Journal of Archaeological Science: Reports 32 (2020) 102423

Contents lists available at ScienceDirect

Journal of Archaeological Science: Reports

journal homepage: www.elsevier.com/locate/jasrep

Modern methods for old data: An overview of some robust methods for outliers detection with applications in osteology

Frédéric Santos

[a] Université de Bordeaux, UMR 5199 PACEA, Bâtiment B8, Allée Geoffroy Saint-Hilaire, CS 50023, 33615 Pessac Cedex, France

ARTICLE INFO

Keywords:
Isolation forests
MAD
Robust Mahalanobis distance
Robust statistics
R language

ABSTRACT

Whereas outlier detection is routinely performed in archaeological sciences and may have a substantial impact on subsequent discussion and interpretations, modern and robust methods are rarely employed in our disciplinary field. The detection of univariate outliers mainly relies on the well-known rule of "sample mean plus or minus two standard deviations", whose the lack of robustness is illustrated in this article. Furthermore, specific and efficient methods for multivariate outliers seem to be very little known and rarely used through the literature published in the *Journal of Archaeological Science: Reports*. To fill this gap, this article aims to present and summarize some robust methods well suited to the data usually gathered in archaeological and anthropological sciences, for both univariate and multivariate outliers. Robust methods for correlation and linear regression, whose results remain correct even in presence of strong outliers, are also illustrated. Methodological guidelines are discussed, in the light of applications in osteology. All the results (figures and tables) presented in this article can be fully reproduced with the companion R code available online, thus providing to the researchers some examples of templates for outliers detection.

1. Introduction

According to the intuitive definition formulated by Hawkins (1980, p.1), an outlier is "*an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*". Detecting outliers is an important step, either upstream of statistical analyses or as a goal in its own right. Outliers may be due to various sources of error such as entry errors, strong measurement errors, or artifacts that may arise at different steps of data acquisition in virtual anthropology. But some outlying values may also reveal "true" anomalies in the data, and then bring important and relevant information, for they can contribute to identify pathological individuals (Dietmeier, 2018) or individuals having too unusual values to be part of a given human group. The latter case is particularly frequent in isotopic studies where outliers (e.g. for $\delta^{13}C$, $\delta^{15}N$ or $^{87}Sr/^{86}Sr$ values) might indicate the presence of non-local individuals, thus allowing to discuss migrations and mobility patterns among human groups (e.g., Santana-Sagredo et al., 2015; Hakenbeck et al., 2010). Outliers detection can thus have important consequences on subsequent interpretations.

When no pre-existing data providing a range of credible values for a given population can be used—which is the general case in archaeological sciences—this range of credible values must be estimated with statistical methods, traditionally using location and scale estimates calculated on the sample itself, that are supposed to accurately reveal the true parameters of the underlying population. The data used for those calculations thus include the potential outliers, which raises a crucial problem: if those location and scale estimates are non-robust, i.e. strongly influenced by the presence of outliers, they may fall far from the true population parameters, thus invalidating the whole procedure.

The handling of outliers often suffers from several problems and misuses in archaeological sciences. First, the number and identity of outliers may vary depending on the method employed (Lightfoot et al., 2014). Nonetheless, the method or criterion used to detect and identify outliers is not always explicitly specified in the scientific literature; this lack of precision is also frequent in other disciplinary subfields of social sciences (Leys et al., 2013). Second, the methods used in the literature are often not robust, and the decision rules used to identify outliers rely either on statistical indicators that are themselves imprecise in the presence of outliers, and/or a normality assumption which is not always clearly met (e.g., Wright, 2005; Webb et al., 2013). Finally, only a few publications utilize efficient and specific methods to detect multivariate outliers (e.g., Harris and Bailit, 1988; Mahoney, 2006; Algee-Hewitt, 2016): both the modern statistical methods for detecting multivariate outliers and their implementation in free software seem to be little known.

E-mail address: frederic.santos@u-bordeaux.fr.

https://doi.org/10.1016/j.jasrep.2020.102423
Received 20 September 2019; Received in revised form 23 April 2020; Accepted 2 June 2020

The problem of outlier detection is also closely related to the robustness of statistical methods. Indeed, outliers are often identified—and sometimes excluded—in search for a more "representative" sample to assess and discuss the correlation between two variables (e.g., Loftus and Sealy, 2012), or to build regression models (e.g., Beck and Smith, 2019). This article will focus on robust methods, both for detecting outliers themselves, and also for getting more precise statistical estimates even when outliers are present in a dataset. Some simple examples where those robust methods outperform more classical and widely used methods will be given. The applications mainly consider osteological data here, but the methods presented will also suit any dataset including one or several continuous variables. Various population samples extracted from the Goldman Data Set freely available online (Auerbach and Ruff, 2004) will be used through the text, and a subsample from the reference sample of the DSP2 sex estimation method (Bružek et al., 2017) will also be utilized in a case study.

The aim of this article is not to provide an exhaustive or practical in-depth review of all available methods of outlier detection. A comparison of several methods, applied on isotopic data, has recently been performed by Lightfoot and O'Connell (2016) in an enlightening article. Leys et al. (2019) recently published a methodological note to "*fill the lack of an accessible overview of best practices*" (p. 1) for outliers detection in the field of psychology. However, there is a strong need that a similar dissemination reaches the field of archaeological sciences. For instance—as of April 2020—, a research within the database of the articles published in *Journal of Archaeological Science: Reports* (https://www.sciencedirect.com/journal/journal-of-archaeological-science-reports/issues) triggered only four results for the keywords "median absolute deviation", three results for "bagplot", one result for "robust Mahalanobis", and no result could be found for requests about "multivariate outliers", "isolation forests" or the S_n estimator. As concerns robust methods, no article matched the keywords "robust regression" or "quantile regression". Those results are nearly identical—and sometimes even lower—in other journals more oriented towards biological anthropology, such as the *American Journal of Physical Anthropology* or the *International Journal of Osteoarchaeology*. Consequently, this article proposes a summary of the recent advances in statistics and provides ready-to-use R templates for modern methods of outlier detection.

Finally, to reinforce the move towards a reproducible research in archaeological and anthropological sciences (Marwick, 2017), this whole article has been written in Org-mode 9.3.6 for Emacs 26.3 (Schulte et al., 2012) and is fully reproducible with the org source file available on GitLab (https://gitlab.com/f-santos/reproducibility-package-for-santos-2020-jasr). Trying to follow the highest current standards of reproducibility (Desquilbet et al., 2019), a Docker image is also made available on DockerHub to provide the computational environment which allowed this study—full documentation on how using it can be retrieved from the GitLab repository. For a basic and simple reuse, the source codes of all tables and figures from this study are also available in separate R files. All the statistical analyses were performed with R 3.6.3 (R Core Team, 2020), and the list of R packages used is given in Appendix C.

2. Univariate outliers

This first section deals with outlying values for one single variable. To present a concrete archaeological case, the left–right differences in humerus maximum length observed on the hunter-gatherers of Ipituaq (US-AK, 1500–1100 BP) are used. Those data are extracted from the Goldman Data Set online. This population sample is known to exhibit a substantial amount of asymmetry for this measurement (Auerbach and Raxter, 2008). Since significant sex differences may be observed on the upper limbs for forager populations (Weiss, 2009), only the 14 male individuals whose humeral length is known on both sides are considered. This small sample also permits the discussion of the robustness of the several methods presented below with the sample sizes usually available in archaeological sciences.

2.1. The classical rule based on the sample mean and standard deviation

In biological anthropology, methods of outlier detection based on the mean and standard deviation are still frequently employed, including in recent research articles (e.g., Bergstrom et al., 2019; Lubritto et al., 2017). Any value out of the range defined by the mean plus or minus two or three standard deviation is then considered as an outlier. This criterion, also known as the "95–99.7 rule", is derived from the properties of the gaussian distribution: it is well known that about 95% and 99.7% of normally distributed values lie within two and three standard deviations from the mean respectively. This rule-of-thumb is both theoretically and practically correct when applied to a large enough sample for which the assumption of normality seems reasonable.

However, this method suffers from a critical lack of robustness in other situations, recently illustrated on real data from various disciplinary fields by Leys et al. (2013) and Lightfoot and O'Connell (2016). The data sets handled in archaeological sciences do not always meet the previous requirements, or it may at least be difficult to check them because of their small sample size. When considering archaeological data, the sample mean and—above all—standard deviation may be drastically distorded by the presence of the extreme outliers themselves, and thus do not provide a good measure of distance to detect outliers.

Fig. 1 provides an illustration of such a situation. The sample mean $\hat{\mu} = -2.929$ and the standard deviation $\hat{\sigma} = 5.129$ are strongly inflated because of the two extreme values located on the right tail. The lack of robustness of the "mean plus or minus two standard deviations" decision rule is revealed by the failure to exclude one of the two outliers, since its value falls within the range $[\hat{\mu} - 2\hat{\sigma} ; \hat{\mu} + 2\hat{\sigma}] = [-13.186 ; 7.329]$.

Albeit not artifical, the example presented here may be seen as peculiar, with a low sample size and two extreme values located on one single tail. However, it shows that this classical rule is clearly non-robust, and should only be used with much precaution and after a careful inspection of the data to ensure that the required assumptions are met.

2.2. Robust alternatives for gaussian data

If the assumption of a normal $\mathcal{N}(\mu, \sigma^2)$ distribution of the data—disregarding some potential extreme values—seems to be reasonable for a given variable, several alternatives sharing the same philosophy do exist. All of them consist in using location and scale estimates for μ and σ which are more robust than the classical sample choice of mean and standard deviation respectively. Consequently, the estimates calculated to define a "credible range of variation" outside of which any value can be considered as an outlier, are themselves less sensitive to the presence of outliers, thus always providing a more accurate estimation of the hidden population parameters.

For all the methods detailed in this section, the credible range of variation is defined by the following general formula, perfectly analagous to the "95–99.7 rule":

$$[m - k \cdot \hat{s} ; m + k \cdot \hat{s}] \quad (1)$$

where m is the sample median—a robust location estimate—, and $\hat{s}$ is a robust scale estimate (D'Orazio, 2017). The choice of a constant k, usually lying between 2 and 3, allows to exclude only clear outliers (if set to a high value, since the interval will be wider) or even slightly suspicious values (if set to a low value, since the interval will be narrower), depending on the goals of the study and the type of data. When dealing with very small sample sizes, a conservative choice $k = 3$ might seem preferable to avoid false positives (Leys et al., 2019).

Among several choices for robust location estimates $\hat{s}$ proposed in statistical literature, three will be compared below: the interquartile range (IQR), the median absolute deviation (MAD), and the S_n

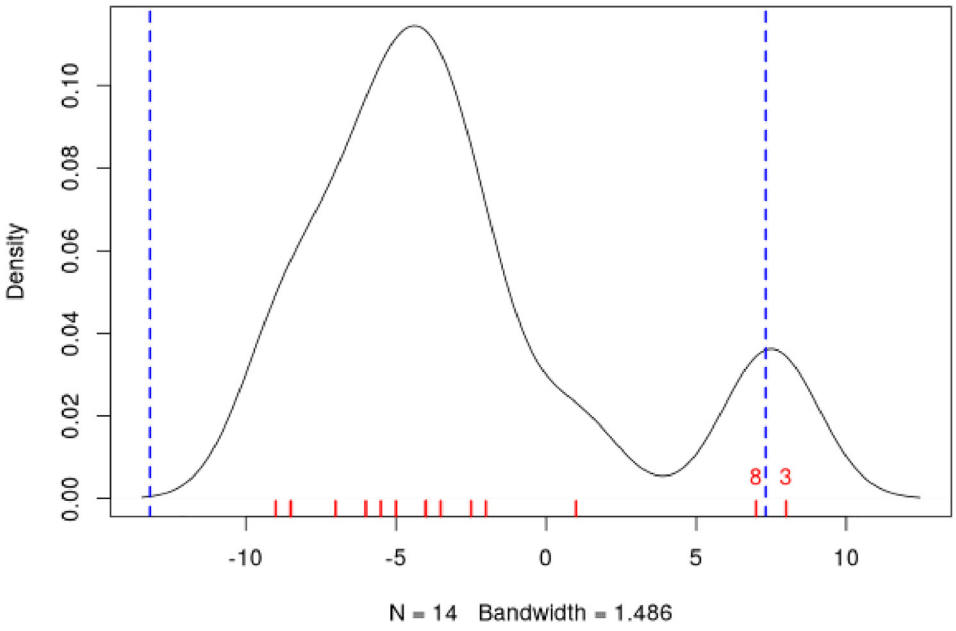

Fig. 1. Kernel density estimation of the vector x of left–right differences (in millimeters) in humeral length observed on the 14 male individuals from the population sample Ipituaq (US-AK, 1500–1100 BP) in the Goldman Data Set. The blue dotted vertical lines represent the exclusion thresholds defined by the classical rule based on the sample mean and standard deviation, equal to $\bar{x} \pm 2 \times \hat{\sigma}_x$. The third and eighth individuals are visual outliers.

Table 1
Comparison of four methods based on various location and scale estimates for outlier detection, applied on the data described in Fig. 1. "Coef" is the user-defined constant k used for the construction of intervals, see Eq. (1). The lower and upper bounds of the intervals built with each method are indicated in the corresponding columns. The last column indicates the ID of the individuals flagged as outliers.

	Location	Scale	Coef	Lower bound	Upper bound	Outliers
mean and sd	−2.929	5.129	2	−13.186	7.329	3
median and IQR	−4	2.78	2	−9.56	1.56	3, 8
median and MAD	−4	2.965	2	−9.93	1.93	3, 8
median and S_n	−4	3.578	2	−11.156	3.156	3, 8

estimator—full mathematical details for each of them are available in Appendix A. Those three estimators provide three different robust variants of formula (1), and therefore three acceptable decision rules for univariate outliers detection. To compare the results obtained with these variants to the results returned by the usual "95–99.7 rule", all four criteria were applied to the 14 male individuals from the Ipituaq population sample. The results can be found in Table 1.

It can be seen that, unlike the usual method based on non-robust estimates, the three robust methods detect both the individuals 3 and 8 as outliers. None of them suffer from the inflation of location and scale estimates—caused by the two outliers located on the right tail—that affects the usual method. As a consequence, at any given value of k, the interval they provide for outlier detection is much narrower, and more accurately captures the range of usual values for the humeral asymmetry in this population sample.

2.3. Robust methods which do not assume normality

In most contexts of archaeological sciences, such as osteometric or isotopic studies, there is almost always a presupposition of normality for all the variables considered—once again, discarding a few potential "true" outliers (e.g., migrants, pathological individuals or entry errors). As noted by Lightfoot and O'Connell (2016, p. 22), skewed data may simply indicate a sample with several outliers on the same distribution tail, as in Fig. 1.

Severely skewed distributions arise almost systematically in some disciplinary fields such as neurosciences (Rousselet and Wilcox, 2019). Specific methods have been proposed for such variables, and numerous formulas do exist depending on the degree of skewness observed on the data (Hubert and Vandervieren, 2008). Conversely, few variables studied by biological anthropologists or archaeologists are intrinsically far from normality. For those reasons, the need of specific methods for non-gaussian data is lower than in other disciplines. Consequently, the methods accounting for skewed distributions are to be used with caution, for they might lead to spurious results as it will be shown below.

As a general rule:

1. If the distribution may at least be considered as symmetric, the three robust variants exposed in Section 2.2 remain valid, albeit more difficult to use since their scale factors (a specific constant required for the computations) must be approximated through computer simulations (Rousseeuw and Croux, 1993).
2. If there is a good reason to suspect an asymmetric or skewed distribution in the whole underlying population, the use of a robust measure of skewness such as the medcouple (Brys et al., 2004) might constitute a useful first step. A high medcouple value (close to 1) may indicate that the variable is intrinsically skewed, i.e. exhibits a substantial skewness that is not only due to a few outliers.

In the general case of no particular assumption about the distribution of the variable, boxplot-based rules are a simple yet efficient way to proceed.

2.3.1. The classical boxplot rule

Boxplots are often used to detect univariate outliers. According to the standard boxplot rule (Tukey, 1977), the credible range of credible values (i.e., the boxplot *fences*) is defined by:

$$[q_1 - k \cdot IQR \,;\, q_3 + k \cdot IQR] \tag{2}$$

where q_1 and q_3 are the first and third empirical quartiles respectively. The constant k is traditionally set to 1.5, although more conservative values such as 2 or 3 are also admissible depending on the goals of the

study. It should be noted that this interval is centered around the arithmetic mean of q_1 and q_3 (which is usually not equal to the median) and is generally not symmetric.

This very general rule does not assume normality, but in the case of a large normal sample, about 0.35% of data points should be flagged as outliers at each end (i.e., 0.70% in total). However, this proportion may be different—much higher—in a symmetric heavy-tailed distribution.

2.3.2. Adjusted boxplots for skewed distributions

Some amendments to the previous rule have been proposed to achieve a better accuracy for skewed distributions. For slightly skewed distributions, Kimber (1990) proposed a rule based on so-called semi-interquartile ranges, and defined the following interval:

$$[q_1 - 2k\cdot(m - q_1)\,;\, q_3 + 2k\cdot(q_3 - m)] \tag{3}$$

using the notations previously introduced in Eq. (2), and a value of k still usually equal to 1.5.

2.3.3. Application to the Goldman Data Set

An example of visually slightly skewed distribution can be given by considering the asymmetry in tibia mediolateral diameter within the population sample of Giza (Egypt, 4700–4200 BP, shortcode in the Goldman Data Set: "Pyramiden, Gizeh"). A kernel density estimation of those values is presented in Fig. 2.

Out of any context, this distribution might simply be regarded as right-skewed, and asymmetric boxplot fences do not detect any outlier—not even the extreme individual 14. This basically means that *if one makes the assumption that tibial asymmetries are intrinsically right-skewed in the whole underlying population*, then no value can be regarded as an outlier in this sample. Such an asymmetry pattern might happen: as various subsets of a given population can present different degrees of directional asymmetry (Graham and Özener, 2016), a complex mixture of fluctuating asymmetry, differential directional asymmetry and/or antisymmetry might indeed end in a skewed distribution. However, if this—strong—assumption is false, accounting for skewness leads to misleading results, since this skewness would not be a characteristic of the underlying population but rather a side-effect of several outliers located on the right tail. Indeed, standard boxplot fences (not adjusted for skewness) do detect the individual 14 as a clear outlier in this population sample.

Accounting for skewed distributions is then a delicate matter and relies on strong biological assumptions that should definitely be supported by previous knowledge. The choice of a given method of outlier detection must not be based only on statistical considerations, but also depends on the biological knowledge about the variable and population studied (Leys et al., 2019).

3. Multivariate outliers

When several variables are involved, using specific methods is mandatory, and one should not rely only on a combination of univariate methods, although it may be a good starting point to get a basic understanding of the data (Unwin, 2019). Among many other available algorithms such as "Dbscan" (Ester et al., 1996) or "hdoutliers" (Wilkinson, 2018), two methods are detailed below, which are both conceptually rather simple and practically easy-to-use, and have efficient implementations in both R and Python languages.

3.1. Robust Mahalanobis distance

Unlike euclidean distance, Mahalanobis distance takes into account the correlation between the variables when computing dissimilarities among individuals. For this reason, it is popular in biological anthropology (Pilloud and Hefner, 2016), where the data suffers almost always from intercorrelation. In a formal way, Mahalanobis distance between an individual x_i (described by p variables) and the multivariate sample mean $\hat{\mu}$ is defined by:

$$D_i = \sqrt{{}^t(x_i - \hat{\mu})\Sigma^{-1}(x_i - \hat{\mu})} \tag{4}$$

where $x_i, \hat{\mu} \in \mathbb{R}^p$, and Σ is the $p \times p$ empirical covariance matrix.

The Mahalanobis distance can be used to detect multivariate outliers (e.g., Stynder, 2009). It is known to be primarily applicable to multivariate normal distributions—or at least elliptically symmetric unimodal distributions—although some studies suggest that its use can be generalized to some extent when the data depart from normality (Warren et al., 2011). The outliers are those individuals whose the distance to the centroid $\hat{\mu}$ is greater than $\sqrt{\chi^2_{p;1-\alpha}}$, i.e. the square-root of the $1-\alpha$ quantile of a Pearson distribution with p degrees of freedom. α may usually vary from 0.001 (for a very conservative rule) to 0.05 (for a not too conservative rule), depending on the aim of the study.

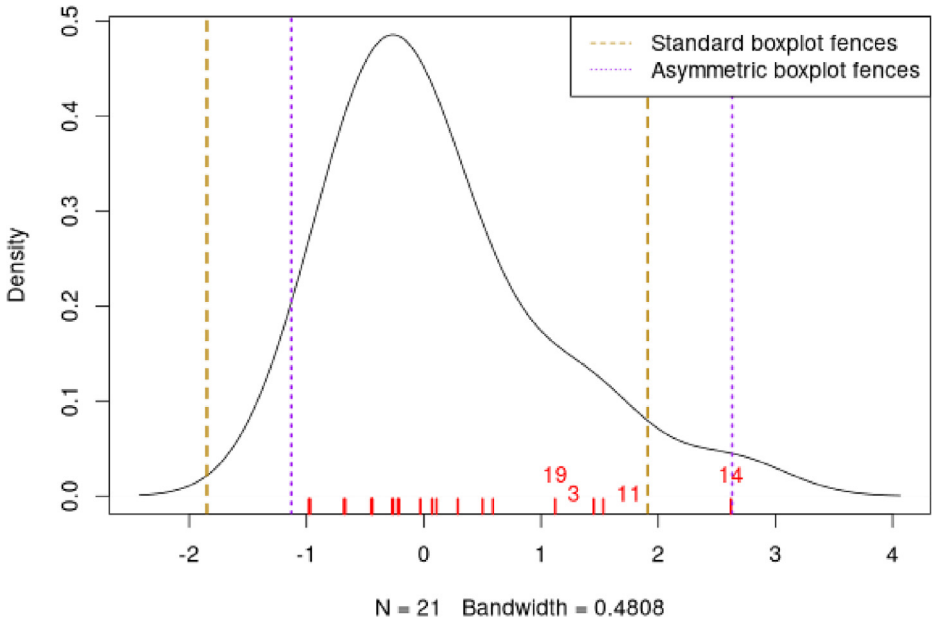

Fig. 2. Kernel density estimation of the vector right-left differences (in millimeters) in tibial mediolateral diameter observed on the 21 individuals from the population sample of Giza (Egypt, 4700–4200 BP) in the Goldman Data Set. The four most extreme individuals on the right tail are labeled in red.

This method is a generalization of the univariate rule relying on the sample mean and standard deviation, described in Section 2.1, and thus it suffers from the same lack of robustness. As for the "95–99.7 rule" in the univariate case, the estimates used in the formula (4) are non-robust and may be distorded by potential outliers, thus making invalid the whole decision rule.

A robust variant of Mahalanobis distance, also known as the MCD (minimum covariance determinant) algorithm, was proposed to circumvent these weaknesses (Rousseeuw and Van Driessen, 1999; Hubert et al., 2018). Intuitively, it can be seen as an iterative method that uses only the "good part of the data" (i.e., uncontaminated data) to derive a robust location estimate $\hat{\mu}_{\text{MCD}}$ and a robust variability estimate $\hat{\Sigma}_{\text{MCD}}$ which will be used instead of the classical $\hat{\mu}$ and $\hat{\Sigma}$ estimates in Eq. (4). As in the case of the classical Mahalanobis distance, the outliers are defined as those individuals whose robust Mahalanobis distance exceeds the threshold $\sqrt{\chi^2_{p;1-\alpha}}$. More mathematical details, along with basic guidelines to determine the "good part of the data", are available in Appendix B.

A simple (and easy to visualize) example may be used to illustrate the differences between the classical and robust versions of the Mahalanobis distance. Fig. 3 represents a three-dimensional scatterplot for the Sayala population sample, retrieved from the Goldman Data Set. The maximal lengths of three long bones, the left femur, humerus and tibia, are considered. Visually, three outliers—the individuals 7, 14 and 20—can be identified.

The presence of those outliers causes an inflation of the generalized variance, i.e. a distorsion of the classical covariance matrix Σ. Consequently, the classical and robust Mahalanobis distances provide different sets of outliers here (Fig. 4). For an α level of 0.01, the classical version detects no outlier at all, whereas the robust version identifies the two individuals 14 and 20. For an α level of 0.05, the robust version also detects the individual 7, which is still far from the exclusion boundary for the classical version.

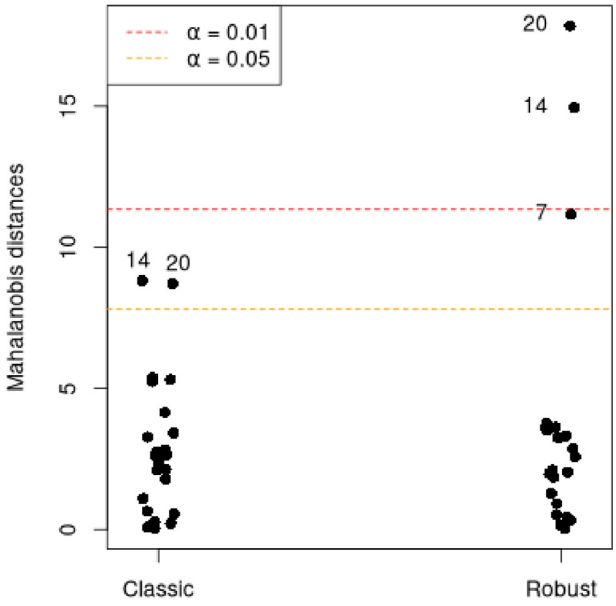

Fig. 4. Stripcharts displaying the squared classical and robust Mahalanobis distances between each individual and the centroid. The dotted lines symbolize the exclusion thresholds $\chi^2_{p;1-\alpha}$ for two different α values. The maximal lengths of three long bones from the population sample of Sayala (Goldman Data Set) were considered (LTML, LHML, LFML).

However, even the robust Mahalanobis distance presents some drawbacks that are likely to be encountered in archaeological sciences. First, Mahalanobis distance can only capture linear relationships between variables, and can deliver spurious results when non-linear

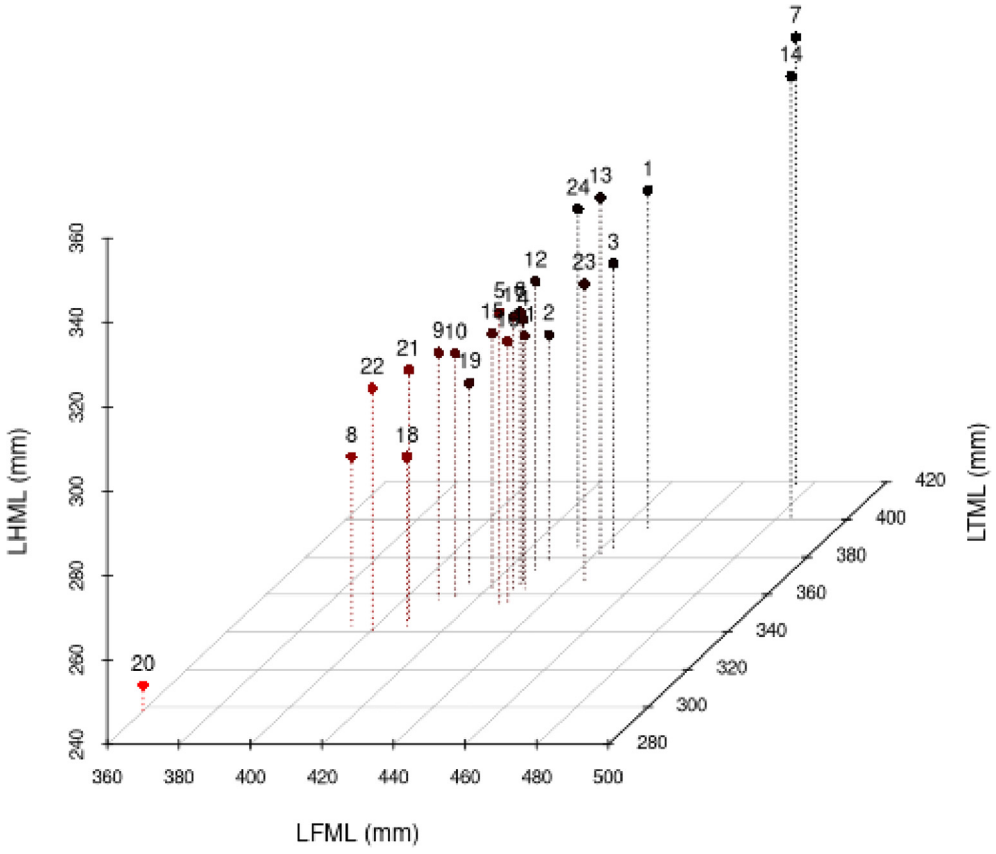

Fig. 3. 3D scatterplot of the population sample of Sayala, drawn from the Goldman Data Set. The maximal lengths of three long bones are represented.

patterns are involved. Second, to achieve a sufficient stability and accuracy in the estimation of the covariance matrix, the number of individuals should be greater than three times the number of variables (Harbottle, 1976). Combining these two limitations, it is safer to use Mahalanobis distances only when dealing with a small number of dimensions. In such a situation, one can verify that there are no complex non-linear relationships in the data—for example using a pairs plot—and it is easier to reach a sufficient sample size to ensure a reliable estimation of Σ.

3.2. Isolation forests

Given the limitations of the classical procedures based on Mahalanobis distances, isolation forests present a useful and very robust alternative, whose the use is safer in higher dimensions. Isolation forests are a recent algorithm of "anomaly detection" (Liu et al., 2012), based on random forests (Breiman, 2001). This method does not rely on any assumption about the distribution of the data, nor any given classical dissimilarity (e.g., euclidean, Mahalanobis).

The general idea is that "anomalies" can be defined by both their unusual values and their rarity, so that they are quite *isolated* in the data, and therefore easy to localize. Indeed, identifying a point located right in the middle of a point cloud will usually require numerous instructions, whereas one single instruction may be sufficient to describe an outlier (e.g., "this is the only individual with $X_5 > 250$").

An isolation forest corresponds to a set of *B isolation trees*, which are themselves randomly built decision trees that are grown until there is one single individual in each terminal leaf. Since outliers are supposed to be easily isolated in the data, they will correspond to the shortest paths in the isolation trees. A measure of credibility for an individual to be outlier is then its corresponding average path length within the *B* isolation trees. An anomaly score, lying in [0, 1] and being a function of the sample size and the average path length, is computed for each individual.

According to Liu et al. (2012), a quick rule-of-thumb can provide a first indication as concerns the presence of outliers: if all the individuals have anomaly scores very close or inferior to 0.5, there is likely no multivariate outlier at all in the data. Conversely, if some anomaly scores depart from 0.5 and raise closer to 1, the corresponding individuals are likely to be outliers.

An isolation forest with 100 isolation trees is built on the same data as in the previous section (Sayala population sample with three variables: LTML, LHML, LFML). The anomaly scores, sorted by decreasing order, can be found in Fig. 5. The isolation forest algorithm provides evidence to consider the individuals 20, 7 and 14 as outliers, since their anomaly scores are the only ones to exhibit a substantial departure from the reference value of 0.50. This conclusion is consistent with the results obtained via the robust Mahalanobis distance (cf. Fig. 4). Isolation forests can thus provide a useful indication about possible multivariate outliers, by studying both the global distribution of anomaly scores (in search for "elbows" or gaps) and their absolute distance to 0.50.

4. Cellwise outliers: a case study

Although they may correspond to different situations, the two multivariate methods presented in Section 3 still have a common drawback. They allow an identification of the most unusual data points, but they do not tell *why* those individuals differ from the typical observations, i.e. on which variables they present anomalous values. Such an investigation is sometimes possible by inspecting several simple graphical outputs, such as a pairs plot—which is a matrix of pairwise bivariate scatterplots. However, this becomes very time-consuming and difficult when the number of variables increases, and it does not allow the identification of all types of multivariate outliers. In such a case, one may think of principal component analysis as a way of finding the variables involved in the "outlyingness" of a given individual. But some outliers may be visible only on the few last principal axes (Jolliffe, 2002), which are usually not inspected. Therefore, in some situations, it may be quite difficult to figure out what is different about an individual detected as suspect by the robust Mahalanobis distance or isolation forests.

This problem is addressed by a recent algorithm called DDC, for Deviating Data Cells (Rousseeuw and Bossche, 2018). This algorithm seems to be particularly promising for osteoarchaeological studies, for it can handle missing values—to some extent—and allow rich and precise interpretations about the unusual measurements observed on an individual. In particular, this algorithm may allow to distinguish the individuals whose outlyingness is only due to their extremity on a single variable, and the individuals whose outlyingness is rather due to an unusual combination of values which would be perfectly acceptable when considered individually—i.e., "shape outliers".

DDC algorithm begins by finding potential extreme values on each single variable, and then looks for unusual combinations of

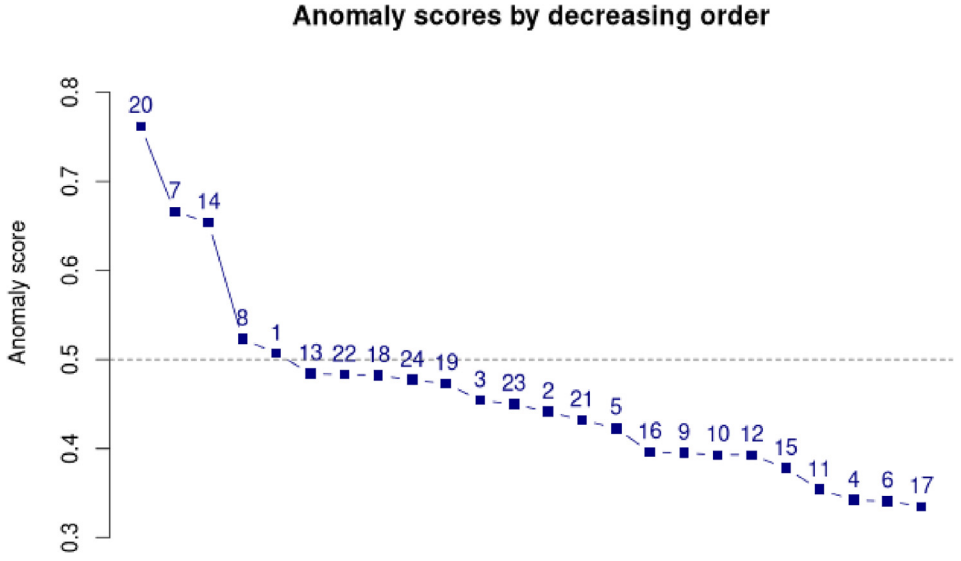

Fig. 5. Plot of the anomaly scores obtained by an isolation forest to detect outliers from the population sample of Sayala (Goldman Data Set), when three maximal lengths are considered (LTML, LHML, LFML). The scores are sorted in decreasing order and the corresponding individual IDs are indicated.

values—e.g., a rather long femur and a rather short tibia—by considering subsets of correlated variables. All data cells exhibiting anomalies are *flagged* in a graphical output: unusually low values are colored in blue, high values are colored in red, and all data cells presenting credible values are indicated in yellow. DDC therefore introduces a new paradigm in outlier detection, moving from *rowwise outliers* (individuals globally considered as anomalies) to *cellwise outliers* (each individual will usually have at most some flagged values, and still a bunch of credible values). One can also set the tolerance probability value, i.e. a cutoff value for flagging only extreme outliers or slightly unusual values (default value is 0.99).

This method can be illustrated on a subset of individuals extracted from the reference sample of DSP2. This subset is composed of 22 left ossa coxae belonging to male individuals from the Cleveland population sample. Following the DSP2 method, ten measurements have been collected on each os coxae, resulting in a small sample with only twice as many individuals than variables. With ten measurements, inspecting the 45 possible bivariate scatterplots is difficult and not necessarily informative, since the anomalies may imply combinations of four or more variables.

A PCA shows no clear outliers on the first three principal axes. When considering each variable separately, only three individuals stand out according to the classical boxplot rule (extensive results available as Supporting Information online). The individual 96 exhibits a low value for the variable PUM, the individual 108 may be seen as an outlier for the variables for SPU and SS, and the individual 64 for the variables SS and VEAC. Unsurprisingly, those three individuals, being easy to "isolate" from the rest of the data, are the best candidates to be regarded as outliers according to the anomaly scores derived by isolation forests (Fig. 6).

However, this not entirely the end of the story: some unusual combinations of variables can also be observed on other individuals. Fig. 7 shows the deviating data cells flagged by the DDC algorithm. The results already known from univariate analysis can usually also be retrieved on this plot: for instance, the individual 108 has indeed be flagged by the algorithm for having high values of SPU and SS, which confirms the results from the well known boxplot rule. However, many other cells are flagged, even for individual that show no univariate anomaly and have low anomaly scores in Fig. 6. For instance, individual

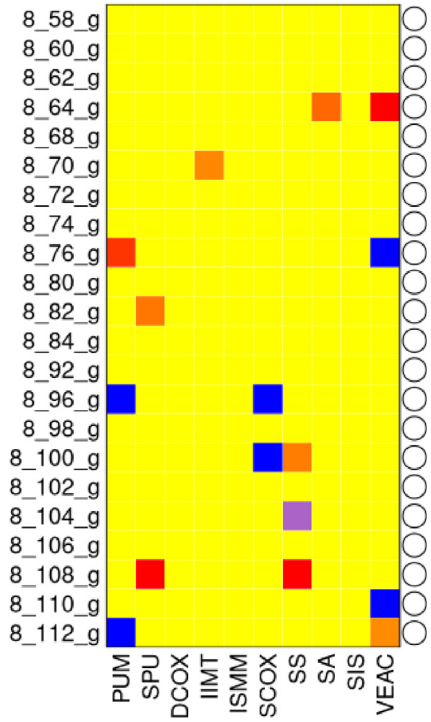

Fig. 7. Deviating data cells flagged by the DDC algorithm on 22 male individuals extracted from the DSP2 reference sample. Unusually low values are colored in blue (if strong anomaly) or purple (if slight), and high values are colored in red or orange. A tolerance probability of 0.975 has been used.

76 exhibits an unusual combination of high PUM and low VEAC measurements: none of those values stand out by themselves but both are atypical with respect to the values taken by the variables most correlated to them. The individual 112 exhibits exactly the reverse combination, with low PUM and high VEAC values. Similarly, the individual 100 exhibits a combination of a rather high SS and very low SCOX, which is also unusual within this population sample. Those peculiarities can indeed be confirmed when going back to the raw data, but the first

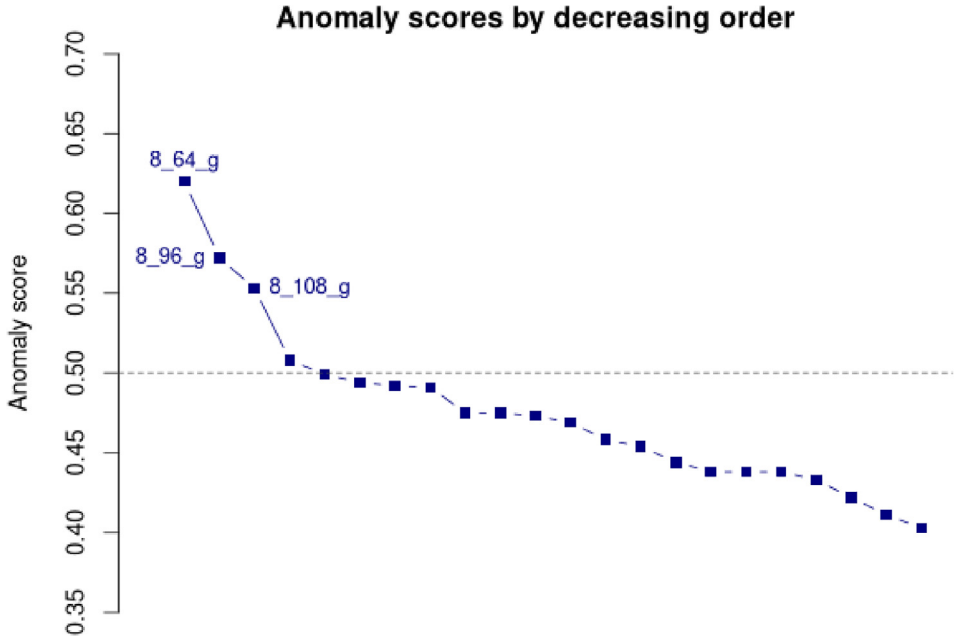

Fig. 6. Anomaly scores obtained with isolation forests for 22 male individuals extracted from the DSP2 reference sample. The three individuals with the highest anomaly scores are identified on the plot.

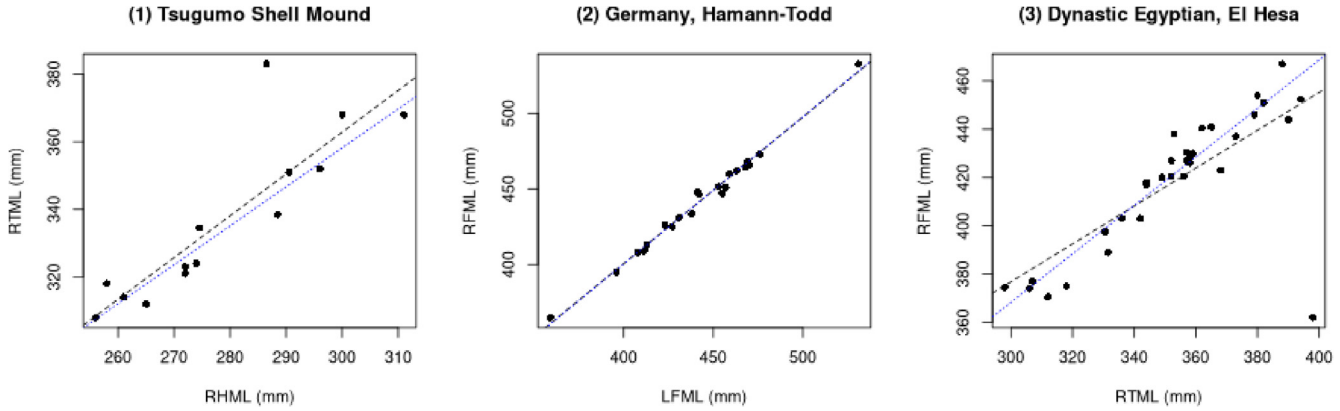

Fig. 8. Illustration of three types of outliers in linear regression, with three different population samples drawn the Goldman Data Set. Their corresponding shortcodes in this dataset are indicated as the main title; the shortcodes of the variables are indicated as axes labels. The black dashed lines are the regression lines including all the individuals; the blue dotted lines are the regression lines excluding the visual outliers.

two principal axes of the PCA were totally unhelpful in identifying those slight anomalies. This highlights a crucial fact: when the anomaly only concerns one given pair of variables among ten possible measurements, the impact may be sufficiently moderate so that multivariate methods cannot consider the individual as *globally* suspect. The DDC algorithm allows to detect the individuals having a slightly different morphology, even if it is restricted to a very precise region of the bone under study.

5. Bivariate outliers

This last section focuses on the particular case of bivariate data. Although general methods for multivariate outliers (especially the Mahalanobis distance, detailed in Section 3.1) can also be used when considering only two variables, some tools were specifically developed for this situation.

5.1. Outliers in the context of correlation and linear regression

When considering the relationship between two continuous variables, three main types of outliers can be defined. In the first panel of Fig. 8, one single individual is far from the regression line, but its position—near the average of the explanatory variable RHML—gives it only a limited influence in the regression model. In the middle panel, two extreme individuals can be identified on the margins of the horizontal axis. However, those two individuals perfectly respect the relationship observed on the other individuals, and the regression lines with or without those two extreme points are indistinguishable. Finally, the right panel shows an *influential point*, i.e. an individual which is both located on the margin of the explanatory variable and has a high residual value: this type of individual may have a strong impact in a regression model, especially when dealing with small sample sizes.

In a regression model, the influential individuals of the type seen in Fig. 8 (3) are the most problematic. Influential individuals can be identified through their high value of Cook's distance, which is provided as a standard diagnostic in most statistical software. A reasonable rule-of-thumb—that should be avoided in the case of a very small sample size—is that influential points have a Cook's distance greater than 1 (Cornillon and Matzner-Løber, 2010).

However, it should be noted that robust methods for correlation and regression do exist (Rousseeuw and Leroy, 1987). Manually excluding outliers is not mandatory with those modern techniques, that have their own built-in way to handle outliers.

A robust version of the correlation coefficient automatically restricts the computation to the "most central" part of the data, using the same minimum covariance determinant algorithm as the robust Mahalanobis distance detailed in Section 3.1 (Fig. 9). In particular, potential outliers

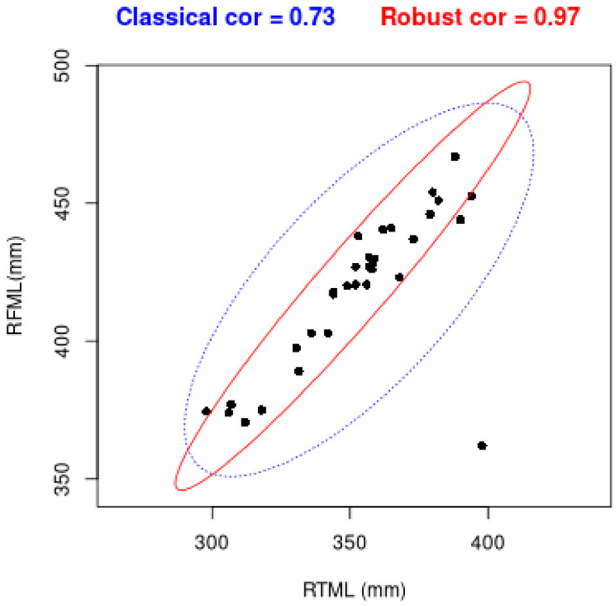

Fig. 9. Classical and robust estimates of the correlation coefficient between the maximal lengths of the right humerus and femur within the population sample "Dynastic Egyptian, El Hesa" drawn from the Goldman Data Set. Correlation ellipsoids are given an α level of 0.95, and a proportion $h = 3/4$ of individuals is used for MCD estimation.

can be kept on the plots, thus allowing to discuss some particular cases without introducing any bias in the computation.

Robust alternatives for linear regression are also implemented in various R packages. The function `MASS::rlm()` implements an algorithm that gives different weights to the individuals according to their distance to the regression line, and iteratively re-fits the model until convergence (Venables and Ripley, 2010). Another option is the quantile regression (Koenker, 2005), implemented in the function `quantreg::rq()`, that replaces the mean by the median within the framework of least squares estimation. As shown on Fig. 10, those two methods are usually consistent with each other, and in this case, with an ordinary linear regression performed after excluding the potential outliers.

5.2. General case: the bagplot

Depending on the aim and context of the study, the two extreme points on the middle panel of Fig. 8 can be seen as clear outliers (they are exceedingly tall and short compared to the other individuals from

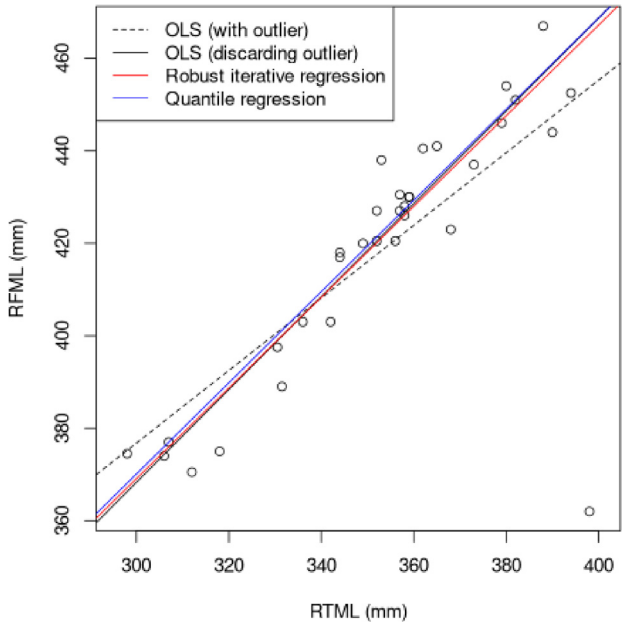

Fig. 10. Comparison of four strategies of linear regression between the right maximum femur and tibia lengths, using the population sample "Dynastic Egyptian, El Hesa" from the Goldman Data Set. Two OLS (ordinary least squares, i.e classical) linear regressions are performed, including or not the clear outlier. Two variants of robust regression are performed with the whole sample, including the outlier.

this population sample) or not (they do respect the relationship between the two measurements). In other words, they are clearly outliers as regards their measurements, but are not outliers in the framework of a regression model.

When one only searches for outliers in a two-dimensional distribution—outside of the context of linear regression or correlation—the bagplot (Rousseeuw et al., 1999) is the appropriate tool. The bagplot is a bivariate generalization of the boxplot. An inner polygon (*bag*) contains about 50% of the individuals which are the closest to the bivariate sample median; an exterior *fence* allows to identify the outliers and is defined by inflating the bag by a factor 3; and an intermediate region (the *loop*) is the convex hull of the outermost individuals that are not outliers. Rarely used in archaeological sciences—O'Connell et al. (2012) and Emery et al. (2018) are two of the few recent instances—, the bagplot provides a simple and visual way to identify bivariate outliers by an *ad hoc* rule (Fig. 11).

6. Discussion and conclusion

As stated by Leys et al. (2019, p. 5), "*there are no universal rules to tell you when to consider a value as "too far" from the others; researchers need to make this decision for themselves*". This statement is in line with the recommendations from Tukey (1977): outliers are data points flagged as somewhat unusual, and then constitute *candidates* for being true—informative—anomalies. Detecting outliers should always lead to thinking, sometimes action, but the final interpretation and conclusion is up to the researcher.

Therefore, any method of outlier detection comes with several arbitrary choices. The constant k in Eqs. (1)–(3) strongly impacts the severity of the decision rule by narrowing or widening the "credibility intervals"; a similar role is played by the α level in Eqs. (4) and (B.1) for Mahalanobis distances. By choosing lower or higher values for such parameters, either only the clearest extreme values or even slightly unusual values will be regarded as outliers. It is not possible to give a universal recommendation to set those parameters at a given value, and the researcher should be prepared to defend the strategy of outlier detection adopted in a study.

Furthermore, it is rather unlikely that an archaeologist can know beforehand the distribution of the variable(s) considered in the underlying population. The gaussian distribution, or at least a symmetric distribution, can be a reasonable assumption in the majority of situations encountered in archaeological sciences. However, one can almost never know with certainty which distribution a given set of values comes from, and this may be a good reason to use modern methods that makes few or even no assumption on the distribution of the data, such as isolation forests.

For all those reasons, outlier detection is strongly user-dependent, and the strategy adopted should be explicitly stated: in some ambiguous situations (cf. Fig. 2), the assumptions made by the researcher may strongly affect the results of outlier detection. Therefore, one should not rely on vague and non-specific assertions such as "after removing four outliers, we performed linear regression […]" without additional details.

Applying several robust methods of outliers detection and comparing their results may also appear as a good practice. In rather simple cases (normally distributed data with sufficient sample size and moderate number of variables), they should lead to the same conclusions (as in Figs. 4 and 5). When dealing with more complex patterns (e.g. involving nonlinear relationships, multimodal or asymmetric distributions), some discordance may appear, calling for an even more careful inspection of the data and of the potential candidates. The different methods of outliers detection all search for different types of outliers, and finding ways to compare them is an active topic in statistical research (e.g., Unwin, 2019). In the multivariate case, robust Mahalanobis distances and isolation forests may be seen as complementary, and can be used in combination, since they have truly different approaches. Indeed, the first method searches for unusual observations in a parametric model assuming roughly multivariate normal data (so that it delivers a "yes/no" answer at a given decision threshold). Conversely, isolation forests rank all individuals in terms of "outlyingness", without making any assumption about the distribution, and does not provide any definitive answer about any individual: it is up to the researcher to inspect carefully the individuals "flagged" by the algorithm, and to make a decision using his or her subjective knowledge.

The recent DDC algorithm may be very helpful in this latest step, by providing a complete map of deviating cells. Those entries may be either strong univariate anomalies or slightly odd combinations of variables. This method is maximally useful when dealing with high-dimensional datasets, both because of its internal logic—that takes advantage of the intercorrelation of the variables—and because it may become hard to understand why an individual is detected by Mahalanobis distance or isolation forests when the number of variables does not allow simple graphical representations anymore. In such a case, the DDC algorithm considerably helps the researcher to identify why some individuals may be regarded as outliers thanks to a very clear and synthetic graphical output (Fig. 7). It should also be noted that this algorithm is improved at a considerable pace, and several of its extensions (Raymaekers and Rousseeuw, 2019; Hubert et al., 2019) should be extremely valuable in osteology, since they allow both outlier detection and imputation of missing values.

Finally, it should be noted that categorical variables might also be considered when performing outlier detection, either by using algorithms which natively handle them (such as "hdoutliers"), or by turning them manually into multivariate numeric values via correspondence analysis (Unwin, 2019).

The focus of the present article was on outlier detection, and not outlier management in a broad sense. The problem of knowing what to do with the individuals that are detected as outliers is extensively covered in Leys et al. (2019). However, numerous robust methods have built-in way to handle outliers, and do not need a controversial manual exclusion. This article focused on robust correlation and regression methods, but most popular methods do have a robust equivalent which

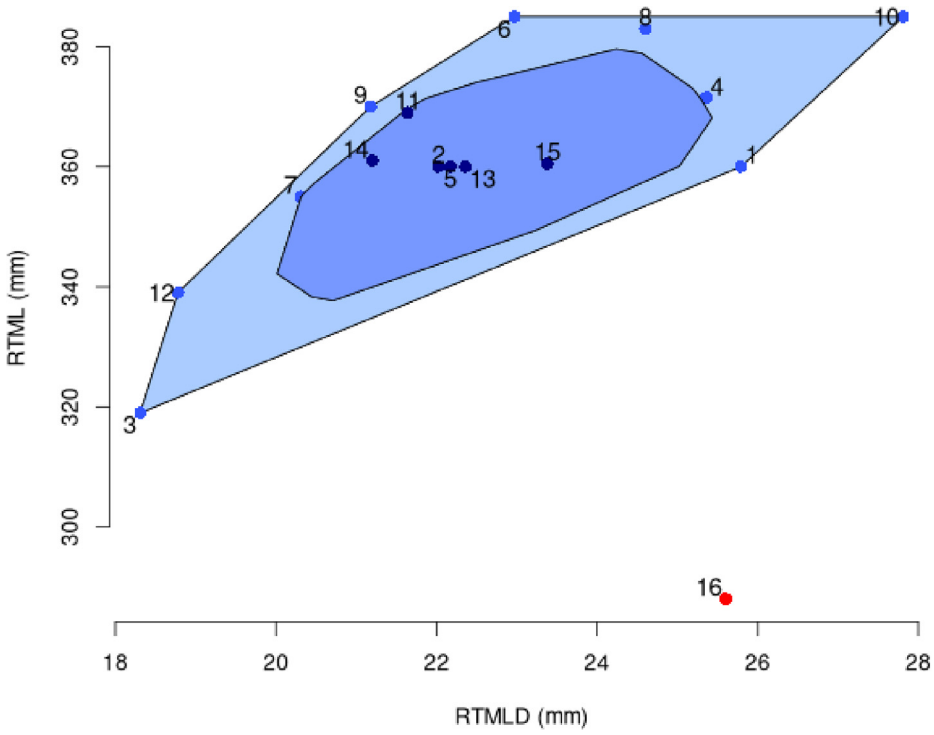

Fig. 11. Bagplot for the maximal length and medio-lateral diameter of the right tibia, measured on the population sample of Delaware (US-NJ, 500 BP) from the Goldman Data Set.

offers a valuable alternative for "contaminated data". Among other examples, robust principal component analysis (Candès et al., 2011) or robust estimation and hypothesis testing (Wilcox, 2012) can be cited. Within the field of robust estimation, winsorization—i.e., replacing all the values exceeding a given threshold t by the value t itself—or trimming—i.e., removing a given percentage of the most extreme values in both directions—could be valuable tools in archaeology, and would offer some new ways to deal with outlying values in statistical inference.

Data availability statement

No new data were created in this study. However, all the datasets used within the text are freely available online, and are the property of their respective authors.

The Goldman Osteometric Data Set is available at https://web.utk.edu/auerbach/GOLD.htm, and those data have been collected by Benjamin Auerbach. This dataset is also included in the R package `bioanth` (Eanes, 2015), and this is the source used in this study.

The DSP2 reference sample has been collected by Jaroslav Bruzek and is available in Bružek et al. (2017) as Supporting Information online. This dataset is also included in the R package `anthrostat`, and this is the source used in this study.

Acknowledgments

I would like to thank Jaroslav Bruzek (University of Bordeaux, France) for allowing me to use part of the DSP2 reference sample in this study.

My warm thanks to Sabrina Granger (Urfist Bordeaux, France), who strongly contributed to put me on the path of reproducible research. The welcoming community of Emacs and Org-mode users helped me to solve some problems encountered while writing this manuscript. Arnaud Legrand (University of Grenoble 1, France) also gave me useful advice about Org-mode.

Finally, the two anonymous reviewers must be acknowledged for providing invaluable and very detailed comments to improve the manuscript, its general structure, and its ability to be fully reproduced. I learned very much from their suggestions. Readers can access the first version of the manuscript on GitLab (https://gitlab.com/f-santos/reproducibility-package-for-santos-2020-jasr) and, by comparing it to the present text, appreciate the significant improvements made thanks to the reviewers' comments.

Appendix A. Formulae of robust scale estimates for univariate outliers detection

Full mathematical details are given here for three possible robust scale estimates $\hat{s}$ which can be used as input in Eq. (1) for univariate outliers detection.

A.1. The interquartile range

The interquartile range (IQR) is defined by the difference between the third and first quartiles of the data. It can be shown that, for a gaussian distribution, $\hat{s} = IQR/a$, with a scale factor $a \approx 1.349$, is a consistent estimate of σ (Wan et al., 2014). Therefore, in this first alternative, the outliers are those extreme values falling outside of the range $[m - k \cdot \frac{IQR}{1.349}; m + k \cdot \frac{IQR}{1.349}]$.

A.2. The median absolute deviation

The median absolute deviation (MAD) provides another estimate of σ which is even more robust than the IQR (Rousseeuw and Croux, 1993). For a given sample x, the MAD is defined as the scaled median of absolute deviations from the sample median:

$$MAD = b \times \text{med}(|x_i - \text{med}(x)|_{1 \leqslant i \leqslant n}) \tag{A.1}$$

The scale factor b depends on the underlying distribution of the data. If the normality assumption is reasonable (disregarding some potential extreme values), b should be set to 1.4826, which is approximately the inverse of the third theoretical quartile of the distribution $\mathcal{N}(0, 1)$. With this method, the outliers are defined as those values that fall outside of the range $[m - k{\cdot}MAD; m + k{\cdot}MAD]$.

A.3. The S_n estimator

A third alternative is the S_n estimator (Rousseeuw and Croux, 1993). S_n is defined by:

$$S_n = c{\cdot}\text{med}_i\{\text{med}_j|x_i - x_j|\} \tag{A.2}$$

and is a very robust estimate of the σ parameter of a gaussian distribution if the scale factor c is set to 1.1926. As for the two previous methods, the outliers are defined as those values that fall outside of the range $[m - k{\cdot}S_n; m + k{\cdot}S_n]$

Appendix B. Theoretical details for robust Mahalanobis distance

This method relies on the concept of generalized variance (Wilks, 1960; Gupta, 2006), which is a measure of multivariate dispersion defined by the determinant of the covariance matrix, $|\Sigma|$. The robust Mahalanobis distance proceeds by iteratively drawing at random h out of the n individuals (with $h \in [n/2, n[$), and finally selecting the subsample of size h that has the minimum generalized variance. Therefore, this can be seen as using only the "good part" of the data—i.e. a "central" part which does not include the potential outliers—to derive robust location and variability estimates. This best subsample of size h is finally used to compute the sample estimates $\hat{\mu}_{\text{MCD}}$ and $\hat{\Sigma}_{\text{MCD}}$ that define the robust Mahalanobis distance:

$$R_i = \sqrt{{}^t(x_i - \hat{\mu}_{\text{MCD}})\hat{\Sigma}^{-1}_{\text{MCD}}(x_i - \hat{\mu}_{\text{MCD}})} \tag{B.1}$$

The choice the parameter h (i.e. the proportion of "good data" used to compute the robust estimators) may have a substantial impact when dealing with small samples. As a general advice, h should be chosen with respect to the anticipated proportion of outliers in the study: if the researcher expects at least one fifth of outliers in his or her sample, h should be less than $4n/5$ to avoid that contaminated data participate to the calculations. A study by Leys et al. (2018) showed that choosing $h = 3n/4$ should be convenient in most situations, and offers a good compromise between robustness and accuracy. This is the value used in the present article.

Appendix C. R packages used in this study

As well as R 3.6.3 itself, the following R packages were used for writing this manuscript:

- `anthrostat` 0.1.5 (Santos, 2020)
- `aplpack` 190512 (Wolf, 2019)
- `bioanth` 0.1.0 (Eanes, 2015)
- `cellWise` 2.1.1 (Raymaekers et al., 2020)
- `FactoMineR` 2.3 (L e et al., 2008)
- `MASS` 7.3–51.5 (Venables and Ripley, 2010)
- `mvoutlier` 2.0.9 (Filzmoser and Gschwandtner, 2018)
- `quantreg` 5.55 (Koenker, 2020)
- `robustbase` 0.93.6 (Todorov and Filzmoser, 2009)
- `scatterplot3d` 0.3–41 (Ligges and Mächler, 2003)
- `solitude` 0.2.1 (Srikanth, 2019)
- `univOutl` 0.1–5 (D'Orazio, 2019)

This exact computational environment is made publicly available through a Docker image that also includes Emacs 26.3, Org-mode 9.3.6, various other Emacs packages, and a LATEX distribution. This ensures that the manuscript can be reproduced in its exact form on any computer, using the source Org file.

Full details are available on the GitLab repository (https://gitlab.com/f-santos/reproducibility-package-for-santos-2020-jasr).

References

Algee-Hewitt, B.F.B., 2016. Population inference from contemporary American craniometrics. Am. J. Phys. Anthropol. 160, 604–624. https://doi.org/10.1002/ajpa.22959.

Auerbach, B.M., Raxter, M.H., 2008. Patterns of clavicular bilateral asymmetry in relation to the humerus: variation among humans. J. Hum. Evol. 54, 663–674. https://doi.org/10.1016/j.jhevol.2007.10.002.

Auerbach, B.M., Ruff, C.B., 2004. Human body mass estimation: a comparison of morphometric and mechanical methods. Am. J. Phys. Anthropol. 125, 331–342. https://doi.org/10.1002/ajpa.20032.

Beck, J., Smith, B.H., 2019. Don't throw the baby teeth out with the bathwater: Estimating subadult age using tooth wear in commingled archaeological assemblages. Int. J. Osteoarchaeol. 29, 831–842. https://doi.org/10.1002/oa.2802.

Bergstrom, M.L., Hogan, J.D., Melin, A.D., Fedigan, L.M., 2019. The nutritional importance of invertebrates to female Cebus capucinus imitator in a highly seasonal tropical dry forest. Am. J. Phys. Anthropol. 170, 207–216. https://doi.org/10.1002/ajpa.23913.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bružek, J., Santos, F., Dutailly, B., Murail, P., Cunha, E., 2017. Validation and reliability of the sex estimation of the human os coxae using freely available DSP2 software for bioarchaeology and forensic anthropology: BRUŽEK et al. Am. J. Phys. Anthropol. 164:440–449. doi:https://doi.org/10.1002/ajpa.23282.

Brys, G., Hubert, M., Struyf, A., 2004. A robust measure of skewness. J. Computat. Graph. Stat. 13, 996–1017. URL: https://www.jstor.org/stable/27594089.
Candès, E.J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? J. ACM 58, 11:1–11:37. https://doi.org/10.1145/1970392.1970395.
Cornillon, P.-A., Matzner-Løber, E., 2010. Régression avec R. Pratique R. Paris: Springer. OCLC: 845859225.
Desquilbet, L., Granger, S., Hejblum, B., Legrand, A., Pernot, P., Rougier, N., 2019. Vers Une Recherche Reproductible. Unité régionale de formation à l'information scientifique et technique de Bordeaux, Bordeaux. URL: https://hal.archives-ouvertes.fr/hal-02144142.
Dietmeier, J.K.C., 2018. The oxen of Oxon Hill Manor: pathological analyses and cattle husbandry in eighteenth-century Maryland. Int. J. Osteoarchaeol. 28, 419–427. https://doi.org/10.1002/oa.2667.
D'Orazio, M., 2017. OutlierDetection in R: Some Remarks. In 5th International Conference "New Challenges for Statistical Software – The Use of R in Official Statistics". Bucharest, Romania.http://www.r-project.ro/conference2017/presentations/D'Orazio-Outlier_Detection_in_R_(slides_v5).pdf.
D'Orazio, M., 2019. univOutl: Detection of Univariate Outliers. R package version 0.1-5. https://CRAN.R-project.org/package=univOutl.
Eanes, G., 2015. Bioanth: Datasets useful in Biological Anthropology. R package version 0.1.0.https://github.com/geanes/bioanth.
Emery, M.V., Stark, R.J., Murchie, T.J., Elford, S., Schwarcz, H.P., Prowse, T.L., 2018. Mapping the origins of Imperial Roman workers (1st–4th century CE) at Vagnari, Southern Italy, using 87Sr/86Sr and δ18O variability. Am. J. Phys. Anthropol. 166, 837–850. https://doi.org/10.1002/ajpa.23473.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, pp. 226–231.
Filzmoser, P., Gschwandtner, M., 2018. Mvoutlier: Multivariate Outlier Detection Based on Robust Methods.https://CRAN.R-project.org/package=mvoutlier.
Graham, J.H., Özener, B., 2016. Fluctuating asymmetry of human populations: a review. Symmetry 8, 154. https://doi.org/10.3390/sym8120154.
Hakenbeck, S., McManus, E., Geisler, H., Grupe, G., O'Connell, T., 2010. Diet and mobility in Early Medieval Bavaria: a study of carbon and nitrogen stable isotopes. Am. J. Phys. Anthropol. 143, 235–249. https://doi.org/10.1002/ajpa.21309.
Harbottle, G., 1976. Activation analysis in archaeology. In G. W. A. Newton (Ed.), Radiochemistryz. Cambridge: Royal Society of Chemistry, vol. 3, pp. 33–72.https://doi.org/10.1039/9781847556882-00033.
Harris, E.F., Bailit, H.L., 1988. A principal components analysis of human odontometrics. Am. J. Phys. Anthropol. 75, 87–99. https://doi.org/10.1002/ajpa.1330750110.
Hawkins, D.M., 1980. Identification of Outliers. Springer, Netherlands, Dordrecht. URL: http://public.eblib.com/choice/publicfullrecord.aspx?p=3106349.
Hubert, M., Debruyne, M., Rousseeuw, P.J., 2018. Minimum covariance determinant and extensions. Wiley Interdisc. Rev.: Comput. Stat. 10https://doi.org/10.1002/wics.1421. e1421.
Hubert, M., Rousseeuw, P.J., den Bossche, W.V., 2019. MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. Technometrics 61, 459–473. https://doi.org/10.1080/00401706.2018.1562989.
Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Comput. Stat. Data Anal. 52, 5186–5201. https://doi.org/10.1016/j.csda.2007.11.008.
Jolliffe, I.T., 2002. Principal Component Analysis. Springer Series in Statistics, second ed. Springer-Verlag, New York. https://doi.org/10.1007/b98835.
Kimber, A.C., 1990. Exploratory data analysis for possibly censored data from skewed distributions. J. R. Stat. Soc. Ser. C (Appl. Stat.) 39, 21–30. https://doi.org/10.2307/2347808.
Koenker, R., 2005. Quantile regression by Roger Koenker. Cambridge University Presshttps://doi.org/10.1017/CBO9780511754098.
Koenker, R., 2020. Quantreg: Quantile Regression. R package version 5.55.https://CRAN.R-project.org/package=quantreg.
Lê, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw. 25. https://doi.org/10.18637/jss.v025.i01.
Leys, C., Delacre, M., Mora, Y.L., Lakens, D., Ley, C., 2019. How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. Int. Rev. Soc. Psychol. 32, 5. https://doi.org/10.5334/irsp.289.
Leys, C., Klein, O., Dominicy, Y., Ley, C., 2018. Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. J. Exp. Soc. Psychol. 74, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011.
Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 49, 764–766. https://doi.org/10.1016/j.jesp.2013.03.013.
Ligges, U., Mächler, M., 2003. Scatterplot3d: an R package for visualizing multivariate data. J. Stat. Softw. 8. https://doi.org/10.18637/jss.v008.i11.
Lightfoot, E., O'Connell, T.C., 2016. On the use of biomineral oxygen isotope data to identify human migrants in the archaeological record: intra-sample variation, statistical methods and geographical considerations. PLOS ONE 11https://doi.org/10.1371/journal.pone.0153850. e0153850.
Lightfoot, E., Šlaus, M., O'Connell, T.C., 2014. Water consumption in Iron Age, Roman, and Early Medieval Croatia. Am. J. Phys. Anthropol. 154, 535–543. https://doi.org/10.1002/ajpa.22544.
Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discovery Data 6, 1–39. https://doi.org/10.1145/2133360.2133363.
Loftus, E., Sealy, J., 2012. Technical note: Interpreting stable carbon isotopes in human tooth enamel: an examination of tissue spacings from South Africa. Am. J. Phys. Anthropol. 147, 499–507. https://doi.org/10.1002/ajpa.22012.
Lubritto, C., García-Collado, M.I., Ricci, P., Altieri, S., Sirignano, C., Castillo, J.A.Q., 2017. New dietary evidence on medieval rural communities of the Basque Country (Spain) and its surroundings from carbon and nitrogen stable isotope analyses: social insights, diachronic changes and geographic comparison. Int. J. Osteoarchaeol. 27, 984–1002. https://doi.org/10.1002/oa.2610.
Mahoney, P., 2006. Dental microwear from Natufian hunter-gatherers and early Neolithic farmers: comparisons within and between samples. Am. J. Phys. Anthropol. 130, 308–319. https://doi.org/10.1002/ajpa.20311.
Marwick, B., 2017. Computational reproducibility in archaeological research: basic principles and a case study of their implementation. J. Archaeol. Method Theory 24, 424–450. https://doi.org/10.1007/s10816-015-9272-9.
O'Connell, T.C., Kneale, C.J., Tasevska, N., Kuhnle, G.G.C., 2012. The diet-body offset in human nitrogen isotopic values: a controlled dietary study. Am. J. Phys. Anthropol. 149, 426–434. https://doi.org/10.1002/ajpa.22140.
Pilloud, M.A., Hefner, J.T. (Eds.), 2016. Biological Distance Analysis: Forensic and Bioarchaeological Perspectives. London, United Kingdom; San Diego, CA, USA: Academic Press. OCLC: ocn951764374.
R Core Team, 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
Raymaekers, J., Rousseeuw, P.J., 2019. Flagging and handling cellwise outliers by robust estimation of a covariance matrix. arXiv:1912.12446 [stat]. arXiv:1912.12446. http://arxiv.org/abs/1912.12446.
Raymaekers, J., Rousseeuw, P.J., Van den Bossche, W., & Hubert, M., 2020. cellWise: Analyzing Data with Cellwise Outliers. R package version 2.1.1.https://CRAN.R-project.org/package=cellWise.
Rousseeuw, P.J., Bossche, W.V.D., 2018. Detecting deviating data cells. Technometrics 60, 135–145. https://doi.org/10.1080/00401706.2017.1340909.
Rousseeuw, P.J., Croux, C., 1993. Alternatives to the median absolute deviation. J. Am. Stat. Assoc. 88, 1273–1283. https://doi.org/10.1080/01621459.1993.10476408.
Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.
Rousseeuw, P.J., Ruts, I., Tukey, J.W., 1999. The Bagplot: a bivariate boxplot. Am. Stat. 53, 382–387. https://doi.org/10.1080/00031305.1999.10474494.
Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223. https://doi.org/10.1080/00401706.1999.10485670.
Rousselet, G.A., Wilcox, R.R., 2019. Reaction times and other skewed distributions: Problems with the mean and the median. bioRxiv, (p. 383935).https://doi.org/10.1101/383935.
Santana-Sagredo, F., Lee-Thorp, J.A., Schulting, R., Uribe, M., 2015. Isotopic evidence for divergent diets and mobility patterns in the Atacama Desert, northern Chile, during the Late Intermediate Period (AD 900–1450). Am. J. Phys. Anthropol. 156, 374–387. https://doi.org/10.1002/ajpa.22663.
Santos, F., 2020. Anthrostat: A Set of Useful Functions for Biological Anthropology and Past Sciences.https://gitlab.com/f-santos/anthrostat/.
Schulte, E., Davison, D., Dye, T., Dominik, C., 2012. A multi-language computing environment for literate programming and reproducible research. J. Stat. Softw. 46, 1–24. https://doi.org/10.18637/jss.v046.i03.
Sen Gupta, A., 2006. Generalized Variance. In: In: Kotz, S., Read, C.B., Balakrishnan, N., Vidakovic, B., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences (p. ess6053.pub2) p John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/0471667196.ess6053.pub2.
Srikanth, K., 2019. Solitude: An Implementation of Isolation Forest. R package version 0.2.1.https://CRAN.R-project.org/package=solitude.
Stynder, D.D., 2009. Craniometric evidence for South African Later Stone Age herders and hunter–gatherers being a single biological population. J. Archaeol. Sci. 36, 798–806. https://doi.org/10.1016/j.jas.2008.11.001.
Todorov, V., Filzmoser, P., 2009. An object-oriented framework for robust multivariate analysis. J. Stat. Softw. 32, 1–47. https://doi.org/10.18637/jss.v032.i03.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley Series in Behavioral Science. Addison-Wesley Pub. Co, Reading, Mass.
Unwin, A., 2019. Multivariate Outliers and the O3 Plot. J. Computat. Graph. Stat. 28, 635–643. https://doi.org/10.1080/10618600.2019.1575226.
Venables, W.N., Ripley, B.D., 2010. Modern Applied Statistics with S. Statistics and Computing (4th ed.). New York: Springer. OCLC: 837651785.
Wan, X., Wang, W., Liu, J., Tong, T., 2014. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med. Res. Methodol. 14. https://doi.org/10.1186/1471-2288-14-135.
Warren, R., Smith, R., Cybenko, A., 2011. Use of Mahalanobis Distance for Detecting Outliers and Outlier Clusters in Markedly Non-Normal Data: A Vehicular Traffic Example. Technical Report AFRL-RH-WP-TR-2011-0070 Air Force Research Laboratory.https://apps.dtic.mil/dtic/tr/fulltext/u2/a545834.pdf.
Webb, E.C., White, C.D., Longstaffe, F.J., 2013. Exploring geographic origins at cahuachi using stable isotopic analysis of archaeological human tissues and modern environmental waters. Int. J. Osteoarchaeol. 23, 698–715. https://doi.org/10.1002/oa.1298.
Weiss, E., 2009. Sex differences in humeral bilateral asymmetry in two hunter-gatherer populations: California Amerinds and British Columbian Amerinds. Am. J. Phys. Anthropol. 140, 19–24. https://doi.org/10.1002/ajpa.21025.
Wilcox, R.R., 2012. Introduction to Robust Estimation and Hypothesis Testing. Statistical Modeling and Decision Science, third ed. Academic Press, Amsterdam, Boston.
Wilkinson, L., 2018. Visualizing big data outliers through distributed aggregation. IEEE Trans. Visual Comput. Graphics 24, 256–266. https://doi.org/10.1109/TVCG.2017.2744685.
Wilks, S., 1960. Multidimensional Statistical Scatter. In Contributions to Probability and Statistics (pp. 486–503). Stanford, US-CA: I. Olkin et al. (Stanford University Press ed.).
Wolf, H.P., 2019. Aplpack: Another Plot Package (version 190512).https://cran.r-project.org/package=aplpack.
Wright, L.E., 2005. Identifying immigrants to Tikal, Guatemala: Defining local variability in strontium isotope ratios of human tooth enamel. J. Archaeol. Sci. 32, 555–566. https://doi.org/10.1016/j.jas.2004.11.011.