an integer as required in PAST for the two-way ANOVA), sex.n (coding males as 1 and females as 2) and CS (centroid size). Other variables are described above and can be ignored. Some (species, sex and indiv) are included only as an aid to identify specimens; sp.nBigN codes species with largest samples with an integer if one wants to repeat the ANOVA after excluding small samples (drag this variable so that it replaces sp.n and then select only the rows with woodchucks, hoary and yellow-bellied marmots). https://doi.org/10.5852/ejt.2024.934.2527.11353

Supp. file 5. TESTING_species_CS_12L_N445.past.txt: this is an example of how to organize data, after pooling sexes, for most univariate analyses/plots in PAST. For instance, it can be used for a one-way ANOVA testing species differences in CS (B3) or for drawing a box-plot of CS (B3). https://doi.org/10.5852/ejt.2024.934.2527.11355

Supp. file 6. TESTING_SDM_SH_12L_N356_sexed.nts: this is the landmark data to run the MANOVA in TPSRegr as explained in the main text of the second paper (B2). They only include the 356 individuals of known sex. https://doi.org/10.5852/ejt.2024.934.2527.11357

Supp. file 7. TESTING_SDM_SH_12L_N356_dummy_variables_MANOVA.nts: this is the landmark design matrix to run the MANOVA in TPSRegr as explained in the main text of the second paper (B2). They only include the 356 individuals of known sex. https://doi.org/10.5852/ejt.2024.934.2527.11359

Supp. file 8. TESTING_SLOPES_etc_STATIC_ALLOMETRY_12L_N445.nts: this is the landmark data to run the MANCOVA in TPSRegr as explained in the main text of the second paper (B6). https://doi.org/10.5852/ejt.2024.934.2527.11361

Supp. file 9. TESTING_SLOPES_etc_STATIC_ALLOMETRY_12L_N445_dummy_variables_MANCOVA.nts: this is the design matrix to run the MANCOVA in TPSRegr as explained in the main text of the second paper (B6). https://doi.org/10.5852/ejt.2024.934.2527.11363

Appendix A

The Appendix of part A opens with an example of how group mean differences tend to be overestimated in small samples, a common problem in most taxonomic studies and one that can lead to misinterpreting results. After this example, I informally discuss the frequentist approach I usually follow in statistical investigation, based on null hypothesis testing and P values. This popular approach is increasingly criticized. I will devote some space to clarify that, for most statisticians, the problem lies with its misuse and misinterpretation and not with the approach itself. In general, statistical models make assumptions which are often neglected. Thus, I also briefly discuss some of the most common assumptions of the methods used in the main study (A and B). Finally, I have included a short section on semilandmarks: they, and the methods used to analyse these 'special points', also make assumptions, which should be carefully considered and, yet, are mostly ignored, overlooked or misreported. At the end of the Appendix, readers can find a short, informal glossary of selected technical terms frequently used in parts A and B: I hope it may be of help for less experienced morphometricians.

Propensity of group mean differences to be overestimated in small samples: an example using yellow-bellied marmots

That group mean differences tend to be overestimated in small samples has already been mentioned. In the chapter on power analysis (A3), I provided a simple explanation for the reason why this happens. A more detailed, concrete example might, however, help to show the effect of sample size on distances between group means.

The example is graphically summarized in the Figure A1 of the Appendix. It is based on a simple randomization experiment in yellow-bellied marmots. I selected this species because it is the one with the largest sample, but results would be analogous in other species. The 'experiment' consists in computing the shape distance between means of balanced random subsamples of females and males, whose size is progressively reduced. I started with 70 individuals per-sex, which is almost the full sample for this species. With N = 70, the mean female to male distance is 0.0113 units of Procrustes distance. Then, I halved the sample size, so that there were four mutually exclusive subsamples with N = 35, two made of females and two of males. I computed the pairwise distances of the female to male means and took their median value, which is 0.0130 (thus, 15% larger than the previous estimate using 70 individuals). I approximately halved again N, using the same rationale, and repeated the computations, which I redid also for random mutually exclusive subsamples of just 10 or five individuals per sex. The median of the estimated mean female to male distance keeps growing until, with N = 5, becomes 0.0297, which is almost three times larger than with the initial N = 70.

The trend is clear, but it is worth observing how, with N = 10 or less, even the smallest female-to-male mean shape distance is larger than all estimates based on 35–70 individuals (Fig. A1). Besides, if observed shape distances between mean females and males within the species with the smallest study samples (i.e., VAN, Olympic and Alaskan marmots) are added to the box-plots of Figure A1 using red filled circles, one might note some interesting patterns. First of all, as sample size becomes smaller (average within sex N is 10 in VAN, 7 in Olympic marmots, and 4 in Alaskan marmots), observed mean differences in these three species increase sharply. This is similar to what is seen in the subsamples of yellow-bellied marmots and suggests that their SDM tends to be overestimated because of sampling error. Yet, for the Olympic marmot and VAN, the bias seems minimal, as they are both below the minimum distance found in randomized subsamples of yellow-bellied marmots of approximately the same size (i.e., 5–10 individuals, respectively). In contrast, the between sex mean difference of the Alaskan marmots is larger than the median found in the smallest subsamples (N = 5) of yellow-bellied marmots and close to the 'top' largest 25%. As suspected, thus, the SDM in mandibular shape of the

Alaskan marmot is likely to be hugely overestimated. The effect of this bias has already been noticed in the discussion of the results of the power analysis (A3).

The randomized subsampling experiment makes the positive bias in estimates of mean shape differences less abstract and, hopefully, clearer. It is also an example of the type of clues a researcher might obtain using simple randomized subsampling experiments. Using random subsamples of large samples to explore the impact of sampling error is less rigorous and generalizable than employing well designed, extensive simulations. Nonetheless, it is much simpler, and therefore doable even by those who have weaker theoretical bases and no knowledge of programming languages. In part B, I will provide more examples of this type of exploratory analysis.

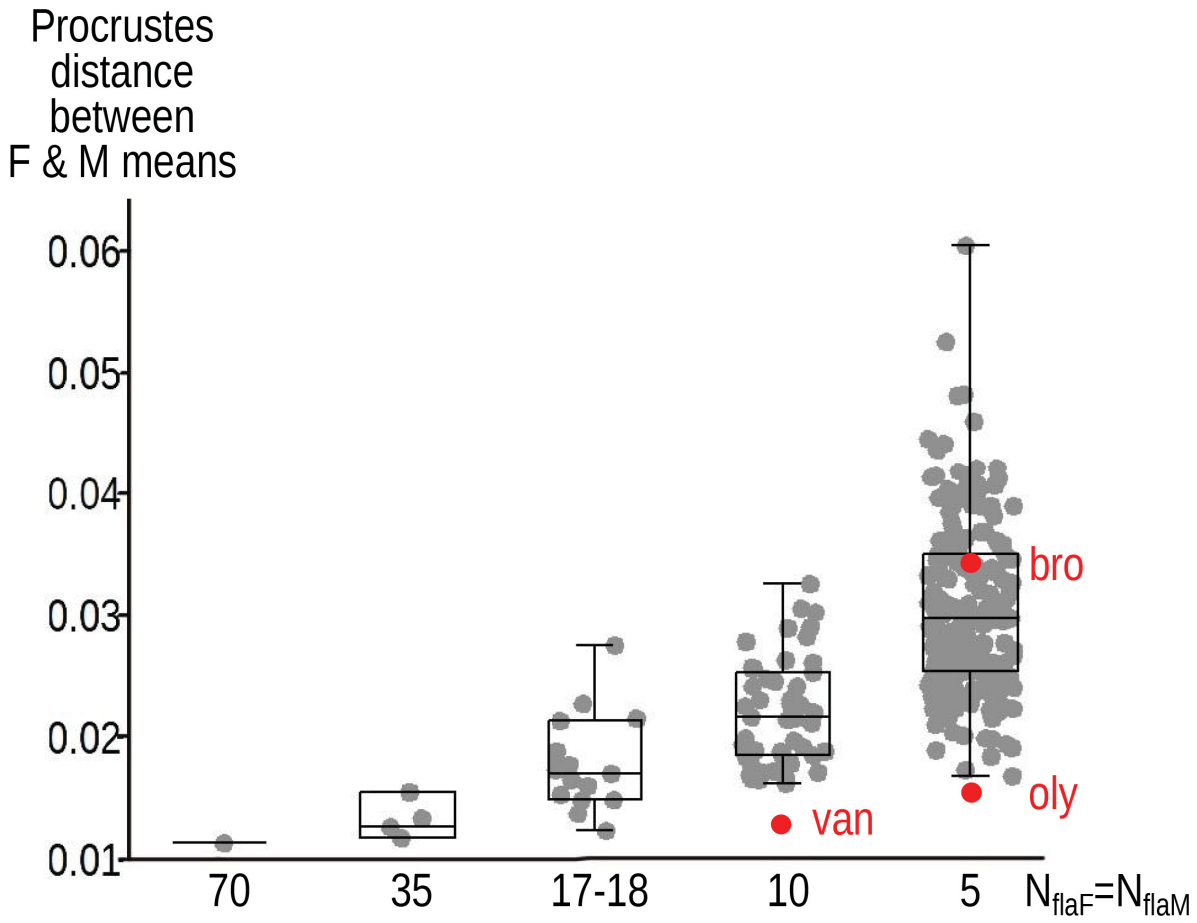


Fig. A1. Box and jitter plot of shape SDM estimated using randomized subsampling experiments in yellow-bellied marmots (whiskers in this figure mark the range from minimum to maximum, with no assessment of outliers, which are irrelevant in the context of this didactic example). The total female sample (F) is split in progressively smaller, mutually exclusive, random subsamples and the same is done for males (M). The first subsample of 70 individuals per sex is almost the same of the total samples; the second consists of two female and two male subsamples with 35 individuals each; the third set of subsamples is made of 17–18 individuals per sex and the fourth and fifth include only 10 or 5 individuals respectively. The vertical axis shows the Procrustes distance between means of females and means of males of each set of subsamples. The red circles show the observed mean female to male Procrustes distance in VIM, Alaskan (bro) and Olympic marmots (oly); they are added to the box and jitter plots of yellow-bellied marmot subsamples whose N is closer to the observed N is these three species.

Frequentist statistics: interpretation, pitfalls and how to avoid them

Testing group differences lies at the heart of virtually all morphometric studies in taxonomy. Statistical testing is part of the frequentist approach, whose origin dates back to the work of Fisher and colleagues in the first half of the 20th century (Goodman 2016; Haaf *et al.* 2019; Betensky 2019). Central to frequentist statistics is the use of P values to decide if a null hypothesis can be rejected. Informally, "a P-value ... represents the probability of obtaining the result (or something more extreme [than observed]) assuming that there was no real effect or difference between the groups or measures being tested (the "null" hypothesis)" (Uttley 2019: 144).

There is a number of common misunderstandings and misuses of P values that are relevant to taxonomists. For example, the null hypothesis is usually framed as the absence of an effect, as in Uttley's definition, but can, in fact, be any size of an effect. Regardless of how the null hypothesis is framed, P is not the probability of the null hypothesis. Besides, statistical significance is based on an arbitrary threshold (alpha, as mentioned in part A), which is not the same as biological significance and, therefore, cannot be the unique reason for claiming a given taxonomic status (subspecies or species, mainly). In the next paragraphs, I try to clarify in plain terms these frequent misunderstandings and briefly discuss some of the potential pitfalls and abuses of P values. If a researcher is aware of the problems, they are easy to avoid and statistical testing and P values can be fruitfully employed.

An accurate summary of the frequentist 'philosophy' and a rigorous clarification of the meaning of P values can be found in Greenland *et al*. (2016). One aim of statistics, they say (Greenland *et al*. 2016: 339) is the "evaluation of certainty or uncertainty regarding the size of an effect"; however, in frequentist statistics, instead of expressing this uncertainty in terms of "probabilities of hypotheses …, 'probability' refers ... to quantities that are hypothetical frequencies of data patterns under an assumed statistical model". The second part of this statement, about data patterns and models, may sound cryptic to biologists without background in statistics, but it is less obscure than it looks.

The statistical model is a set of assumptions, among which, typically, the main one is that there is no effect at all (e.g., no group differences or no correlation between variables). However, as mentioned, and somewhat counter-intuitively given its name, a null hypothesis can be made also for specific non-zero effect sizes. For instance, one could test that men are on average at least, say, 5 cm taller than women, instead of testing that men and women have the same height. Once the null hypothesis is clearly formulated, a P value summarizes the compatibility between observed data, obtained using a sample from a larger population (all human beings, in the made-up example I used above, or all individuals of a species or subspecies in a taxonomic study) and the expectations we have if the null hypothesis is correct. More accurately, the expectations are those obtained if (Greenland *et al*. 2016: 339) "we knew the entire statistical model (all the assumptions used to compute the P-value [including the null hypothesis!]) were correct". Thus, $P = 1$ implies complete compatibility between the data and the model, and $P = 0$ complete incompatibility. Intermediate Ps ($1 > P > 0$) indicate where the data lie in the continuous range of their potential compatibility with the null hypothesis.

These definitions have profound implications and are, as anticipated, easily misunderstood. They are still difficult for me after almost 25 years since I started using statistics as a self-taught morphometrician with no intuition for the mathematical language. My personal recipe to get the concept right and avoid misinterpretations of P values relies on the reasoning behind a simple permutation test. The test could be that of the correlation between two variables or a test for the mean differences between two groups. I will use the correlation, as an example, but readers can find the permutation test of group means clearly explained in the chapter on Computer-Based Statistical Methods of Zelditch *et al.* (2004). The rationale behind Ps is the same in other tests, including parametric ones.

I emphasize using *italics* some of the main points in this description. In a permutation test, one uses the available data to simulate what results they (i.e., the specific data that were sampled) would produce if the null hypothesis was true. This is a first reminder that *P values are computed by considering the null hypothesis as if it is the correct one*. Now, let us say, as an example, that I am interested in the correlation between cranial CS and body mass, in a representative random sample of 20 adults of hoary marmots. That the sample should be random is important, because having *random independent observations* is another assumption of the model used for the test (for more on this, see next section on assumptions). Also, the test will be one-tailed, because I expect the correlation, if present, to be positive: the larger the body mass, the bigger the cranium. Thus, using the data in the sample of 20 individuals, I compute the correlation and find, for instance, r = 0.4. This r is the *'observed correlation'*. The question then is: is 0.4 *large enough, given the variability in the available data, to reasonably exclude that it was obtained just because of sampling error in that specific sample?* If, in fact, it was just a 'by-product' of sampling error, by measuring another sample, I should obtain a different value, sometimes smaller and sometimes larger. By extracting random samples many times, when the null hypothesis of no effect is correct, the average of all estimated correlations should be zero. To simulate this type of 'experiment', I need to *create data compatible with the null hypothesis of no correlation*. These data will be used to estimate the range of r I expect in a sample of 20 hoary marmots, when there is no association between CS and body mass, except by chance, purely because of sampling error.

Then, how do I *simulate 'random chance' correlations*? Easily: by randomizing the order of the 20 individuals in one of the two variables before recomputing r. Randomizing means that, if data were organized in two columns, with individuals in rows, as in an Excel spreadsheet, I randomly reorder the individuals in the CS column (or those in the body mass column; it is irrelevant which variable is randomized). The *reordering breaks any potential true association* between CS and body mass because individual one will now 'have' the CS of, for instance, individual three; individual two might have the CS of individual 18, and so on and so forth. If I repeat the randomization 1000 times, occasionally I might get large (positive or negative) values, but most of the time they will be close to zero and their average will be centered on zero, because I generated data with random correspondences between CS and body mass and, thus, with no 'true' correlation.

To visualize the distribution of r from many runs of randomizations one can use a frequency histogram. This type of plot will show the *empirical approximation of the probability distribution of r for data compatible with the null hypothesis of no correlation* between CS and body mass. I stress again that *all the r values obtained using permutations are compatible with the null hypothesis* of no correlation, because they are all obtained by creating random chance associations. Finally, I plot the original observed r = 0.4 (i.e., the r of the sample before any randomization) on the histogram. *The observed r is significant when* it is so extreme that it *'looks like an outlier'*, which, in practice, means that it is found among (or beyond) the most extreme values in the histogram. In my specific example, where I expect r to be positive, the observed r = 0.4 should be in the positive tail of the empirical distribution of 'random chance' correlations to be significant. If it is there, far from r values expected when there is no real correlation, then the observed data have produced a correlation that is highly incompatible with the null hypothesis. Precisely, the P value in the test is computed as the *number of times randomized data produced r ≥ 0.4*, divided by the number of randomizations in the simulation[20]. Phrased this way,

[20] More accurately, one (the observed r) plus the number of times randomized data produced r ≥ 0.4. The observed r must be incorporated in both the denominator and the numerator to avoid having zero divided by the number of randomizations, which could happen if none of the randomizations produces r ≥ 0.4. For this reason, generally one uses 999 randomizations (instead of 1000) and adds one (the observed) both to the numerator and denominator. Also, the test, computed the way I wrote, is one-tailed (i.e., the null hypothesis is rejected only by positive correlations ≥ 0.4) because I expect r to be positive and, thus, I consider the sign of the correlations in the randomized data. If I wanted a two-tailed test (the null hypothesis is rejected whenever r is ≥ 0.4 or ≤ -0.4), then I just need to use the absolute value of r when counting how many times randomized data produced r ≥ 0.4.

I am reminded that *P is the probability of the data producing r as large as or larger than the observed r.* Thus, it is not the probability of the null hypothesis! If, out of, for instance, 1000 permutations, which generated 'random chance' correlations, I find P = 0.03, that means that only 3% of the time random associations can be as high or higher than observed in the available sample, which suggests that the observed data are unlikely to be compatible with (a null hypothesis of) no correlation.

Up to this point, I have explained how we infer how much the data are incompatible with the null hypothesis. However, I have not yet said what is the precise P value below which one can confidently consider that they are so incompatible to reject the null hypothesis. In the example using the correlation of CS and body mass, this means deciding how rare large 'random chance' correlations have to be in order to confidently reject the null hypothesis. This is decided by arbitrarily choosing a significance threshold, which conventionally is indicated with alpha. If I am happy to take a relatively large risk of claiming a positive association between CS and body mass (liberal threshold), when in fact there is no real association, I could use alpha = 0.1. Then, I accept that, even if up to 10% of the randomizations produced $r \geq 0.4$, I do not believe that 0.4 is likely to be just a consequence of sampling error. However, if I am more cautious, and want the risk of potentially spurious claims based on my data to be really small (conservative threshold), I may choose alpha = 0.005 [21]. In this case, among the 1000 r from the randomizations, I tolerate less than five cases of $r \geq 0.4$, by chance, in order to be able to state that the observed correlation is significant. With this alpha, only if $P < 0.005$, I will consider data in the sample to be so highly incompatible with the null hypothesis of no correlation to have confidence that the association between CS and body mass is likely to be genuine.

Frequentist statistics is often criticized and criticisms seem to have increased in the last two decades (as an example, see references in Muff *et al.* 2022). Yet, it is the dominant approach in user-friendly GMM programs, as well as in their R counterparts. In fact, frequentist statistics is valid, if correctly used and interpreted. For readers who want to learn more, there is an abundant literature on the matter, including some relatively short introductory papers aimed at a broad readership (Weinberg 2001; Wasserstein & Lazar 2016, as well as, in the SI of Wasserstein & Lazar 2016: Senn 2016 and Benjamini 2016; Goodman 2016; Ioannidis 2019; Amrhein *et al.* 2019; Greenland 2019). As these readings show, frequentist statistics has pitfalls, but alternatives are not devoid of their own problems and many issues are, in fact, general. A cautious application of hypothesis testing and a reasoned interpretation of its results will not be misleading. Likely, thus, the morphometric evidence from the application of frequentist methods will lead to the same taxonomic conclusions as other approaches.

Briefly, I list and comment on some of the main caveats in the use of frequentist statistics. I use, again, italics to underscore important points. To start, as already stressed, *statistical and biological significance are not the same*. Biological significance is typically based on many studies and multiple lines of evidence (Adams 2019). In taxonomy, for instance, significant morphometric differences are rarely enough to make strong claims on whether a population is a subspecies or species (see Introduction in A and Conclusions in B). However, they can provide clues for further studies or help to corroborate or refute previous ones.

A null hypothesis of no effect is often the default choice, but sometimes *may not be the most appropriate* way of phrasing a research question. With taxonomic variation, most of the time testing the absence of differences between taxa seems meaningful even if one is not expecting zero differences. However, if there is a good justification for it, a researcher could use a null hypothesis that corresponds to a certain amount of expected differences and, then, set up the test to see if the data in the sample produce results incompatible with that specific threshold. I am not suggesting to do it, but, as thought example, one

[21] A lower alpha might also offer some protection for the possibility of inflating type I error rates in samples with violations of the assumption of non-independence of the observations (Stevens & James 2015).

could test whether a specific population has mean differences larger than found on average in other populations of the same species. If true, it could be argued that this population is unusually distinctive. In terms of software, this type of test likely requires a programming environment such as R as all the user-friendly programs I know use a null hypothesis of no differences in their tests.

Other common criticisms to frequentist statistics concern the misuse of the outcome of the tests of hypothesis. Selectively reporting results, for instance by cherry picking significant findings, is called '*P hacking*'[22] (Wasserstein & Lazar 2016). Making or modifying hypotheses after seeing the results of tests (i.e., creating 'post-hoc' hypotheses) is, instead, called *HARKING* (Hypothesizing After Results are Known). Both are bad practices that should be avoided, but none is strictly specific to frequentist statistics. In fact, HARKING may not be always wrong, if the decision is clearly motivated and transparently acknowledged (Hollenbeck & Wright 2017).

In general, *negative results should not be overlooked* (Fanelli 2012) and, when negative findings are the consequence of a non-significant P value, they must be interpreted correctly. Significance thresholds are arbitrary cut-off points and non-significance simply implies a lack of evidence using the available data. *A lack of evidence, however, does not simply translates into evidence for the absence of a given effect.* Non-significant differences between two taxa indicate either negligible differences or samples which are too small for confidently rejecting the null hypothesis using a specific alpha. Thus, *the arbitrary alpha threshold for significance should not lead to "dichotomania"* (Amrhein *et al.* 2019: 265), which is the belief that P values can be rigidly interpreted as a yes or no answer. *P values represent a continuum*, which is why it is better to report the actual value rather than saying if a test produced a P below or above a certain alpha. Precisely reporting Ps is, in fact, requested by many journals and the philosophy I followed in this study. The exceptions, where one can report P as less than a given threshold, is when they are very low (Greenland 2019) or the P values are those representing the minimum obtainable in a permutation test (i.e., one divided by the number of permutations – e.g., with 10000 permutations, $P < 0.0001$).

Finally, a researcher should always be aware that non-significance, but also significance, may sometimes happen for the 'wrong' reasons. The null hypothesis is the main assumption of a model in a test, but there are *other assumptions*, which *could have been violated*. For instance, with taxonomic data, observations are typically not independent and often may not even be truly representative of the variation in a population (Cardini 2020a). Many tests of group differences require homogeneous variance (see below on assumptions) and parametric methods usually need normally distributed data. Assumptions are often overlooked, but, at least for techniques which are fairly standard, are explained in clear terms in most introductory manuals. For instance, Hair *et al.* (2013) provide a graphical summary for each of the multivariate methods they cover, with a dedicated box for the model assumptions. This box is a concise reminder, but detailed explanations are available in the main text of each chapter in the book.

Even when properly used and interpreted, P values provide just one side of the story in a statistical analysis (Maher *et al.* 2013; Zhang 2019). As I have often mentioned, effect size, which in taxonomy measures how large differences are, is at least as important as the P value for those differences. *Estimates of effect size*, therefore, *must be always provided* together with the other results of a test. Ideally, these estimates should be accompanied by confidence intervals, which have been briefly discussed in the section on power analyses. Why, however, are confidence intervals so useful in relation to estimates of effect size? *Confidence intervals are also called compatibility intervals, because they cover the range of effect sizes*

[22] An informal way of describing P hacking is in the citation of Jeffreys (1961) cited in Ioannidis (SI in Wasserstein & Lazar 2016: 1): "A null hypothesis ... set up and 'tested' against data … is merely something set up like a coconut to stand until it is hit" (Jeffreys 1961) [with] many scientific fields ... accustomed to taking an endless number of shots until they (unfortunately) hit the coconut".

compatible with the data (not to be confused with the range that includes the true effect size) (Greenland 2019). The usefulness of confidence intervals for effect sizes, as well as a reasoned interpretation of the results of null hypothesis testing, are well exemplified using ecological data by Stephens *et al.* (2007: 193), which I quote at length: "Consider the simplest effect size statistic, the counter-null. This is the non-null magnitude of effect size that is supported by the same amount of evidence as the null. If we estimate the growth rate of a population, for example, our estimate might be -15% y^{-1}, with a 95% confidence interval from -32% to +2%. Some would interpret this to mean that we cannot reject the null hypothesis (of a zero growth rate). By contrast, if our estimate of decline is subject to normally distributed error, the counter-null indicates that a rate of decline of 30% per annum [which has the same distance from 15% as 0% but lies on the opposite side of the interval] is just as well supported as an estimate of zero. *The counter-null thus reminds researchers that a failure to reject the null does not mean that the null effect is more plausible than alternatives*". As already discussed, confidence intervals are, for now, rarely estimated, especially for multivariate shape, by user-friendly GMM programs. Commercial software and R (either by coding or using specific packages) might be necessary to add confidence intervals. Replicating analyses many times after bootstrapping samples offers a relatively simple approach to estimate confidence intervals (Manly 2007). Randomized subsampling experiments, such as those I exemplified in Figure A1 and in the analyses of mean shapes similarity relationships in part B (see also Cardini *et al.* (2021), and references therein), can also help to preliminary explore the sensitivity of results to sampling error and the robustness of the findings. However, randomized subsampling experiments do not have the same aim as, and cannot replace, confidence intervals.

P values and estimates of effect size should be complemented with graphical summaries. This is a point I stress often in both part A and B and is not specific to preliminary analyses, such as the assessment of ME or the detection of outliers. By plotting the data, one can gain insight into patterns of variability, spot potential problems and provide more accurate interpretations of the outcome of statistical tests. Les Marcus, one of the 'fathers' of multivariate morphometrics and GMM (Neff & Marcus 1980; Marcus 1990; Rohlf & Marcus 1993) used to tell colleagues to "plot the hell out of your variables" (from the "1990 – Stony Brook Morphometric Workshop at SUNY: memorable statements" opening the Yellow Book of GMM – Cardini & Loy 2013). For shape data, plots might be accompanied by shape diagrams (Klingenberg 2013), which are generally part of the analytical output in user-friendly GMM software, such as MorphoJ or the TPS Series. For univariate and multivariate analyses, PAST also offers a range of plots either in the *Plot* menu or as options of specific analyses. Programs usually allow some basic editing of the graphical output, but, for better quality, the plots can be saved as EMF or SVG and edited with a vector graphics software (for instance, the freeware open source Inkscape: https://inkscape.org/).

In conclusion, the temptation of relying exclusively on P values, and the apparently simple numerical outputs of tests of hypothesis, might be strong, but often represents a misuse of otherwise valid techniques (Zhang 2019). Other approaches may help to overcome some of the problems with null hypothesis statistical testing. With some caveats (and further assumptions - Benjamin & Berger 2019; Senn 2016 in the SI of Wasserstein & Lazar 2016), P values can be converted into more intuitive quantities. For instance, P values can be associated to upper bound Bayes factors. Bayes factors are "the largest odds in favour of the alternative hypothesis relative to the null hypothesis that is consistent with the observed data", which is, thus, their relative likelihood ratio (Benjamin & Berger 2019: 187). In turn (with one more assumption, explained in Benjamin & Berger 2019), Bayes factors can be translated into the probability of the alternative hypothesis being true given the available data. It is also possible to convert P values into S values ($S = -\log_2(P)$), which are the amount (bits, more precisely) of information against the null hypothesis (Greenland *et al.* 2016; Greenland 2019). This is, to my knowledge, probably the most intuitive way of expressing what a P value is saying, given the model of a specific test. For instance, $P = 0.05$ translates into $S = 4$, which is the probability of a fair coin to always give head, when tossed four consecutive times. Four consecutive heads, when tossing a coin, is unlikely, but not impossible.

Thus, using an alpha of 0.05 sets a probability threshold equivalent to using at least four consecutive heads as evidence to decide that a coin (the data) is not compatible with fairness (the null hypothesis, whose expectation is equal chances of heads or tails). In contrast, P = 0.005 gives S = 8, which is like having heads eight times in a row, a result much less compatible with the hypothesis that the coin was fair. However, as mentioned in the introductory paragraph of this section of the Appendix (see corresponding references), none of the alternatives to statistical inference is without potential pitfalls and, if used unwisely, all bring in their own sets of problems.

Statistical assumptions

In the previous section, I have already clarified that "a statistical model is a set of assumptions …[and] the model matches reality to the degree that assumptions are met" (Amrhein *et al.* 2019: 262–263). All statistical models make assumptions, but many of the assumptions are rarely tested or even reported in taxonomic studies using morphometrics. This criticism applies to my own work. Even if I rarely (can) rigorously test the main assumptions of the methods I use to compare groups, I make an effort to bear in mind their importance. The frequent violation of assumptions in taxonomic research using GMM is a potential, but sometimes inevitable source of inaccuracies (Cardini 2020a).

A long digression on all the assumptions of the methods used in this study is beyond my degree of expertise and the scope of this work. I suggest a few introductory references and, in the next paragraphs, briefly discuss some common issues in the context of taxonomic comparisons using GMM. Extensive discussions on many of the assumptions of statistical tests can be found in basic textbooks on univariate and multivariate statistics (Moore & McCabe 2005; Manly 2007; Sokal & Rohlf 2011; Hair *et al.* 2013; Howell 2013). For univariate data, Uttley (2019) is also a brief, but clear introduction to the main assumptions of popular methods, such as ANOVAs and regressions; likewise, for multivariate data, Stevens & James (2015) have an excellent chapter on the assumptions of the MANOVA. For multivariate statistics, also Neff & Marcus' chapter on study design, in their "Survey of multivariate methods for systematics" (Neff & Marcus 1980), is still an excellent starting point, although it does not cover all common techniques and later developments in statistical analysis of morphometric data. For instance, Neff & Marcus wrote before the development of the "comparative approach" that deals with the non-independence of species and lineages in macroevolutionary analyses (Felsenstein 1985; Huey *et al.* 2019). For those interested, however, a concise introduction to the 'comparative approach' in GMM is available in Monteiro (Monteiro 2013).

Statistical assumptions are typically difficult to verify in taxonomic data. For instance, univariate normality for parametric tests (t-tests, ANOVAs etc.) can be explored using a variety of techniques, including simple box-plots and frequency histograms. However, these plots must be done in each group and the assessment of normality (as with most other assumptions) is hard or impossible in small samples, simply because there are not enough data. If assessing normality is not always easy for univariate data, such as CS, the problem is far more serious for multivariate shape, for which the assumption must hold for all variables, as well as their linear combinations both in the total datasets and its subsets, when split by groups (Hair *et al.* 2013; Stevens & James 2015). Statistical programs, including PAST, offer some methods to formally test univariate (*Statistics, Normality tests*) and multivariate normality (*Multivar, Multivariate normality test*), but one should not blindly rely on these tests, which might make their own assumptions. There are also analyses that do not require normally distributed data. Permutation tests, and other resampling methods, do not assume normality because the frequency distribution of the test statistic is simulated from the available data (Zelditch *et al.* 2004). Resampling statistics is common, but not always available in user-friendly GMM software. Nevertheless, there are options to combine parametric and resampling approaches, although they might be testing slightly different hypotheses. This is something I exemplify in part B for group mean differences by using parametric ANOVAs (to simultaneously test sex and species) together with pairwise post-hoc permutation tests.

Homoscedasticity is another common assumption which concerns the similarity (sometimes called homogeneity) of variance and, for multivariate data, also covariance across groups. Homogeneous variance means that there is a similar amount of variability within each sample. For covariance, homogeneity means that, if two variables are correlated in one group, the sign and magnitude of the covariation must be similar in other groups, and this should hold for all pairs of variables. Homoscedasticity, unlike normality, is assumed not only by most parametric analyses of group mean differences, but also by the majority of tests using resampling statistics. For univariate analysis, the Levene's test of homogeneity of variance (included in both the two sample T-test and one-way ANOVA in PAST) is probably the best-known test for homogeneity of variances. For multivariate data, there are fewer options in the free software I used, although a two-group Box's M is available in PAST. Box's M is analogous to Levene, but designed to compare variance-covariance matrices of data with a multivariate normal distribution (Stevens & James 2015). The manual of PAST has a brief introduction on Box's M, including important caveats on its interpretation. Homoscedasticity can also be graphically explored. For CS, for instance, box and jitter-plots should be similar across groups. For shape, a canonical variate analysis, which is a multi-group DA used for summary scatterplots of group differences (as better explained in B4) should produce similar patterns of variation in all groups. In practice, this means having groups with circular variation of similar magnitude in the pairwise scatterplots that account for most of the between group differences (Albrecht 1992).

As with normality, small samples and heterogeneous sample size makes it harder to assess homoscedasticity in both univariate and multivariate data. Plotting the variables, however, is helpful to detect strong violations of homoscedasticity. For CS, box and whiskers plots should be of similar size across all groups. Figure 6 in part A clearly suggests that this is not the case in my North American marmot samples. Differences are not huge, but unequal sample size makes violations of homoscedasticity more serious in univariate and also in multivariate data (Howell 2013; Stevens & James 2015). Replicating analyses after excluding the smallest samples (as I do in part B) might help to assess the impact of heteroscedasticity, as well as of unequal N. Analyses could also be replicated using perfectly balanced random subsamples of each taxon (e.g., Seetah *et al.* 2016). However, when there is at least one very small sample, that might set the N of random balanced subsamples to a size that is too small for meaningful tests. The smallest sample(s) may, thus, be excluded before trying a design with random balanced subsamples. Yet, which is potentially more problematic, using subsamples reduces statistical power and, for shape, it is more likely lead to unfavourable p / N ratios. It is almost a truism that for robust results in the analysis of small variation one needs large and reasonably homogeneous sample sizes.

Linear regressions also assume homogeneity of variance, although the assumption concerns the distribution of residuals[23] (i.e., the variance unaccounted for by the predictor) in relation to the independent variable. In a bivariate regression, such as body mass on cranial CS (if one was, say, interested to predict weight using cranial size), the scatter of the data above and below the regression line should be similar across the entire range of CS. In a multivariate regression, as in the analysis of allometry, where Procrustes shape data are regressed on CS, exploring the homogeneity of residuals is less straightforward because they encompass as many dimensions as the original shape data (e.g., the 20 PCs of the Procrustes shape coordinates of the 12 landmarks on marmot mandibles). One might be tempted to explore if regression scores (see also B6), computed in MorphoJ, suggest a homogenous scatter, but regression scores only visualize shape information that maximally covaries with the predictor (CS, in allometric regression). Therefore, they cannot be used to assess homoscedasticity in the full space of the multivariate residuals. Other methods are appropriate to verify homoscedasticity in multivariate regression residuals (Caroni 1987), but none seems available in free user-friendly programs.

[23] Regression residuals should also be normally distributed and independent (i.e. uncorrelated with each other) (Hair *et al.* 2013).

Finally, a basic but fundamental assumption of the majority of statistical analyses is that of independent observations, which should be representative of the main patterns of variation in each of the study populations. Taxonomic data, however, almost always violate this assumption, as frequently stressed in this paper and discussed, with further examples, in Cardini (2020a). Groups, be they different species or different populations within a species, are not independent because of phylogeny and ancestry (Felsenstein 2002). Comparative methods, such as phylogenetic independent contrasts and phylogenetic generalized least squares have been developed to take the non-independence of species measurements into account (see Monteiro 2013, and references therein). However, not even the populations of a species, or the individuals within them, are really independent, because of potential autocorrelations. Autocorrelation, which is the correlation among the observations in a sample (instead of the variables, as in the common use of 'correlation'), might happen for a variety of reasons. Individuals can be autocorrelated because of genetics, in relation to different degrees of kinship. They can also be correlated because of environmental factors. For instance, food scarcity may limit size variation in a specific locality or population, thus potentially introducing a correlation between geographic distribution and size. Similarly, if a lineage follows the Bergmann's rule (Meiri 2011), size will covary with temperature, so that individuals living in colder climates might be on average larger than in milder regions.

Compared to the extensive development of comparative methods for interspecific analyses, much less work seems to have been done to address the problem of non-independence among populations and individuals within a species. In fact, in taxonomic comparisons, one needs to control for non-independence both between and within species. Thus, even if, for instance, using methods developed in spatial data analysis (Hawkins 2012), geographic proximity is used as a crude proxy for modelling non-independence among individuals in a species, that would do nothing to address the issue of the phylogenetic hierarchy among species. Using genetic data, rather than geography, to control for within-species autocorrelation may be more accurate, but, for now, there seem to be more problems than solutions. Thus, Felsenstein (2002) urges to avoid the temptation "to use molecular sequences to infer a "phylogeny" within the species and then to form [phylogenetically independent] contrasts [or equivalent comparative methods] based on that". This is because, he writes, "the phylogeny would reflect the coalescent genealogy of that particular locus [and] a different locus is expected to show a different coalescent. If populations have been exchanging migrants for a long time, there is no reason to assume that there is an underlying treelike genealogy". In his 2002 paper, Felsenstein suggests a method that, assuming the availability of a migration matrix, which accurately estimates gene flow among populations, can control for the within-species non-independence due to genetics. However, he acknowledges that there are multiple levels of non-independence. One would have to combine phylogenies and migration matrices to allow both between and within-species inferences, which is difficult and, to my knowledge, has never been tried in taxonomic comparisons.

Various types of strong autocorrelation, but also poor representativeness of variation in and among populations, happen also in archaeological and palaeontological material. Taxonomic comparisons in palaeontology are, almost by definition, harder than in neontology. Fossils are rare, often fragmentary and typically occur in heterochronic clusters across localities, thus making the issue of non-independence even more serious. Problems with gaps in the distribution, as well as a degree of heterochronicity of the individuals in a sample, are, in fact, not unlikely also in studies of modern species. Especially when working on endangered taxa or large animals, taxonomists often rely opportunistically on what is available in museums. Even when specimens are collected ad hoc, it is infrequent that a research can obtain a synchronic sample large enough that it uniformly covers at an appropriate spatial resolution the entire geographic range of a taxon. These are all serious problems that should be considered and mitigated as much as possible by improving sampling. When a potential bias in sampling cannot be controlled, it must be acknowledged (Cardini 2020a).

Semilandmarks: pros and cons?

A main 'omission' in the study case of the North American marmots, among some of the topics that can be of interest for taxonomists, is semilandmarks. Semilandmarks, as briefly anticipated in the main text, are a 'trick' to quantify curves and surfaces with no clearly corresponding landmarks. How a set of landmarks might represent 'homologous' features, and the meaning of homology in morphometrics (Smith 1990), is a complex and controversial argument in itself (Klingenberg 2008; Oxnard & O'Higgins 2009; Meik *et al.* 2020). Semilandmarks further increase the complexity of this discussion (Cardini 2020b). They may be useful, when really needed. For instance, they potentially add important details in specific contexts, such as individual identification (e.g., Kieser *et al.* 2007; Baralle *et al.* 2021), the virtual reconstruction of fossils (e.g., Gunz *et al.* 2009; Schlager *et al.* 2018), specific biomechanical applications (O'Higgins *et al.* 2011), and, with limitations in terms of biological insight and interpretation, the comparison of outlines or surfaces with virtually no landmarks (e.g., Ponton 2006; Hublin *et al.* 2009; Sanfilippo *et al.* 2010; Ros *et al.* 2014). In contrast, the morphological details captured by semilandmarks, when used in taxonomy, might often imprecisely pick up 'noisy' within-species variation, irrelevant at the level of population or species differences. This problem is not meant to be there all the time and its impact will vary from case to case, but should be borne in mind as a potential risk when the landmark configuration for a specific study is designed.

Despite a persistent denialism of the peculiarity and limitations of these special points (see references in Cardini 2020b), semilandmarks cannot by definition produce the same information as landmarks (Cardini 2013, 2020a, 2020b; Cardini & Loy 2013). They add a level of arbitrariness, which is not 'fixed' by claiming the equivalence of whole curves or surfaces. Within the continuous curvature of the inferior part of the marmot mandible horizontal ramus, for example, if measured using semilandmarks, the same semilandmark (or even a group of closely spaced semilandmarks) may be mapping on different anatomical features across specimens. The i-th point on the curvature could be approximately where the imprecise L12 is in Figure 1, but lie slightly (or much) to the left of the incisura vasorum facialum in one individual and to the right in another one. In the first individual, this semilandmark maps onto the incisor alveolus, but in the second individual is on the masseteric ridge, a developmentally and functionally different region of the mandible (Atchley *et al.* 1992; Klingenberg *et al.* 2001). Even if the researcher could split the series of semilandmarks in two, so that one series is on the lower margin of the incisor alvelous and the other marks the curvature of the region of insertion of the masseter, the problem is not solved, but simply shifted within each of the two regions. Each semilandmark inaccurately maps the anatomical correspondence among individuals, otherwise it would be replaced by a landmark. The visualization and the shape distances one obtains are a function of the specific position and density of each and every semilandmark; if not, following those who claim that what matters is only the overall correspondence of the curve/surface (Gunz & Mitteroecker 2013), we should get identical shape distances using semilandmarks regardless of density and the specific position of each point, but this does not happen. As one varies the number, density and mathematical treatment of semilandmarks, the quantitative description of the same shape changes. Curves (or surfaces) may be homologous, but shape distances depend on each point in a configuration: modify their number or position and shape distances are changed as well. Even if varying the semilandmarks leaves shape distances proportional, and thus does not alter appreciably the description of the similarity relationship in a sample, that demonstrates precision and is no proof of accuracy: all those descriptions could be similarly biased and inaccurate (Cardini 2020b). Sliding semilandmarks by minimizing bending energy or any other purely mathematical quantity, devoid of any biological interpretation, does not fix the issue of 'homology' either, contrary to the erroneous claim by the proponents of this approach. For instance, it has been misleadingly stated (Gunz *et al.* 2005: 25) that "semilandmarks like these [i.e., slid using the minimum bending energy criterion] can then be treated as homologous, without artifact", and also (Gunz & Mitteroecker 2013: 107) that "for larger shape variation and more extensive sliding, minimizing bending energy usually leads to better results that are in line with our notion of biological

homology". Yet, behind these authoritative statements, no proof of their accuracy was given and this is because there is no such proof, except, and dubiously, in special, carefully built, *ad-hoc* examples, such as the famous transformation of a shorter rectangle into a longer one, discussed in Cardini (2020b). Sliding changes the relative positions of the semilandmarks, but the algorithm 'knows nothing' about biological correspondence. In contrast, it certainly changes the covariance structure in the configuration, with a change, that might look small, but has no biological model behind and can be large enough to dramatically alter the results of analyses, such as a simple PCA or some of the analyses of modularity and integration (Cardini 2019, 2023; Zelditch & Swiderski 2023).

There is at least one more reason for a cautious use of semilandmarks. They necessarily increase the dimensionality of the data, with all the inevitable problems this brings in. In a highly dimensional space, it becomes harder to accurately summarize variation. Patterns may emerge, in results, that are largely an artefact of picking up noise in spaces of huge dimensionality (Cardini 2019; Cardini *et al.* 2019; Rohlf 2021). Putative increases in the 'signal to noise ratio', because of the larger amount of information due to the inclusion of semilandmarks, have never, to my knowledge, been demonstrated (Cardini 2020b). They may happen, but this is not by default. Depending on the case, semilandmarks may be informative or just make the data noisier and less accurate. Measuring more simply does not equate to measuring better. Semilandmarks should not be added because they make a nicer visualization, are fashionable and, now, easy to digitize manually or even automatically (Zhang *et al.* 2022). Researchers should carefully evaluate the pros and cons of semilandmarks and then decide whether they are really important for a specific research question (a taxonomic one, in our case).

Glossary of selected terms

I provide informal, simplified definitions for some of the main technical, mostly statistical, terms that I am using in the papers (part A and B). Rigorous definitions can be found in textbooks, as well as introductory papers written by professional statisticians and morphometricians.

- Accuracy and precision: in a scientific study, accuracy refers to how close one gets to the true answer, whereas precision is about how often one gets the same answer (regardless of it being correct or not). The assessment of ME, outlined in this study, is about precision.

- Alpha, significance threshold: the arbitrary cutoff chosen to reject a null hypothesis based on the P value of a test. It can be interpreted as the maximum risk one is ready to take that the data lead to reject the null hypothesis by mistake. For instance, with alpha = 0.05, the null hypothesis is rejected only when P, the probability of the data to be compatible with the null hypothesis, is less than 5%.

- Autocorrelation: positive autocorrelation may occur when, for instance, geographically closer specimens tend to be more similar to each other, for a specific variable, than expected by chance alone. Similarly, observations that are closer in time may be more similar than random expectations. In these examples, the individual observations, that are not independent, might behave like 'pseudo-replicates' (in a taxonomic comparison, a type of inaccurate replication of the 'unit of analysis', which can inflate sample size and lead to inaccurate estimates of degrees of freedom, and other issues, in statistical tests). In taxonomic samples, the causes of non-independence could be higher gene flow in a local population (endogenous cause, because of kinship) or locally similar environmental pressures (exogenous cause, because of sympatry), which both potentially lead to stronger morphological similarity among specimens from neighbouring locations than between distant ones.

- Bias: directional or systematic error. This term in the paper is mostly used in context of ME, but its meaning is broader. If there is a bias that affects similarly all observations, like, for instance, an inaccurate scale factor that leads to the same relative overestimate of CS in the entire study sample, the bias is

undesirable, but unlikely to change results. It is like adding a constant to all measurements. However, if the bias varies, that is potentially much more problematic. For instance, if CS is overestimated in a species sample, but not in those of the other species, differences between that species and all others will be inflated or deflated. How serious the problem is, thus, depends on the type and direction of the bias, and also on its magnitude.

- Covariate: any type of numerical variable. It can be a continuous variable (temperature, year, latitude, body mass etc.) or an ordinal one (a rank, usually coded with discrete numbers, like 1, 2, 3 etc., where 3 > 2 > 1 etc.). However, it can also be a dummy variable used to code groups (see part B). With dummy variables, the coding is arbitrary, but one has to be careful of the consequences on specific analyses: for instance, a multivariate regression on 0 for group A and 1 for group B is identical to a regression on 1 (for A) and 0 (for B), but it reverses the sign of the regression coefficients (which may matter if one is computing vector angles).

- Generalizability: it is about whether results (or, more generally, a claim) are valid not only within the context of a specific study (i.e., a specific scientific question, taxon, study material and method), but also have external validity, which means that they are accurate also in other contexts (e.g., different groups, measurements or structures).

- Group: it is a categorical variable; it could be sex, taxon or something else. Taxon, in GMM applied to taxonomy, is the most interesting type of group. Depending on the study, it could be populations within a species, subspecies or species. In MorphoJ groups are called classifiers, whereas in R they are called factors.

- Isotropic variation: in GMM, it means random uncorrelated differences of the same magnitude around each landmark; sometimes it is also referred to as ‘circular’ variation or ‘noise’.

- Multiple vs multivariate: these two terms are often used interchangeably; they both refer to a set of observations with many variables. However, in the context of regressions, multiple is when there is a single dependent variable and many predictors, whereas multivariate is when there is a single predictor (e.g., CS) and many dependent variables (e.g., shape). One can have both many dependent and many independent variables, thus making the regression multivariate multiple.

- p/N ratio: number of variables (p) relative to sample size (N). For instance, in the main analysis sample (N = 445), considering the four degrees of freedom lost in the Procrustes superimposition of 2D data, with 12 pairs of Procrustes shape coordinates, p / N = (12*2 – 4) / 445 = 0.045 or ~ 1 / 22, which means that there are 22 individuals for each variable in the analysis. However, within species, the average p / N is smaller (20 / 74 = 0.3, i.e. ~ three individuals per variable).

- P value: probability of the data being compatible with the null hypothesis, assuming the null hypothesis is correct (and all other assumptions of a model are met).

- R square (Rsq): variance accounted for by predictors in a statistical model (e.g., groups in an ANOVA or independent variables in a regression).

- SS: sum of squared deviations from the mean. See also variance.

- Statistical power: probability of correctly rejecting the null hypothesis, when really false. It is equal to one minus the type II error rate.

- Test statistics: the 'quantity' being tested for significance (e.g., F ratio, Rsq or the simple Euclidean distance between two means).

- Type I error rate: probability of incorrectly rejecting the null hypothesis. If a test is valid, the type I error rate should be less than or equal to alpha (i.e., using alpha = 0.05, if the test is repeated many times on new samples and the null hypothesis is correct, the null hypothesis should not be rejected more than 5% of the time).

- Type II error rate: probability of failing to reject the null hypothesis, when it was incorrect.

- Variance: for univariate data, the definition is the usual one: the sum of squared deviations from the mean (SS) divided by N - 1, which becomes the standard deviation (SD) if one takes the square root. Variance estimates the dispersion of the values of a variable around its mean. When I informally use terms such as "variation" or "variability", I am referring in general to any measure of dispersion of the data (variance, SD but also, for instance, the range of a variable). For multivariate data, the full pattern of variability in a set of variables is captured by the variance-covariance matrix (the symmetric square matrix with variances, of each variable, on the main diagonal and pairwise covariances off the main diagonal). When one is only interested in the overall magnitude of multivariate variance, however, that can be measured using different statistics. The most common one is the sum of the variances of each variable, which is also called the trace of the variance-covariance matrix. Summing the diagonal elements of the variance-covariance matrix produces the same results as as the sum of the eigenvalues in a PCA done using the variance-covariance matrix. In GMM, the magnitude of multivariate variance is often used to estimate shape disparity, which is the amount of variability in shape data in a taxon or lineage.

References

Adams H.H. 2019. Stats: a trillion P values and counting. *Nature* 569 (7756): 336–337.

Albrecht G. 1992. Assessing the affinities of fossils using canonical variates and generalized distances. *Human Evolution* 7 (4): 49–69.

Amrhein V., Trafimow D. & Greenland S. 2019. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *The American Statistician* 73 (sup1): 262–270. https://doi.org/10.1080/00031305.2018.1543137

Atchley W.R., Cowley D.E., Vogl C. & McLellan T. 1992. Evolutionary divergence, shape change, and genetic correlation structure in the rodent mandible. *Systematic Biology* 41 (2): 196–221. https://doi.org/10.1093/sysbio/41.2.196

Baralle G., Marchal A.F.J., Lejeune P. & Michez A. 2021. Individual identification of cheetah (*Acinonyx jubatus*) based on close-range remote sensing: first steps of a new monitoring technique. *Remote Sensing* 13 (6): 1090. https://doi.org/10.3390/rs13061090

Benjamin D.J. & Berger J.O. 2019. Three recommendations for improving the use of p-values. *The American Statistician* 73 (sup1): 186–191. https://doi.org/10.1080/00031305.2018.1543135

Betensky R.A. 2019. The p-value requires context, not a threshold. *The American Statistician* 73 (sup1): 115–117. https://doi.org/10.1080/00031305.2018.1529624

Cardini A. 2013. *Geometric Morphometrics.* Biological Science Fundamental and Systematics. UNESCO, Encyclopedia of Life Support Systems (EOLSS), Oxford, UK.
Available from http://www.eolss.net [accessed 5 Apr. 2024].

Cardini A. 2019. Integration and modularity in Procrustes shape data: is there a risk of spurious results? *Evolutionary Biology* (46): 90–105. https://doi.org/10.1007/s11692-018-9463-x

Cardini A. 2020a. Modern morphometrics and the study of population differences: good data behind clever analyses and cool pictures? *The Anatomical Record* 303 (11): 2747–2765. https://doi.org/10.1002/ar.24397

Cardini A. 2020b. Less tautology, more biology? A comment on "high-density" morphometrics. *Zoomorphology* 139 (4): 513–529. https://doi.org/10.1007/s00435-020-00499-w

Cardini A. 2023. Shall we all adopt, with no worries, the 'within a configuration' approach in geometric morphometrics? A comment on claims that the effect of the superimposition and sliding on shape data is "not an obstacle to analyses of integration and modularity". *EcoEvoRxiv*, unpublished preprint. https://doi.org/10.32942/X2002C

Cardini A. & Loy A. 2013. On growth and form in the computer era: from geometric to biological morphometrics. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 1–5. https://doi.org/10.4404/hystrix-24.1-8749

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 (4): 303–316. https://doi.org/10.1007/s11692-019-09487-5

Cardini A., Elton S., Kovarovic K., Strand Viðarsdóttir U. & Polly P.D. 2021. On the misidentification of species: sampling error in primates and other mammals using geometric morphometrics in more than 4000 individuals. *Evolutionary Biology* 48 (2): 190–220. https://doi.org/10.1007/s11692-021-09531-3

Caroni C. 1987. Residuals and influence in the Multivariate Linear Model. *Journal of the Royal Statistical Society: Series D (The Statistician)* 36 (4): 365–370. https://doi.org/10.2307/2348833

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90 (3): 891–904.

Felsenstein J. 1985. Phylogenies and the Comparative Method. *The American Naturalist* 125 (1): 1–15. https://doi.org/10.1086/284325

Felsenstein J. 2002. Contrasts for a within-species comparative method. *In*: Slatkin M. & Veuille M. (eds) *Modern Developments in Theoretical Population Genetics: the legacy of Gustave Malecot*. New York: Oxford University Press.

Goodman S.N. 2016. Aligning statistical and scientific reasoning. *Science* 352 (6290): 1180–1181. https://doi.org/10.1126/science.aaf5406

Greenland S. 2019. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. *The American Statistician* 73 (sup1): 106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland S., Senn S.J., Rothman K.J., Carlin J.B., Poole C., Goodman S.N. & Altman D.G. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31 (4): 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gunz P. & Mitteroecker P. 2013. Semilandmarks: a method for quantifying curves and surfaces. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 103–109. https://doi.org/10.4404/hystrix-24.1-6292

Gunz P., Mitteroecker P. & Bookstein F.L. 2005. Semilandmarks in three dimensions. *In*: Slice D.E. (ed.) *Modern Morphometrics in Physical Anthropology*: 73–98. Kluwer Academic Publishers-Plenum Publishers, New York.

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 (1): 48–62. https://doi.org/10.1016/j.jhevol.2009.04.004

Haaf J.M., Ly A. & Wagenmakers E.-J. 2019. Retire significance, but still test hypotheses. *Nature* 567 (7749): 461–461. https://doi.org/10.1038/d41586-019-00972-7

Hair J.F., Black W.C., Babin B.J. & Anderson R.E. 2013. *Multivariate Data Analysis*. Pearson Education Limited.

Hawkins B.A. 2012. Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography* 39 (1): 1–9. https://doi.org/10.1111/j.1365-2699.2011.02637.x

Hollenbeck J.R. & Wright P.M. 2017. Harking, sharking, and tharking: making the case for post hoc analysis of scientific data. *Journal of Management* 43 (1): 5–18. https://doi.org/10.1177/0149206316679487

Howell D.C. 2013. Statistical methods for psychology (Eight Edition). *Vermont: Wadsworth*.

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 (6): 777–785. https://doi.org/10.1016/j.jhevol.2009.09.001

Huey R.B., Garland T. & Turelli M. 2019. Revisiting a key innovation in evolutionary biology: Felsenstein's "Phylogenies and the comparative method". *The American Naturalist* 193 (6): 755–772. https://doi.org/10.1086/703055

Ioannidis J.P. 2019. Retiring statistical significance would give bias a free pass. *Nature* 567 (7749): 461–462. https://doi.org/10.1038/d41586-019-00969-2

Jeffreys H. 1961. *Theory of Probability*. Oxford, Oxford University Press.

Kieser J.A., Bernal V., Neil Waddell J. & Raju S. 2007. The uniqueness of the human anterior dentition: a geometric morphometric analysis. *Journal of Forensic Sciences* 52 (3): 671–677. https://doi.org/10.1111/j.1556-4029.2007.00403.x

Klingenberg C.P. 2008. Novelty and "homology-free" morphometrics: what's in a name? *Evolutionary Biology* 35 (3): 186–190. https://doi.org/10.1007/s11692-008-9029-4

Klingenberg C.P. 2013. Visualizations in geometric morphometrics: how to read and how to make graphs showing shape changes. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 15–24. https://doi.org/10.4404/hystrix-24.1-7691

Klingenberg C.P., Leamy L.J., Routman E.J. & Cheverud J.M. 2001. Genetic architecture of mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics. *Genetics* 157 (2): 785–802. https://doi.org/10.1093/genetics/157.2.785

Maher J.M., Markey J.C. & Ebert-May D. 2013. The other half of the story: effect size analysis in quantitative research. *CBE—Life Sciences Education* 12 (3): 345–351. https://doi.org/10.1187/cbe.13-04-0082

Manly B.F.J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. CRC Press, Boca Raton, FL.

Marcus L.F. 1990. Traditional morphometrics. *In*: Rohlf F.J. & Bookstein F.L. (eds) *Proceedings of the Michigan Morphometrics Workshop*: 77–122. Special Publication 2, University of Michigan Museum of Zoology.

Meik J.M., Lawing A.M. & Watson J.A. 2020. Use of scalation landmarks in geometric morphometrics of squamate reptiles: a comment on homology. *Zootaxa* 4816 (3) 397–400. https://doi.org/10.11646/ZOOTAXA.4816.3.12

Meiri S. 2011. Bergmann's Rule – what's in a name? *Global Ecology and Biogeography* 20 (1): 203–207. https://doi.org/10.1111/j.1466-8238.2010.00577.x

Monteiro L. 2013. Morphometrics and the comparative method: studying the evolution of biological shape. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 25–32. https://doi.org/10.4404/hystrix-24.1-6282

Moore D.S. & McCabe G.P. 2005. *Introduction to the Practice of Statistics*. WH Freeman & Co.

Muff S., Nilsen E.B., O'Hara R.B. & Nater C.R. 2022. Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution* 37 (3): 203–210. https://doi.org/10.1016/j.tree.2021.10.009

Neff N.A. & Marcus L.F. 1980. *A Survey of Multivariate Methods for Systematics*. American Museum of Natural History, New York.

O'Higgins P., Cobb S.N., Fitton L.C., Gröning F., Phillips R., Liu J. & Fagan M.J. 2011. Combining geometric morphometrics and functional simulation: an emerging toolkit for virtual functional analyses. *Journal of Anatomy* 218 (1): 3–15. https://doi.org/10.1111/j.1469-7580.2010.01301.x

Oxnard C. & O'Higgins P. 2009. Biology clearly needs morphometrics. does morphometrics need biology? *Biological Theory* 4 (1): 84–97. https://doi.org/10.1162/biot.2009.4.1.84

Ponton D. 2006. Is geometric morphometrics efficient for comparing otolith shape of different fish species? *Journal of Morphology* 267 (6): 750–757. https://doi.org/10.1002/jmor.10439

Rohlf F.J. 2021. Why clusters and other patterns can seem to be found in analyses of high-dimensional data. *Evolutionary Biology* 48 (1): 1–16. https://doi.org/10.1007/s11692-020-09518-6

Rohlf F.J. & Marcus L.F. 1993. A revolution morphometrics. *Trends in Ecology & Evolution* 8 (4): 129–132. https://doi.org/10.1016/0169-5347(93)90024-J

Ros J., Evin A., Bouby L. & Ruas M.-P. 2014. Geometric morphometric analysis of grain shape and the identification of two-rowed barley (*Hordeum vulgare* subsp. *distichum* L.) in southern France. *Journal of Archaeological Science* 41: 568–575. https://doi.org/10.1016/j.jas.2013.09.015

Sanfilippo P.G., Cardini A., Sigal I.A., Ruddle J.B., Chua B.E., Hewitt A.W. & Mackey D.A. 2010. A geometric morphometric assessment of the optic cup in glaucoma. *Experimental Eye Research* 91 (3): 405–414. https://doi.org/10.1016/j.exer.2010.06.014

Schlager S., Profico A., Vincenzo F.D. & Manzi G. 2018. Retrodeformation of fossil specimens based on 3D bilateral semi-landmarks: implementation in the R package "Morpho". *PLoS One* 13 (3): e0194073. https://doi.org/10.1371/journal.pone.0194073

Seetah K., Cardini A. & Barker G. 2016. A 'long-fuse domestication' of the horse? Tooth shape suggests explosive change in modern breeds compared with extinct populations and living Przewalski's horses. *The Holocene* 26 (8): 1326–1333. https://doi.org/10.1177/0959683616638436

Smith G.R. 1990. Homology in morphometrics and phylogenetics. *In*: *Proceedings of the Michigan Morphometrics Workshop. University of Michigan Museum of Zoology, Ann Arbor*: 325–338. Special Publication 2, University of Michigan Museum of Zoology.

Sokal R.R. & Rohlf F.J. 2011. *Biometry*. Macmillan Higher Education.

Stevens K.A.P. James P. 2015. *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS, Sixth Edition*. Sixth Edition. Routledge, New York.

Stephens P.A., Buskirk S.W. & del Rio C.M. 2007. Inference in ecology and evolution. *Trends in Ecology & Evolution* 22 (4): 192–197. https://doi.org/10.1016/j.tree.2006.12.003

Uttley J. 2019. Power analysis, sample size, and assessment of statistical assumptions — improving the evidential value of lighting research. *LEUKOS* 15 (2–3): 143–162. https://doi.org/10.1080/15502724.2018.1533851

Wasserstein R.L. & Lazar N.A. 2016. The ASA statement on p-values: context, process, and purpose. *The American Statistician* 70 (2): 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weinberg C.R. 2001. It's time to rehabilitate the P-value. *Epidemiology* 12 (3): 288–290. https://doi.org/10.1097/00001648-200105000-00004

Zelditch M., Swiderski D., Sheets D. & Fink W. 2004. *Geometric Morphometrics for Biologists: A Primer*. Elsevier Academic Press, Waltham, MA. https://doi.org/10.1016/B978-0-12-778460-1.X5000-5

Zelditch M.L. & Swiderski D.L. 2023. Effects of Procrustes superimposition and semilandmark sliding on modularity and integration: an investigation using simulations of biological data. *Evolutionary Biology* 50 (2): 147–169. https://doi.org/10.1007/s11692-023-09600-9

Zhang C., Porto A., Rolfe S., Kocatulum A. & Maga A.M. 2022. Automated landmarking via multiple templates. *PLoS One* 17 (12): e0278035. https://doi.org/10.1371/journal.pone.0278035

Zhang H. 2019. Stats: educate P-value abusers. *Nature* 569 (7756): 336–336. https://doi.org/10.1038/d41586-019-01530-x