An overview of some robust methods for univariate and multivariate outliers detection with applications to archaeological samples

Frédéric Santos*

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

Abstract

Whereas outlier detection is routinely performed in archaeological sciences and may have a substantial impact of 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 on osteometric data extracted from the Goldman Data Set online. 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.

Keywords: isolation forests, MAD, robust Mahalanobis distance, robust statistics, R language

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. Some 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

*Corresponding author
Email address: frederic.santos@u-bordeaux.fr (Frédéric Santos)

too unusual values to be part of a given human group. The latter case is particularly frequent in isotopic studies where outliers for $\delta^{13}C$, $\delta^{15}N$ or $\delta^{18}O$ values might indicate the presence of non-local individuals, thus allowing to discuss migrations and mobility patterns among human groups (Hakenbeck et al., 2010; Kendall et al., 2013; Knudson & Tung, 2011; Santana-Sagredo et al., 2015). 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 past 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 oftenest 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., Webb et al., 2013; Wright, 2005). Third, although a few publications do utilize efficient and specific methods to detect multivariate outliers (e.g., Algee-Hewitt, 2016; Harris & Bailit, 1988; Mahoney, 2006), numerous articles rely on a combination of univariate methods applied separately on each variable, which is sub-optimal and can even lead to misleading results. Finally, both the modern statistical methods for detecting outliers and their implementation in free software seem to be little known.

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 & Sealy, 2012), or to build regression models (e.g., **?**). 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 examples where those robust methods outperform more classical and widely used methods will be given, using population samples extracted from the Goldman Data Set freely available online (Auerbach & Ruff, 2004).

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 & 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 september 2019—, a research within the database of the articles published in *Journal of Archaeological Science: Reports* triggered only three results for the keywords "median absolute deviation", two results for "bagplot", one result for "robust Mahalanobis", and no result could be found for requests about "multivariate outliers" or the S_n estimator. As concerns robust methods, the keywords "robust regression" or "quantile regression" triggered one single result—which furthermore only deals with local polynomial regression (LOESS). 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, illustrates some caveats of the most used methods, and provides ready-to-use R templates for modern methods of outlier detection. Several excellent R packages already provide useful functions implementing the robust methods described in this article, such as `Routliers` (Klein & Delacre, 2019) or `univOutl` (D'Orazio, 2018). Some additional functions, mainly bringing more options of data vizualisation, are implemented in an R package introduced in the present article, `anthrostat`, which is available on GitLab (`https://gitlab.com/f.santos/anthrostat`).

Finally, to reinforce the move towards a reproducible research in archaeological and anthropological sciences (Marwick, 2017a,b), this whole article has been written in Org-mode 9.3.1 for Emacs 26.3 (Schulte et al., 2012; Stanisic et al., 2015) and is fully reproducible with the org source file available on GitLab (`https://gitlab.com/f.santos/reproducibility-package-for-santos-2020-jasr`). For non Emacs users, the source code of all tables and figures from this study has also been made available in separate R files. All the statistical analyses were performed with R 3.6.2 (R Core Team, 2019).

2. Univariate outliers

This first section deals with outlying values for one sigle 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 & 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 allows to discuss 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.; 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 & O'Connell (2016). The data sets handled in past sciences do not always meet the previous requirements, oftenest 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.

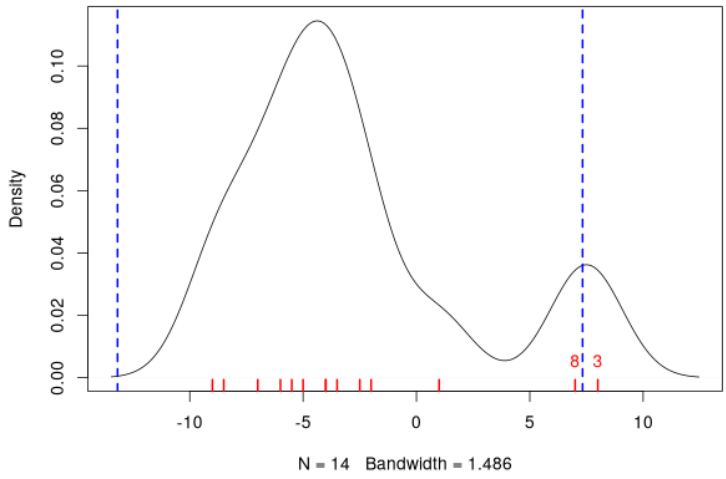

Figure 1: Kernel density estimation of the vector x of left-right differences in humeral length observed on the 14 male individuals from the population sample Ipituaq (US-AK, 1500–1100 BP) in the Goldman Data Set. The blue dotted vertical lines represent the exclusion thresholds defined by the classical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

Figure 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 more robust estimates for μ and σ than the classical sample mean and standard deviation. Consequently, the estimates calculated to define a "credible range of variation" outside of which any value can be considered as an outlier, are themselves fewer 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":

$$[m - k \cdot \hat{s}\,;\, m + k \cdot \hat{s}] \tag{1}$$

where m is the sample median—a robust estimate for the expectation μ—, and $\hat{s}$ is a robust estimate for σ (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. With very small sample sizes, $k = 3$ seems recommendable to avoid false positives (Leys et al., 2019).

2.2.1. 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 $[m - k \cdot \frac{IQR}{1.349}\,;\, m + k \cdot \frac{IQR}{1.349}]$.

2.2.2. The median absolute deviation

The median absolute deviation (MAD) provides another estimate of σ which is even more robust than the IQR (Rousseeuw & Croux, 1993). Although defined more than 40 years ago by Hampel (1974), this estimate is still rarely used in archaeological sciences. For a given sample x, the MAD is defined as the scaled median of absolute deviations from the sample median:

$$MAD = b \times \text{med}(|x_i - \text{med}(x)|_{1 \leqslant i \leqslant n}) \tag{2}$$

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 opposite 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 $[m - k \cdot MAD\,;\, m + k \cdot MAD]$

2.2.3. The S_n estimator

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

$$S_n = c \cdot \text{med}_i \left\{ \text{med}_j |x_i - x_j| \right\} \tag{3}$$

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 $[m - k \cdot S_n \,;\, m + k \cdot S_n]$

2.2.4. Application to the Goldman Data Set

To compare the three robust methods described above with 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 on Table 1.

	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

Table 1: Comparison of four methods based on location and scale parameters for outlier detection, applied on the data described in Figure 1. "Coef" is the user-defined constant k used for the construction of intervals, see equation (1). The lower and upper bounds of the intervals built with each method are indicated in the corresponding columns.

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 parameters—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 for non-gaussian data

In most contexts of past sciences, such as osteometric or isotopic studies, there is almost always a strong 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 & O'Connell (2016, p. 22), skewed data may simply indicate a sample with several outliers on the same distribution tail, as in Figure 1.

Severely skewed distributions arise almost systematically in some disciplinary fields such as neurosciences (Rousselet & 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 & 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 symmetrical, the previous rules based on the MAD and S_n estimator remain valid, albeit more difficult to use since their scale factors must be approximated through computer simulations (Rousseeuw & Croux, 1993).
2. If an asymmetric or skewed distribution is suspected, 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, boxplot-based rules are a simple yet efficient way to proceed.

2.3.1. The classical boxplot rule

Boxplots (Tukey, 1977) are often used to detect univariate outliers. Widely used in past sciences (e.g., Pickard et al., 2017), this rule makes no particular assumption about the underlying distribution. The standard boxplot rule does not use one single location estimate and a scale estimate as previous methods. Instead, the credible range of credible values (i.e., the boxplot *fences*) is defined by:

$$[q_1 - k \cdot IQR\,;\, q_3 + k \cdot IQR] \tag{4}$$

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 not symmetrical.

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:

$$[q_1 - 2k \cdot (m - q_1)\,;\, q_3 + 2k \cdot (q_3 - m)] \tag{5}$$

using previous notations, 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 Figure 2.

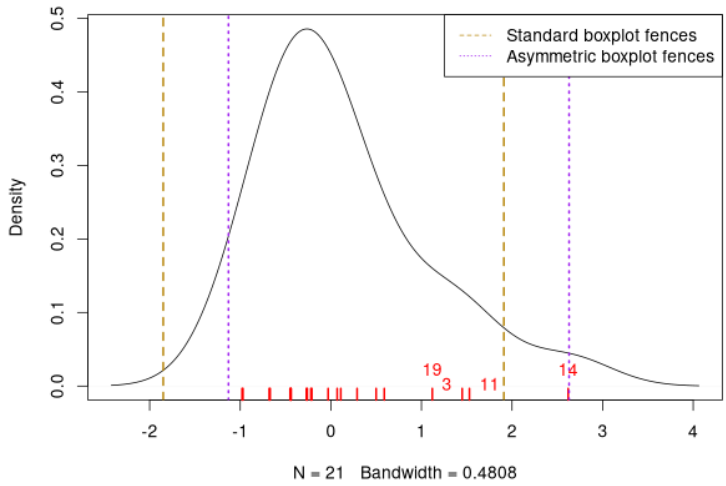

Figure 2: Kernel density estimation of the vector right-left differences in tibial mediolateral diameter observed on the 21 individuals from the population sample of Giza (Egypt, 4700–4200 BP) in the Goldman Data Set. The four most extreme individuals on the right tail are labeled in red.

Out of any context, this distribution might simply be regarded as right-skewed. Actually, 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 & 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 knwoledge. 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 on a combination of univariate methods (Leys et al., 2018). Among other available algorithms such as Dbscan (Ester et al., 1996), two methods are detailed below, which are both conceptually rather simple and practically easy-to-use, thanks to very 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 & Hefner, 2016), where the data suffers almost always from a great 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:

$$D_i = \sqrt{{}^t(x_i - \hat{\mu})\Sigma^{-1}(x_i - \hat{\mu})} \tag{6}$$

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

The Mahalanobis distance can be used to detect multivariate outliers (e.g., Stynder, 2009): 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.

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 univariate case, the estimates used in the formula (6) are non-robust and may be distorded by potential outliers, thus making invalid the whole decision rule.

A robust variant of Mahalanobis distance was proposed by Hubert et al. (2018). Their method rely on the concept of generalized variance (Oja, 1983; Sen Gupta, 2006; Wilks, 1960), 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. Intuitively, this can be seen as working only on a "good part" of the data, i.e. a "central" part which does not include the potential outliers. This best subsample of size h is finally used to compute the sample estimates $\hat{\mu}_{\text{MCD}}$, $\hat{\Sigma}_{\text{MCD}}$ that define the robust Mahalanobis distance:

$$R_i = \sqrt{{}^t(x_i - \hat{\mu}_{\text{MCD}})\,\hat{\Sigma}^{-1}_{\text{MCD}}\,(x_i - \hat{\mu}_{\text{MCD}})} \tag{7}$$

This procedure is also known as the MCD (minimum covariance determinant) algorithm. As in the case of the usual Mahalanobis distance, the outliers are defined as those individuals whose robust Mahalanobis distance R_i exceeds $\sqrt{\chi^2_{p;1-\alpha}}$. 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.

An implementation of robust Mahalanobis distance is available in the R package `robustbase` (Todorov & Filzmoser, 2009). This package will be used to illustrate the differences between the classical and robust versions of the Mahalanobis distance. Figure 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.

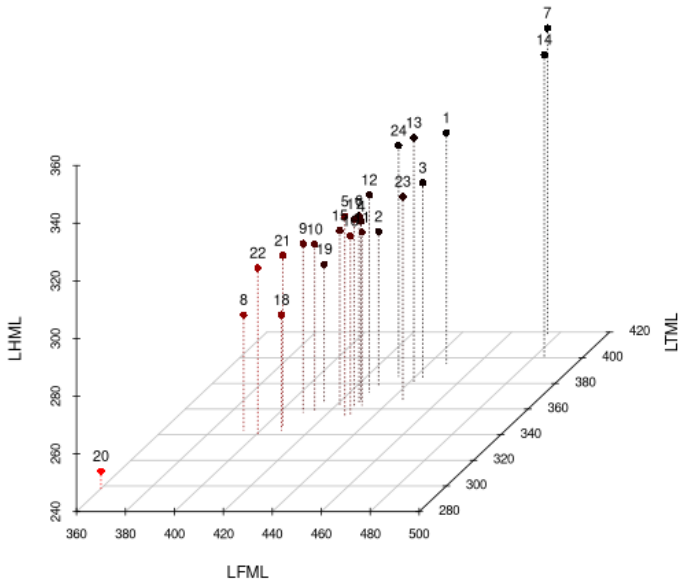

Figure 3: 3D scatterplot of the population sample of Sayala, drawn from the Goldman Data Set. The maximal lengths are three long bones are represented.

The presence of those outliers causes an inflation of the generalized variance, i.e. the determinant of the classical covariance matrix, $|\Sigma|$. 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.

3.2. ***TODO*** *Isolation forests*

Isolation forests are a very 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,

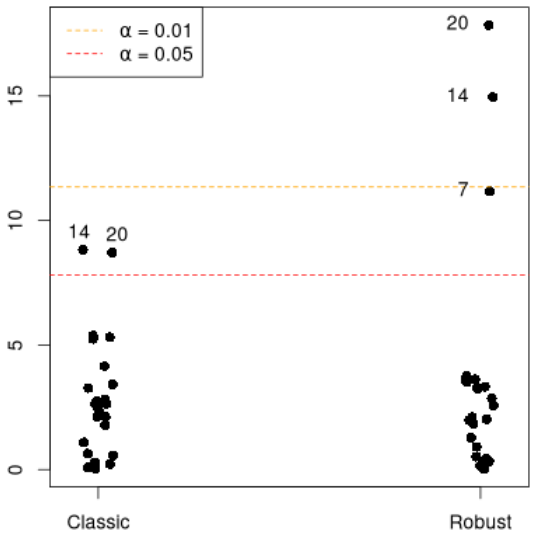

Figure 4: Stripcharts displaying the squared classical and robust Mahalanobis distances between each individual and the centroid. The dotted lines symbolize the exclusion thresholds $\chi^2_{p;1-\alpha}$ for two different α values. The maximal lengths of three long bones from the population sample of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

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 weak number, 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 Figure 5. The isolation forest algorithm provides a moderate 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.

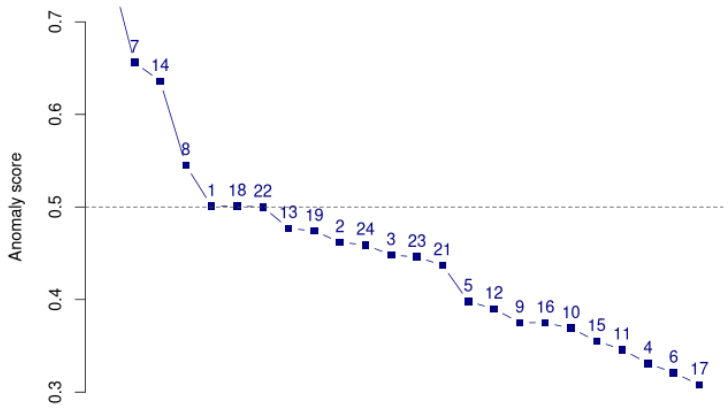

Figure 5: Plot of the anomaly scores obtained by an isolation forest to detect outliers from the population sample of Sayala (Goldman Data Set), when three maximal lengths are considered (LTML, LHML, LFML). The scores are sorted in decreasing order and the corresponding individual IDs are indicated.

4. Bivariate outliers

Although the general methods for multivariate outliers detailed in section 3 can also be used when considering only two variables, some tools were specifically developed for this situation.

4.1. 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 Figure 6, 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 withour those two extreme points are indistinguishable. Finally, the right panel shows a *leverage* individual, 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 has a great influence in a regression model, especially when dealing with small sample sizes.

In a regression model, only the leverage individuals corresponding to the right panel of Figure 6 are problematic. Leverage individuals can be identified through their high value of Cook's distance, which is

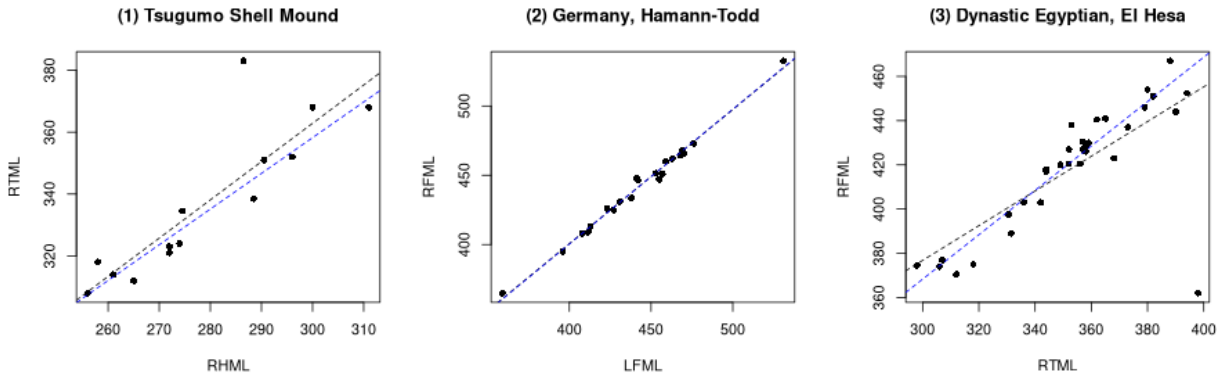

Figure 6: Illustration of three types of outliers in linear regression, with three different population samples drawn the Goldman Data Set. Their corresponding shortcodes in this dataset are indicated as the main title; the shortcodes of the variables are indicated as axes labels. The black dotted lines are the regression lines including all the individuals; the blue dotted lines are the regression lines excluding the visual outliers.

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 leverage points have a Cook's distance greater than 1 (Cornillon & Matzner-Løber, 2010).

However, it should be noted that robust methods for correlation and regression do exist (**?**). 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 MCD algorithm as the robust Mahalanobis distance detailed in section 3.1 (Fig. 7). In particular, potential outliers can be lefted in 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 R. The function `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 & Ripley, 2010). Another option is the quantile regression (Koenker, 2005), that replaces the mean by the median wihtin the framework of least squares estimation. As shown on Figure 8, those two methods are usually consistent with each other, and with an ordinary linear regression performed after excluding the potential outliers.

4.2. General case: the bagplot

Depending on the aim and context of the study, the two extreme points on the middle panel of Figure 6 can be seen as clear outliers (they are exceedingly tall and short compared to the other individuals from this population sample) or not (they do respect the relationship between the two measurements). In other

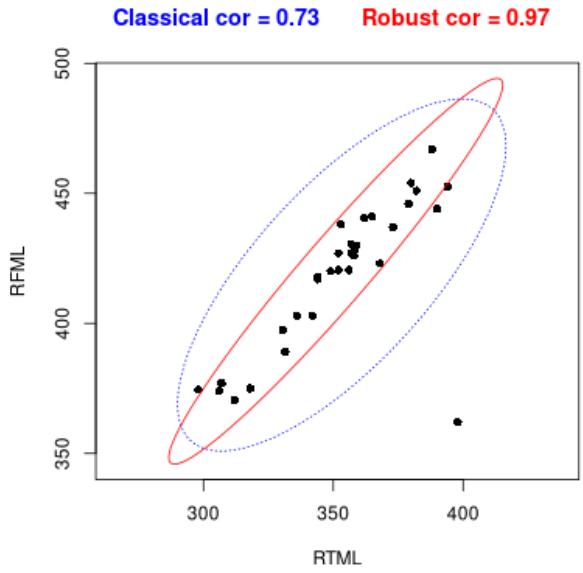

Figure 7: Classical and robust estimates of the correlation coefficient between the maximal lengths of the right humerus and femur within the population sample "Dynastic Egyptian, El Hesa" drawn from the Goldman Data Set. Correlation ellipsoids are given an α level of 0.95, and a proportion $h = 3/4$ of individuals is used for MCD estimation.

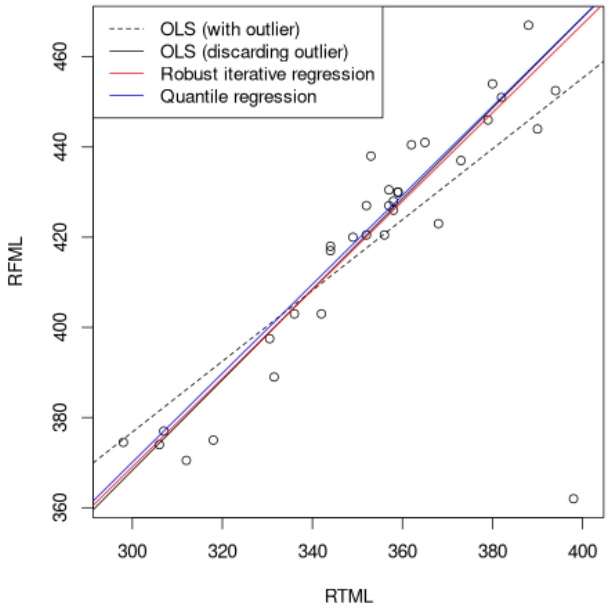

Figure 8: Comparison of four strategies of linear regression between the right maximum femur and tibia lengths, using the population sample "Dynastic Egyptian, El Hesa" from the Goldman Data Set.

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 bivariate outliers 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. 9).

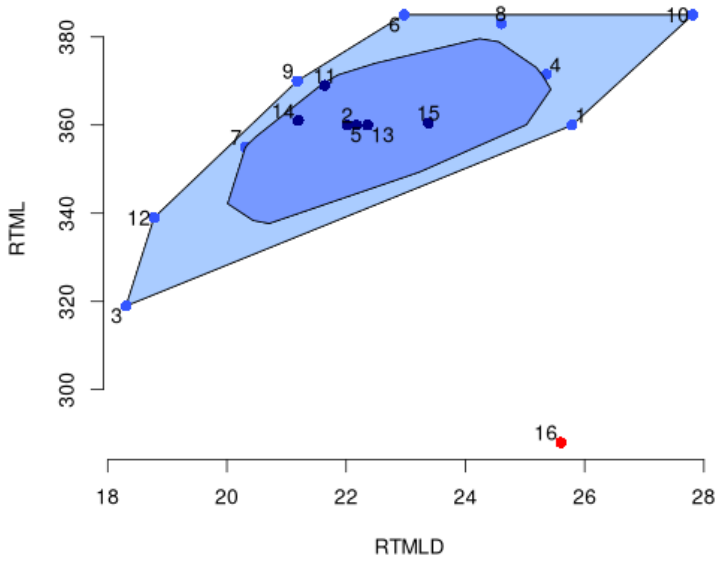

Figure 9: Bagplot for the the maximal length and medio-lateral diameter of the right tibia, measured on the population sample of Delaware (US-NJ, 500 BP) from the Goldman Data Set.

5. 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*". Any method of outlier detection comes from several arbitrary choices from the researcher. The constant k in equations (1) to (5) strongly impact the severity of the decision rule by narrowing or widening the "credibility intervals"; a similar role is played by the α level in equations (6) and (7) for Mahalanobis distances. By chosing lower or higher values for such parameters, either only the clearest extreme values or even slightly unusual values will be

regarded as outliers. There is no possibility 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 symmetrical distribution, can be a reasonable assumption in the large majority of situations encountered in past sciences. However, one can almost never know with certainty which distribution a given set of values comes from. In some ambiguous situations (cf. Fig. 2), the assumptions made by the researcher also greatly impact the results of outlier detection.

For all those reasons, outlier detection is strongly user-dependent, and the strategy adopted should be explicitly stated. One should not rely on vague and non-specific assertions such as "after removing four outliers, we performed linear regression [...]" without additional details.

The focus of the present article was on outlier detection, and not outlier management in a broad sense. The problem of kwowing 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 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.

TODO Acknowledgments

- Reviewers
- Granger
- Legrand

TODO Data availability statement

References

Algee-Hewitt, B. F. B. (2016). Population inference from contemporary American craniometrics. *American Journal of Physical Anthropology*, *160*, 604–624. doi:10.1002/ajpa.22959.

Auerbach, B. M., & Raxter, M. H. (2008). Patterns of clavicular bilateral asymmetry in relation to the humerus: Variation among humans. *Journal of Human Evolution*, *54*, 663–674. doi: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. *American Journal of Physical Anthropology*, *125*, 331–342. doi:10.1002/ajpa.20032.

Bergstrom, M. L., Hogan, J. D., Melin, A. D., & Fedigan, L. M. (). The nutritional importance of invertebrates to female Cebus capucinus imitator in a highly seasonal tropical dry forest. *American Journal of Physical Anthropology*, *0*. doi:10.1002/ajpa.23913.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. doi:10.1023/A:1010933404324.

Brys, G., Hubert, M., & Struyf, A. (2004). A Robust Measure of Skewness. *Journal of Computational and Graphical Statistics*, *13*, 996–1017.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust Principal Component Analysis? *J. ACM*, *58*, 11:1–11:37. doi:10.1145/1970392.1970395.

Cornillon, P.-A., & Matzner-Løber, E. (2010). *Régression avec R*. Pratique R. Paris: Springer. OCLC: 845859225.

Dietmeier, J. K. C. (2018). The oxen of Oxon Hill Manor: Pathological analyses and cattle husbandry in eighteenth-century Maryland. *International Journal of Osteoarchaeology*, *28*, 419–427. doi: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.

D'Orazio, M. (2018). univOutl: Detection of Univariate Outliers. R package version 0.1-4.

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. *American Journal of Physical Anthropology*, *166*, 837–850. doi: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. (pp. 226–231). AAAI Press.

Graham, J. H., & Özener, B. (2016). Fluctuating Asymmetry of Human Populations: A Review. *Symmetry*, *8*, 154. doi: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. *American Journal of Physical Anthropology*, *143*, 235–249. doi:10.1002/ajpa.21309.

Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, *69*, 383–393. doi:10.1080/01621459.1974.10482962.

Harris, E. F., & Bailit, H. L. (1988). A principal components analysis of human odontometrics. *American Journal of Physical Anthropology*, *75*, 87–99. doi:10.1002/ajpa.1330750110.

Hawkins, D. M. (1980). *Identification of Outliers*. Dordrecht: Springer Netherlands. OCLC: 851385856.

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*, e1421. doi:10.1002/wics.1421.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*, 5186–5201. doi:10.1016/j.csda.2007.11.008.

Kendall, E. J., Montgomery, J., Evans, J. A., Stantis, C., & Mueller, V. (2013). Mobility, mortality, and the middle ages: Identification of migrant individuals in a 14th century black death cemetery population. *American Journal of Physical Anthropology*, *150*, 210–222. doi:10.1002/ajpa.22194.

Kimber, A. C. (1990). Exploratory Data Analysis for Possibly Censored Data From Skewed Distributions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *39*, 21–30. doi:10.2307/2347808.

Klein, O., & Delacre, M. (2019). Routliers: Robust Outliers Detection. R package version 0.0.0.3.

Knudson, K. J., & Tung, T. A. (2011). Investigating regional mobility in the southern hinterland of the Wari empire: Biogeochemistry at the site of Beringa, Peru. *American Journal of Physical Anthropology*, *145*, 299–310. doi:10.1002/ajpa.21494.

Koenker, R. (2005). *Quantile Regression by Roger Koenker*. Cambridge University Press. doi:10.1017/CBO9780511754098.

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. *International Review of Social Psychology*, *32*, 5. doi:10.5334/irsp.289.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. doi: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. *Journal of Experimental Social Psychology*, *49*, 764–766. doi:10.1016/j.jesp.2013.03.013.

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*, *11*, e0153850. doi:10.1371/journal.pone.0153850.

Lightfoot, E., Šlaus, M., & O'Connell, T. C. (2014). Water consumption in Iron Age, Roman, and Early Medieval Croatia. *American Journal of Physical Anthropology*, *154*, 535–543. doi:10.1002/ajpa.22544.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, *6*, 1–39. doi: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. *American Journal of Physical Anthropology*, *147*, 499–507. doi: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. *International Journal of Osteoarchaeology*, *27*, 984–1002. doi:10.1002/oa.2610.

Mahoney, P. (2006). Dental microwear from Natufian hunter-gatherers and early Neolithic farmers: Comparisons within and between samples. *American Journal of Physical Anthropology*, *130*, 308–319. doi:10.1002/ajpa.20311.

Marwick, B. (2017a). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*, *24*, 424–450. doi:10.1007/s10816-015-9272-9.

Marwick, B. (2017b). Open Science in Archaeology, . doi:10.17605/OSF.IO/3D6XX.

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. *American Journal of Physical Anthropology*, *149*, 426–434. doi:10.1002/ajpa.22140.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, *1*, 327–332. doi:10.1016/0167-7152(83)90054-8.

Pickard, C., Girdwood, L.-K., Kranioti, E., Márquez-Grant, N., Richards, M. P., & Fuller, B. T. (2017). Isotopic evidence for dietary diversity at the mediaeval Islamic necropolis of Can Fonoll (10th to 13th centuries CE), Ibiza, Spain. *Journal of Archaeological Science: Reports*, *13*, 1–10. doi:10.1016/j.jasrep.2017.03.027.

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 (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, *88*, 1273–1283. doi:10.1080/01621459.1993.10476408.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The Bagplot: A Bivariate Boxplot. *The American Statistician*, *53*, 382–387.

doi:10.1080/00031305.1999.10474494.

Rousselet, G. A., & Wilcox, R. R. (2019). Reaction times and other skewed distributions: Problems with the mean and the median. *bioRxiv*, (p. 383935). doi: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). *American Journal of Physical Anthropology*, *156*, 374–387. doi:10.1002/ajpa.22663.

Schulte, E., Davison, D., Dye, T., & Dominik, C. (2012). A Multi-Language Computing Environment for Literate Programming and Reproducible Research. *Journal of Statistical Software*, *46*, 1–24. doi:10.18637/jss.v046.i03.

Sen Gupta, A. (2006). Generalized Variance. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (p. ess6053.pub2). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471667196.ess6053.pub2.

Stanisic, L., Legrand, A., & Danjean, V. (2015). An Effective Git And Org-Mode Based Workflow For Reproducible Research. *SIGOPS Oper. Syst. Rev.*, *49*, 61–70. doi:10.1145/2723872.2723881.

Stynder, D. D. (2009). Craniometric evidence for South African Later Stone Age herders and hunter–gatherers being a single biological population. *Journal of Archaeological Science*, *36*, 798–806. doi:10.1016/j.jas.2008.11.001.

Todorov, V., & Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, *32*, 1–47. doi:10.18637/jss.v032.i03.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science. Reading, Mass: Addison-Wesley Pub. Co.

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 Medical Research Methodology*, *14*. doi:10.1186/1471-2288-14-135.

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. *International Journal of Osteoarchaeology*, *23*, 698–715. doi:10.1002/oa.1298.

Weiss, E. (2009). Sex differences in humeral bilateral asymmetry in two hunter-gatherer populations: California Amerinds and British Columbian Amerinds. *American Journal of Physical Anthropology*, *140*, 19–24. doi:10.1002/ajpa.21025.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*. Statistical Modeling and Decision Science (3rd ed.). Amsterdam ; Boston: Academic Press.

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.).

Wright, L. E. (2005). Identifying immigrants to Tikal, Guatemala: Defining local variability in strontium isotope ratios of human tooth enamel. *Journal of Archaeological Science*, *32*, 555–566. doi:10.1016/j.jas.2004.11.011.