Supplementary material

Enamel surface topography analysis for diet discrimination.

A methodology to enhance and select discriminative parameters.

Arthur Francisco[1*], Cécile Blondel[2], Noël Brunetière[1], Anusha Ramdarshan[2], Gildas Merceron[2*]

[1] Institut Pprime, CNRS, Université de Poitiers, ISAE-ENSMA, F-86962 Futuroscope Chasseneuil, France

[2] Institut de Paléoprimatologie et Paléontologie Humaine: Evolution et Paléoenvironnements UMR 7262,

CNRS and Université de Poitiers, 86073 Poitiers Cedex 9, France

[*]corresponding authors: arthur.francisco@univ-poitiers.fr and gildas.merceron@univ-poitiers.fr

Outline

Nomenclature

Surfaces

PS2, *PS8*	least square polynomial surfaces of degree two, and degree eight
S1	primary extracted surface
SA, SB, SC	cleaned *S1* surface, *SA* minus its *PS2*, *SA* minus its *PS8*

Parameters (in order of appearance)

*fst_, lst_, mea_, med_, ent_*	05 percentile, 95 percentile, mean, median, value on the entire surface
*max_, min_, std_, MAX_, MIN_*	ten highest and lowest value mean, standard deviation, maximum, minimum
Sa, Sp,	arithmetic mean of the absolute of the heights, highest height
Sq, Sv, Sz	height standard deviation, absolute of the smallest height, amplitude of the heights
Sku, Ssk	surface kurtosis, skewness
f_{ACF}*(tx,ty), Rmax, Rmin*	autocorrelation function, f_{ACF} ellipse major(minor) axis
s, Sal, Str	height of the f_{ACF} ellipsis, *Rmin*, *Rmin/Rmax*
Sk, Smr1, Smr2, Spk, Svk	core height, end(beginning) of the hill area, hill(dale) area equivalent
Asfc	area-scale fractal analysis complexity parameter
Sbc	surface box counting dimension
Sar, Sm, Smd	surface relative area, height mean, height median
Sres, Ssa, Ssb	residual of Abbott-Firestone tangent fit, Abbott-Firestone tangent fit limits
Stp	ratio amplitude from 0.49 *Sz* to 0.51 *Sz*
Sh	percentage of nearly horizontal surface

Species

AB, AA, CS	*Alcelaphus buselaphus*, *Alces alces*, *Cephalophus silvicultor*

Statistics

α, β	risk of kind I and II *resp.*
λ	Box-Cox exponent

F', F	SSD_B/SSD_W , $(SSD_B/(k\text{-}1))/(SSD_W/(n\text{-}1))$ (Fisher's F statistic)
H_0, H_1	null, alternative hypotheses
k, p, n	number of groups, number of parameters, number of individuals
LSD, HSD	Fisher's *post hoc* Least Significant Difference, Tukey's *post hoc* Honest Significant Difference
MAD	Median Absolute Deviation
Ses	skewness standard error
SSD_B, SSD_W	sum of square of the deviations of groups: between and within
SSD_{total}	total sum of square of the deviations of groups: SSD_B+SSD_W
z_i , $\bar{x}$, $\tilde{x}$, $\hat{\sigma}$	z-score of individual i, parameter mean, parameter median, parameter estimated standard deviation

1. Preparation

1.1. Surface acquisition

The tooth surfaces are scanned using the Leica DCM8 optical profiler. The instrument can provide the benefits of two different technologies: high definition confocal microscopy for high lateral resolution and interferometry to reach sub-nanometer vertical resolution [1]. The device being utilized in its confocal configuration, surface elevations for each specimen are collected at a lateral interval of 0.129 µm with a vertical numerical step of 1 nm. Each measured surface is a raw digitized surface called "primary extracted surface", *S1*, according the ISO 25178 terminology [2]: it is a scale-limited surface in the way that it embeds a finite number of wavelengths.

1.2. Surface outlier cleaning

As recalled by Grubbs [3] an outlying observation, or "outlier", is one that appears to deviate markedly from other members of the sample in which it occurs. Therefore no universal procedure exists to remove extra points: it depends on the kind of outliers and the surrounding data. In the present case several procedures have been tested and the one that best suits our need, illustrated with Fig.S1.a), is the following:

1 A 5x5 kernel median filter is applied to *S1*, giving *S2*.

2 *S3* = *S1* -*S2* represents the *S1* deviation from the median.

3 *S3* is divided in 10 parts in each direction, and for each part the standard deviation is calculated. The global *S3* deviation σ is defined as the median value of the 10x10 standard deviations

4 inspired by the normal law, the procedure ends with the substitution of heights, for which abs(*S1*-*S2*) > 3σ, by median heights. The cleaned surface which is obtained will be called *SA* in what follows.

Eventhough the procedure is unusual, it provides satisfactory results in cleaning the primary extracted surfaces, without altering "real" points.

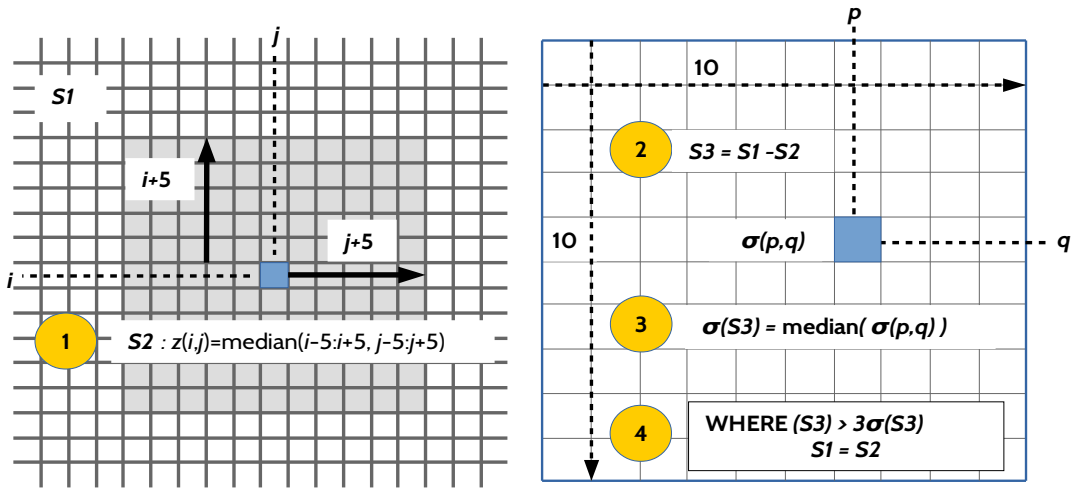


Figure S1: surface cleaning for outlier removal. The procedure is based on median kernels: whenever a height appears as suspicious (beyond the classical limits of a normal law) it is replaced by the kernel value it belongs to.

1.3. The autocorrelation function $f_{ACF}(tx,ty)$

The autocorrelation function $f_{ACF}(tx,ty)$, quantifies the self similarity of a surface shifted along a (tx, ty) vector. As an example, two profiles extracted from a real surface, Fig.S2.a), can be analyzed with 1D autocorrelation. Profile #1 is a wavy profile, Fig.S2.b); if the lag tx of the shifted profile is about 50µm, the signal roughly repeats, then at tx=50µm it reaches a local maximum, Fig.S2.c). The global maximum of f_{ACF} is obviously reached for tx=0, because the profile perfectly matches itself. The profile #2 exhibits less 'periodicity' but instead, a long wavelength component; therefore f_{ACF} decreases slowly, with no maximum value after tx=0. If the surface heights are white noise, *ie* the surface has no pattern, $f_{ACF}(tx,ty)=\delta(tx,ty)$: $f_{ACF}(0,0)=1$, $f_{ACF}(tx,ty)=0$ if $tx\neq0$ and $ty\neq0$. In contrast, surfaces with large scratches have a slow decaying f_{ACF} in the scratch direction, and wavy f_{ACF} across the scratches.

Fig.S2.d) illustrates the bidimensional case, and particularly the parameters *Rmin* and *Rmax* that are a signature of anisotropy when they differ from each other.

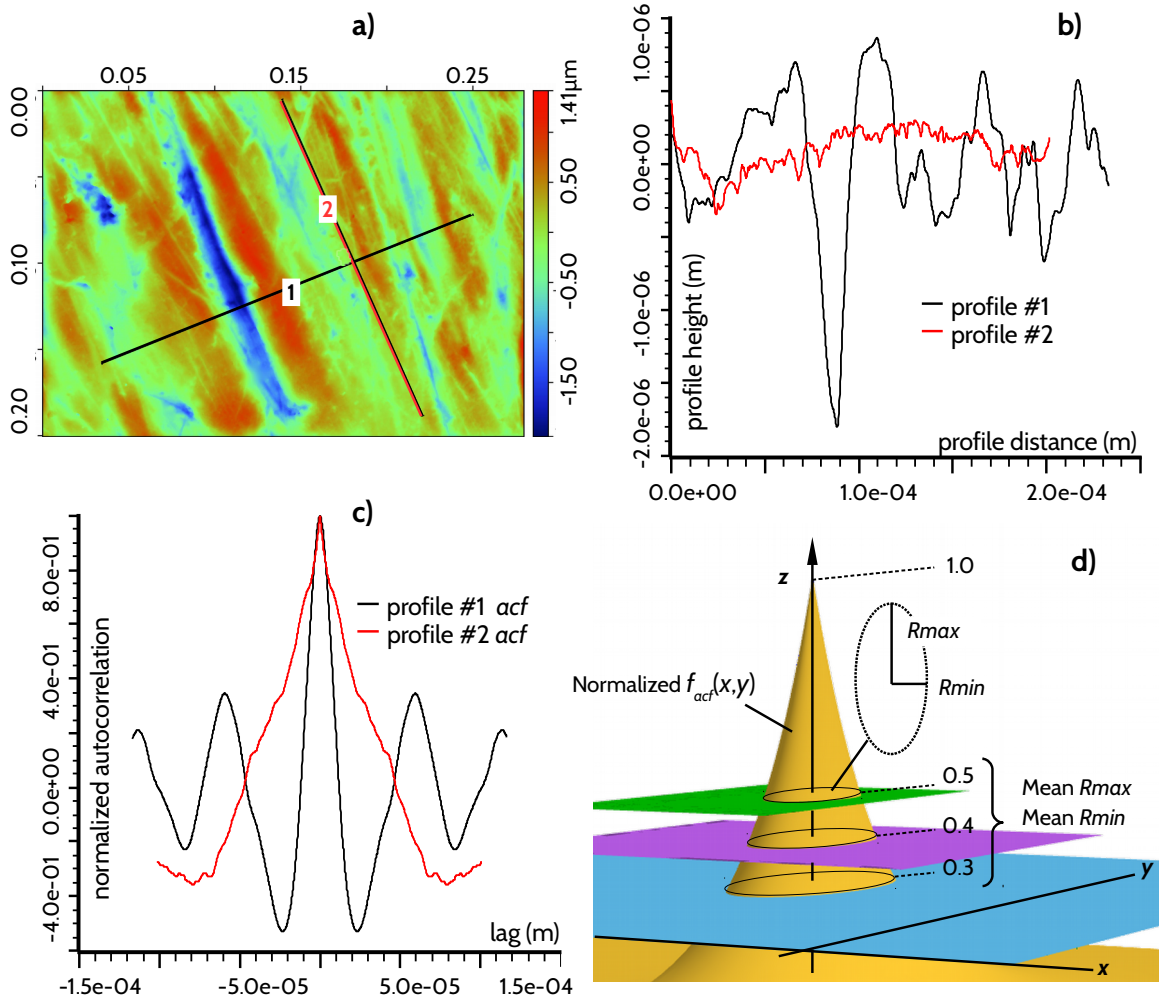
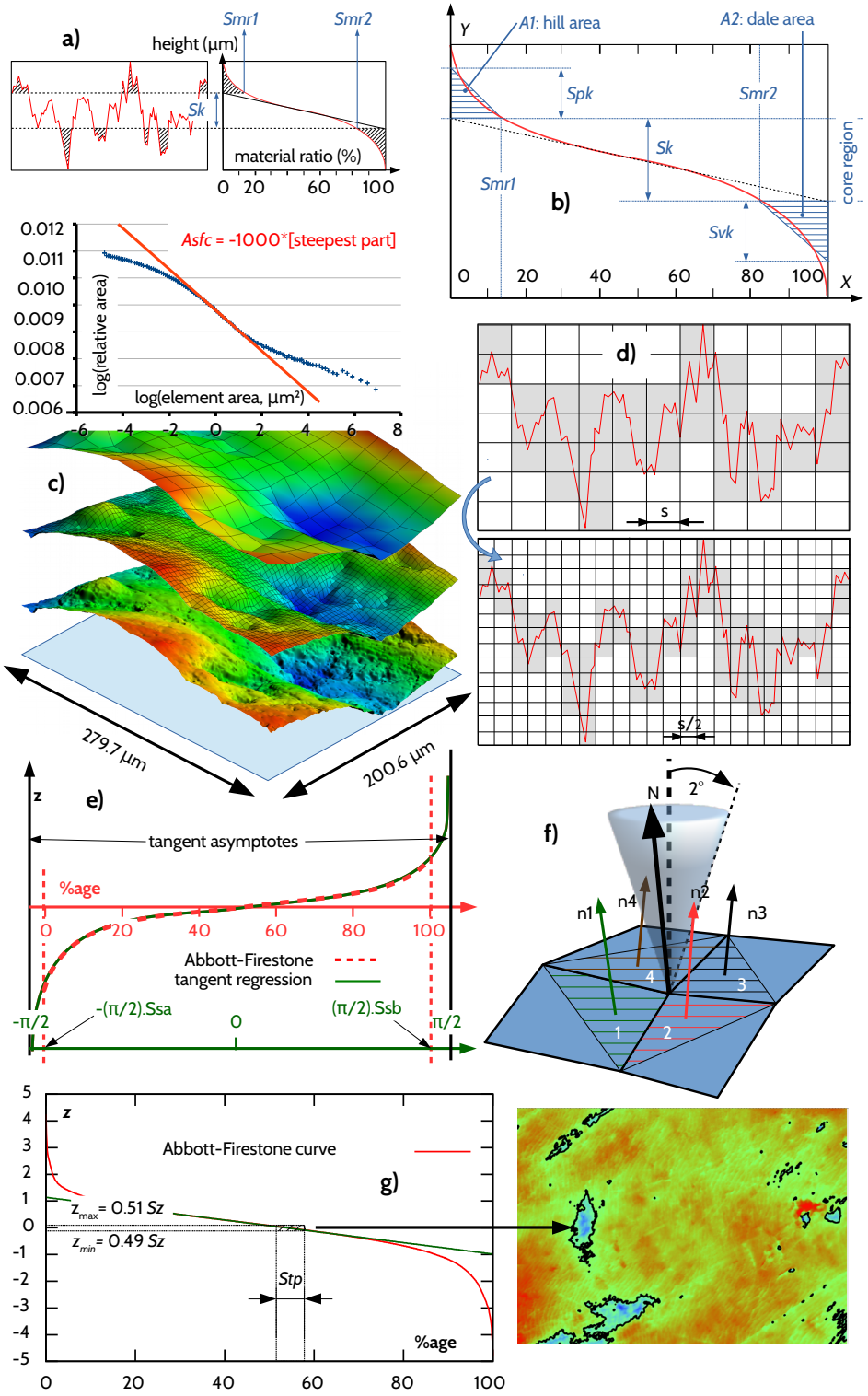


Figure S2: a) a typical dental surface to be analyzed. Two 1D profiles #1 and #2 are represented, #1 across the scratches and #2 along the scratches.
b) The profile #1 and #2 heights are represented; it clearly appears that the profile #1 shows more self-similarity than the profile #2, because of the long wavelengths.
c) Profile autocorrelations. When a profile doesn't repeat, its self-similarity decreases monotonously with the lag. Conversely, a wavy profile exhibits *acf* secondary peaks at the lag of the 'pseudo' periodicity.
d) 2D situation with a typical autocorrelation function related to an anisotropic surface and focused on the central peak. *Rmax* and *Rmin* are the ellipsis axes.

1.4. Functions, fractals and miscellaneous parameters

Fig.S3.a) illustrates the Abbott-Firestone material curve of a profile, from which the parameters *Smr1* and *Smr2* are defined. Fig.S3.b) goes deeper in the ISO 25178, with a representation of the parameters *Svk*, *Sk* and *Spk*. Fig.S3.c) illustrates the calculation scheme leading to *Asfc* determination. Instead of a straight line – for an ideal fractal surface – the curve log(relative area)=f(log(element area)) looks much more like a "*S*" shaped curve. That is the reason why the definition of *Asfc* specifies that the slope is measured on the steepest part of the curve. It is the parameter that needs the most computation time; it is fast when a few surfaces are analyzed, but it becomes slower with increased surface sampling (three hours for a 1024 sampled surface). Fig.S3.d) represents two stages in the *Sbc* determination; starting a few 3D blocks and refining until the maximum of blocks is reached (within the surface definition, 2168×1555 pt), the

function log(number of boxes)=f(log(box size)) can be plotted (not represented here). The function is almost linear for every studied surface and *Sbc* is the least square line slope. Fig.S3.e) illustrates the material curve fit with a tangent function. This way, even if there are remaining outlier heights, the parameters *Ssa* and *Ssb* – linked to the height skewness and kurtosis – are not affected. Fig.S3.f) shows a verticality 4° cone: if a normal region falls into the cone, it contributes to the global surface horizontality, *Sh*, expressed as a percentage. Fig.S3.g) illustrates the *Stp* determination; *Stp* catches the points in the surface mid-height.



FigureS3: a) Abbott-Firestone material curve of a profile, explaining *Smr1* and *Smr2*
b) Material curve parameters, following ISO 25178, *Svk*, *Sk* and *Spk*
c) *Asfc* calculation; three steps are illustrated
d) One step in the box-counting dimension *Sbc* calculation scheme; for fractal profiles the number of grayed squares increases as a power of their dimension.
e) Material curve fit: the horizontal axis of the Abbott-Firestone is reversed to allow for a tangent function fit. *Ssa* and *Ssb* are a measure of the distance of the material curve to the tangent asymptotes.
f) Flatness parameter *Sh*: percentage of quasi-horizontal faces (normal within a 4° cone)
g) *Stp* parameter determination – the steepest curve, the smallest *Stp*

2. The conservative ANOVA-based procedure

2.1. ANOVA background

There is a different approach regarding group discrimination. The Fig.S4.a) shows a typical situation where group responses seem to be different. However the variability inside a group can alter this *a priori* conclusion. Moreover, the sampling may not be accurate enough – with too few individuals – to suppose that the variability is representative of the whole initial population. ANOVA, as a discriminative tool, starts from the intuitive idea that separating groups will be easier if the dispersions (data range δ in the figure) are "small" compared to the differences between the locations (data mean μ in the figure). In more mathematical terms, the groups are considered as different if the ratio "variation between groups" over "variation within groups" is high enough.

An ANOVA toolbox can obviously provide descriptive statistics such as means, variations, … However, upon some assumptions on the data distribution, ANOVA can also make predictions on the initial population. In that case, it becomes an inference statistical tool. This situation is summed up by Eisenhart [4]: when the formulas are used to summarize properties of the data, no assumptions are needed to validate them, on the other hand, when ANOVA is used for inferring properties of the population, then certain assumptions, about the population and the sampling procedure must be fulfilled if the inferences are to be valid.

To make predictions about the group separation, one has to know how the data are spread around their mean. The simplest way to "prove" that the groups are different, under an α% risk, is to prove that there is little chance that they come from the same population. The subsequent null hypothesis H_0 is that the groups are samples of the same population, the differences being due to the inherent variability of the probability distribution.

In order to quantify the chances for that, the population is supposed Gaussian, as it often occurs in the nature. The group variances are obviously supposed equivalent otherwise, whether the groups are different – and no further test becomes necessary – or it is due to the sampling procedure – and any further interpretation becomes risky. On the basis of these assumptions, the F statistic previously defined is an F-distribution variable. Beyond a given value of F, the variable has less than α% chances (common threshold $F\alpha$) to occur. Thus, at a given risk α%, one should be able to bet on the group discrimination thanks to the F value.

The Fig.S4.b) illustrates the two situations under the null-hypothesis H_0. For low values of F, the variance between groups is small or of the same order than the variance within groups, the null-hypothesis can not be rejected. For high values of F, the variance between the groups is high enough to reject the null-hypothesis: one or more groups are sampled from other populations. The transition between the two states is usually set to common values: 5% or 2%. In

the present study α=5%. A convenient way to quantify the null-hypothesis rejection is to use the so-called *p*-value. The *p*-value just gives the probability of the null-hypothesis: even if groups can be visually distinguished, it can be due to random sampling error and *vice versa*.

2.2. Limitations

About the ANOVA: generalities. In many research fields, people are prone to use ANOVA for selecting discriminative parameters, but it has been shown that, carrying out carefully ANOVA *F*-tests needs specific pre-tests that, when added, increase the overall type II errors. Even if rules of thumb are also used to recover erroneously rejected parameters, some interesting parameters may be dropped. In addition, transforming the data makes the result interpretation uneasy, even if the data can easily be back transformed. The second point is that removing highly correlated parameters can make us reject truly discriminative parameters. Finally, selecting parameters on the basis of *post hoc* tests increases, once more, the overall type II error.

About the F-test: a power analysis. To illustrate the α and β risks in a *F*-test, let's consider two situations, where two samples of unity variances, are obtained from two populations P_0 and P_1, Fig.S4.c) The *F*-distribution, $F_{2,42}$, curve is the graph of the study *F* ratio probability. If P_0 and P_1 are statistically the same and the group means very different (on the right of the graph), it is concluded that they belong to two different populations (Type I error, risk α): the studied parameter is erroneously identified as discriminative. If P_0 and P_1 are statistically different but the group means close to each other (on the left of the graph), there is not enough evidence to reject H_0 (Type II error, risk β): the parameter is dropped from the parameter set.

As concerns the performed tests in the present study, the accepted risk is α=5%: we accept that 5% of the time, an error is committed in seeing a difference where there actually isn't. In the field of epidemiology, the α error is the most important one. Indeed, in comparing the efficiency of two vaccines – the usual one, and a candidate – if a false positive effect is detected it can lead to a decrease in the disease cover, unexpected side effects, etc.: the risk is usually greater than missing a progress (β risk). It is therefore common to set β as four times than α. However, when used for filtering parameters, the β risk is more important because erroneously considering a parameter as non-discriminative may result in poorer models.

The statistical power, 1-β, of a test is defined as the probability of detecting an existing difference: it is the statistical parameter on which one should focus for discriminative parameter selection. In a synthetic paper on the statistical power importance, Hallahan and Rosenthal [5], report studies, mainly from Cohen's work [6], for which the median power of the tests used to detect medium sized effects *d*=0.5 (in a *t*-test, *d* is the standardized difference of the means)

was below 50%. In other words, if true effects of medium size did exist, more than half of the studies would had less than 50% chance to detect them. According to Cohen [6], on the basis of 70 statistical studies published in the Journal of Abnormal and Social Psychology, 1960(6), “when one posits medium effects in the population the studies average slightly less than a 50-50 chance of successfully rejecting their major null hypotheses. No more than one-quarter of these studies have as good as three chances in five of succeeding under these conditions, and another quarter have less than one chance in three.”

In our work, the group size has been set to 15 individuals for practical reasons: specimen number, acquisition duration, etc. Therefore it is a *post hoc* statistical power that it is calculated (as Cohen did). Several free tools exist for that task, like *G*power* [7] or R packages. Fig.S4.d) shows how the group size, the statistical power and the effect size impact each other. For a desired 80% power, a 5% α risk and sample sizes set to 15, the effect size is qualified as large (d>0.4 in the conventional terminology): the tests that are performed on the parameter set are not statistically able to see small differences between the three groups. It can also be concluded that fine parameters able to find small differences between the groups (d=0.1) are likely to be rejected.

To conclude on the topic, if we are willing to catch differences between the groups, related to medium effect sizes (d≈0.3), the ANOVA F-test should be turned very permissive with α=35%, Fig.S4.e) In a prospective approach, with a 10% α risk and a 10% β risk, and medium effect size, the sample sizes should be around 170, which is not realistic: the mathematical rigor related to statistics may be not suitable for the kind of present study.

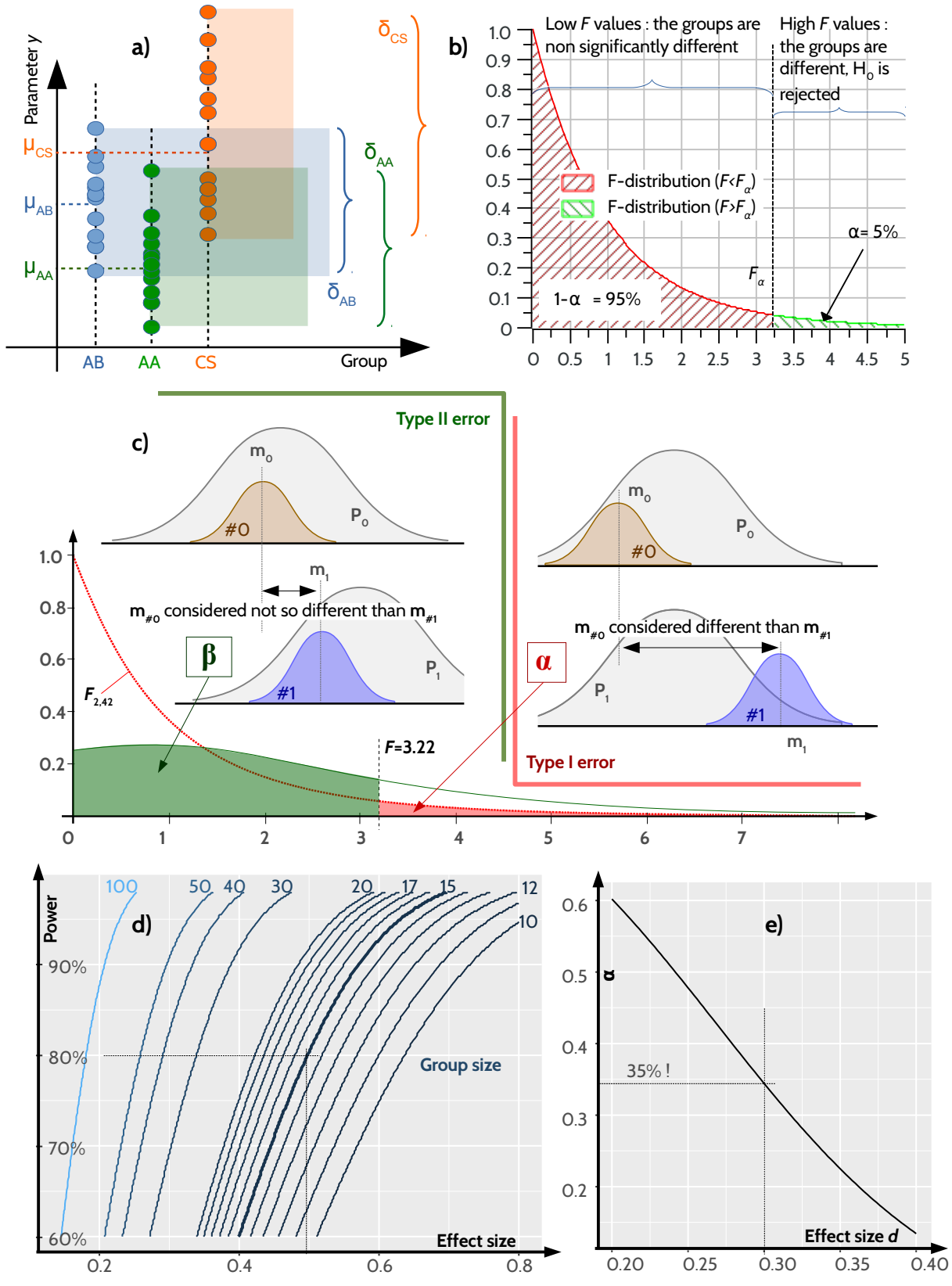


Figure S4: a) group discrimination as a function of position μ and dispersion δ
b) Density of *F*-statistic under H_0 ; three groups for 45 individuals, so *F*=*F*(2,42).
c) Type I/II errors and H_0/H_1 curves. On the right the drawing illustrates the risk I "seeing a difference where there actually is not" and the left one illustrates the risk II "not seeing any difference where there actually is"
d) Present study *post hoc* statistical power as a function of the effect size (standardized difference of the means).
e) α as a function of the effect size, power=80%, sample size=15

2.3. Data transformations

When one or more ANOVA *F*-test assumptions are violated, an alternative test can be carried out. The most common of it is a nonparametric test – the Kruskal-Wallis test, even if it is not specifically designed for normal distributions – which is based on the rank sorted data. On the subject McDonald [8] p158, doesn't recommend it as an alternative to one-way ANOVA. First, according to his experience the ANOVA *F*-test is quite robust and second, the Kruskal-Wallis test reveals sometimes to be less robust that the ANOVA *F*-test against heteroscedasticity. As concerns the other alternatives, they also suffer from weaknesses, that have not been yet fully investigated [9] p324. That is why data transformation is chosen in the present study rather than alternative tests.

The main drawback of mathematical data transformation, from skewed to normal for instance, is the loss of further understanding of the results such as the factor influences. Thus many authors restrict the transformation functions to some integer powers – from -2 to +2 – the square root and the logarithm. If the results are to be interpreted in terms of mean, influence, …, reversing the transformation once the analyzes are done makes the drawback vanish. Here, the transformations are only used for test purposes, the results being presented with the untransformed form. Instead of randomly try transformations to increase normality and/or variance homogeneity, Osborne [10] suggests to use an *ad hoc* designed tool, the so-called Box-Cox transformation, detailed by Eq.1

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y_i) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

The λ parameter is adjusted by maximum likelihood estimation, as the purpose is to fit a Gaussian distribution.

As will be shown in later sections, the parameter ent_Rmax_{SC} is of high importance for the group discrimination. However, without any transformation the 45-individual set is, at a significance level of 0.05, non-normal and furthermore heteroscedastic, Fig.S5.a) The "best" λ value is -0.5, meaning that the data should be transformed with an inverse square root function. As a result, the data become suitable for the ANOVA *F*-test. The effect of the Box-Cox transformation on the initial set of parameters is detailed in Fig.S5.b)

A more complete set of transformation functions is proposed by Johnson [11], the Johnson's translation system, that transforms to normality using the *Z* family of distributions, Eq.2, implemented in R package "Johnson" [40].

$$Z = \gamma + \delta f\left(\frac{X-\xi}{\lambda}\right) \quad f(x) \in \left\{\ln(x), \ln\left(\frac{x}{1-x}\right), \sinh^{-1}(x)\right\} \tag{2}$$

It performs the Johnson's transformation based on the method of the percentiles. Strange as it may seem, the Box-Cox

transformation has provided slightly better results than Johnson's. We think that it is mainly due to the implementation in the R package and the widely recognized difficulties to tune the transformation parameters. Indeed, the Johnson's system encloses three parameters and distinguishes three regions depending on the input data statistics. Fine tuning Johnson's parameters can be tricky and can result in unstable transformations. This explains why some authors still work to improve Johnson's parameter determination methods, [13].

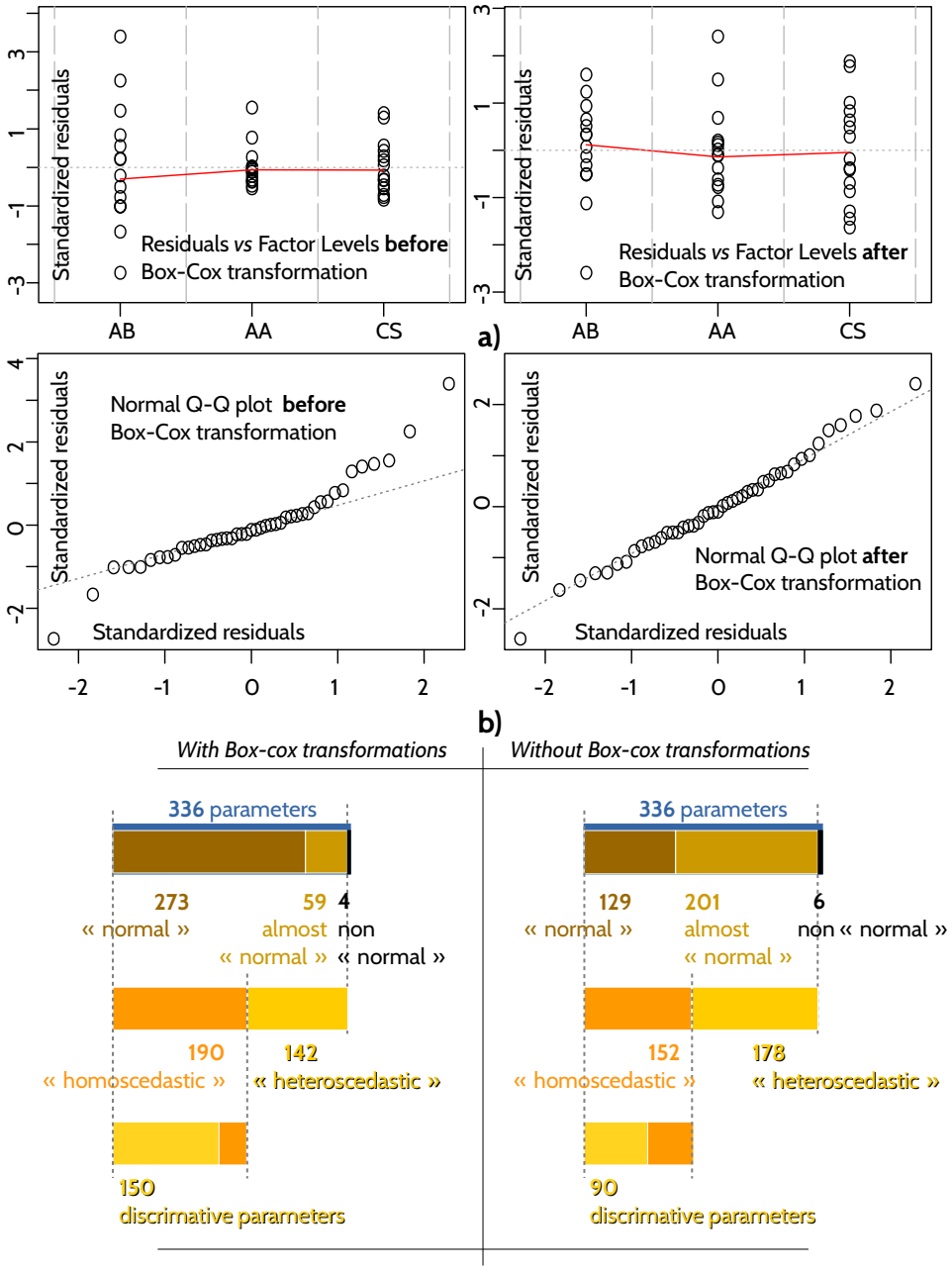


Figure S5: a) *ent_Rmax*$_{SC}$ group variances (top) and Q-Q plots (bottom) before transformation (left) and after transformation (right). It can be seen that the Box-Cox transformation increases the normality and the homoscedasticity. b) Overview of the benefits of the Box-Cox transformation: with the transformation (left) there are many more discriminative parameters than without transformation (right)

2.4. The different tests

Test for normality. Among the most used tests, the Shapiro-Wilks' test appears as the most powerful, in most cases [14–16]. It is the chosen test for normality in the present study. Nonetheless, it should be noted that the samples are quite small, and therefore choosing a test rather than an other may be not mandatory. To avoid a too conservative test, if a parameter fails, its skewness is compared to twice its skewness standard error. Then, if the value is below the skewness limit, the parameter is kept, with a "warning flag".

Test for variance homogeneity. Bartlett's test is widely used to test if the samples share nearly the same variance. Bartlett's test is sensitive to departures from normality so it is suggested to use Levene's test whenever the situation occurs [17]. In an automated procedure, it takes place if the skewness check has been used and has successfully passed – parameter with a "warning flag". Both tests are utilized because for nearly normal distributions, the Bartlett's test has a better performance [17]. Although homogeneity of variance is critical, a chance to recover the eligibility of the data is added with the variance rule of thumb, if the data has not been recovered thanks to the skewness test.

Test for outliers. The intuitive definition of an outlier would be "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism", [18] p1. But it is just 'suspicions' because long-tailed distributions can lead to outlier presence in the samples. When such a suspicion arises two behaviors are encountered: considering the outliers as contaminants that must be systematically removed [19] p70, followed by Schulz *et al.* [20], or as data to be kept because potentially carrying information, unless it is proved to be bad data [21] p68.

The most popular criterion is the 3σ rule which consists in labeling as outliers any data n°i with a z_i score above 3, Eq.3

$$z_i = \frac{x_i - \bar{x}}{\hat{\sigma}} \tag{3}$$

However as recalled by Pearson [22] p75, the historical and popularity of this convenient rule – also named ESD identifier (Extreme Studentized Deviation) – hides a major drawback: both $\bar{x}$ and $\hat{\sigma}$ are determined with the whole dataset – maybe including outliers. In addition with 15 individuals the detection can be erroneous. Shiffler [23] showed that z_i is bounded above by $\frac{n-1}{\sqrt{n}}$, *ie* 3.6, so a 3σ threshold can lead to too many outliers. Conversely, Pearson shows that the maximum detectable contamination is 10% with the 3σ rule; one single point in the present case.

To overcome the problem of influential outliers, the median statistic can be used instead of the mean. The standard

deviation estimator is then replaced by the MAD (Median Absolute Deviation) which is even more robust to outliers than the average absolute deviation. The modified z score is therefore $m_i = \frac{x_i - \tilde{x}}{\text{MAD}}$ but in order for the MAD estimator to converge towards σ for Gaussian datasets, a correction is brought to the previous expression, Eq.4

$$M_i = 0.6745 \frac{x_i - \tilde{x}}{\text{MAD}} \tag{4}$$

Indeed, for Gaussian data, $MAD \approx 0.6745\,\sigma$. Iglewicz and Hoaglin [24] recommend that modified z scores with an absolute value of greater than 3.5 be labeled as potential outliers, which is the chosen threshold in the present study.

There exists efficient statistical tests for multiple outlier detection, see Rosner [25] but they are not suitable for small samples. Hence, a *Q*-test (known as Dixon's test [26,27], with Rorabacher's corrections [28]) is also performed to test the greatest value and the lowest value against the null-hypothesis – H_0: there is no outlier in the sample. Dixon's *Q*-test examines the difference between the supposed outlier and the next closest observation relative to the overall range of the data. As the two aforementioned tests suppose an underlying normal distribution, an outliers is labeled 'NA' ('Not Available' following R convention) if both tests suggest it and if the data has successfully passed the normal test.

The ANOVA F-test. Once the required assumptions are met, the so-called "one way ANOVA" *F*-test is carried out. It is reminded that it tests if the group means are significantly 'non equal'. If the parameter has been recovered with the homoscedasticity rule of thumb, *ie* the variances are not 'so' different, the Welch's ANOVA [29] is used instead of classical ANOVA: the means are weighted by the reciprocal of the group mean variances.

2.5. About the statistical p-values

The LSD and HSD test *p*-values are presented on a log-log plot, Fig.S6 and it can be seen that:

- for three groups, the tests give very similar results; below the 5% level of significance the relationship is linear, the HSD test *p*-values being greater than the LSD tests' because the HSD test is more conservative. So, when there are only three groups, there is no need for carrying out both tests.
- The smallest *p*-values are related to the first and second groups, AB and AA *resp*.

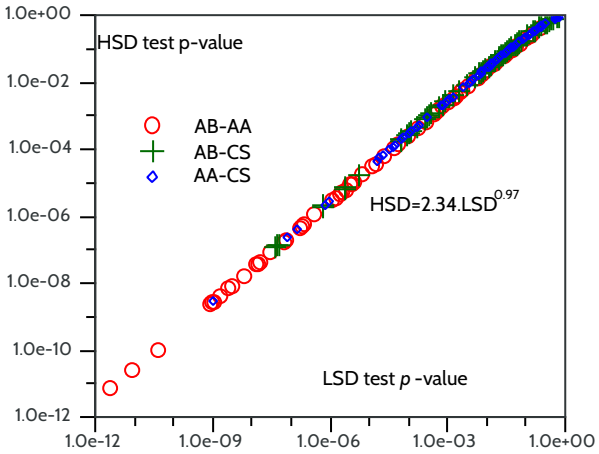


Figure S6: HSD *vs* LSD *post hoc* test

Concerning the latter point, exploiting the *p*-values – magnitude, comparison, … – is a controversial subject. On the one hand, the *p*-value represents the probability of obtaining a test statistic at least as extreme as the one that is performed: it can be expected that betting on an equality of means, when the *p*-value is dramatically low, is risky. On the other hand, the tests are designed in a yes/no manner regarding the null hypothesis along with a 95% significance level. According to the authors' experience, the lowest *p*-values, the more different the means.

3. The ANOVA-based simplification

Results without the last correlation step. The within group correlation step has been introduced in the global ANOVA procedure because it makes sense: if a parameter is strongly linked to others, it can be thought useless and may be removed safely. However, the correlation has to be considered on the whole surface set. Hence removing a parameter, that is found globally too much correlated to other parameters, may affect the positioning of hard to classify surfaces.

Figs.S7.a-c) detail the three biplots for the triplet (ent_Rmax_{SC}, fst_Ssk_{SC}, med_Sku_{SC}). The plots are obtained from the Top9 set keeping for each group the smallest *p*-valued parameter. The results are better than before (full ANOVA procedure), with the couple of parameters (ent_Rmax_{SC}, fst_Ssk_{SC}) which proves that focusing on correlations inside the groups does not make choosing the best discriminative parameters. The commonsense step which consists in suppressing the 'redundant' parameters is not suitable here: it may rather be introduced when bigger groups are used for building analytical predictive models.

Results without the first correlation step. The goal of removing highly correlated parameters (above 95%) is to significantly reduce the parameter set. Doing so, some parameters that discriminate slightly better than others may be dropped. Skipping this step, the new Top9 is detailed in Table S1. It is based on the *p*-values of the parameters that have successfully passed the *F*-test.

parameter	physical meaning	type	AB-AA	AB-CS	AA-CS
**ent_Rmax$_{SC}$*	*acf* ellipsis major axis value calculated on the entire *SC* surface	spatial	3.76E-11	3.81E-08	3.79E-02
**fst_Rmax$_{SC}$*	*acf* ellipsis major axis 5-percentile value of the resampled *SC* surface	spatial	8.27E-12	4.41E-08	1.05E-02
**fst_Str$_{SC}$*	*acf* ellipsis axis ratio 5-percentile value of the resampled *SC* surface	spatial	1.40E-08	4.71E-08	~~7.15E-01~~
**fst_Ssk$_{SC}$*	skewness 5-percentile value of the resampled *SC* surface	height	2.38E-12	~~6.16E-02~~	9.73E-10
MIN_Sp$_{SC}$	highest height lowest value of the resampled *SC* surface	height	2.13E-06	~~2.33E-01~~	3.87E-08
min_Sp$_{SC}$	highest height ten lowest value mean of the resampled *SC* surface	height	3.43E-06	~~2.56E-01~~	7.63E-08
mea_Sku$_{SC}$	kurtosis mean value of the resampled *SC* surface	height	6.78E-11	2.33E-03	2.68E-06
med_Rmax$_{SC}$	*acf* ellipsis major axis median value of the resampled *SC* surface	spatial	6.08E-10	2.46E-07	~~7.43E-02~~
! lst_Sku$_{SC}$	kurtosis 95-percentile value of the resampled *SC* surface	height	7.86E-10	1.13E-02	4.78E-06

Table S1 – *post hoc* results. A crossed *p*-value (above 5%) means non-significant differences
An asterisk prefixes the common parameters with the previous study.
The exclamation mark prefixes the parameters with an outlier value.

The final triplet is (*ent_Rmax$_{SC}$*, *fst_Ssk$_{SC}$*, *mea_Sku$_{SC}$*); it is very close to the previous one and does not bring more separability power, Figs.S7.d-e). The first correlation step can therefore be safely kept for screening purposes (when dealing with huge parameter sets), but for the present case, it appears to be unnecessary, it is deleted from the global procedure.

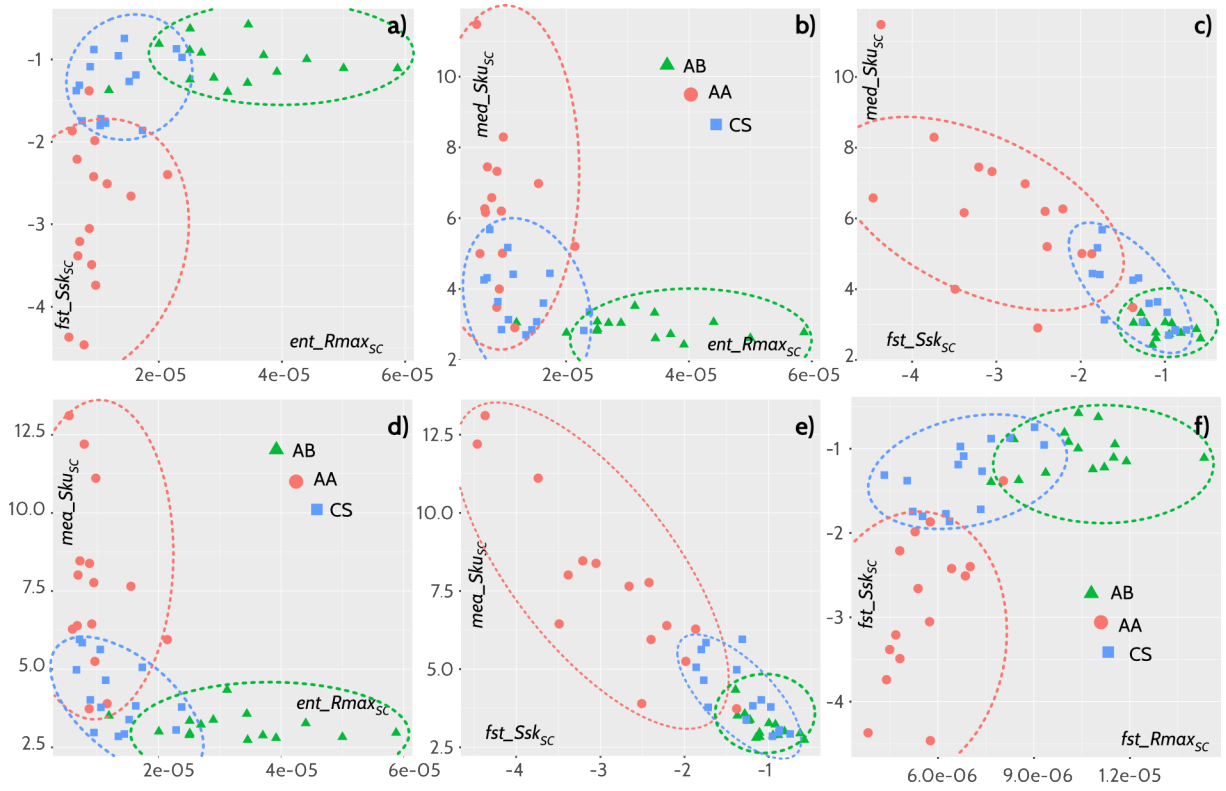


Figure S7: a)-b)-c) biplots with the best *p*-valued parameters, without the last correlation step. d)-e) Biplots with the best *p*-valued parameters, without the first correlation step. The missing biplot (*ent_Rmax*$_{SC}$, *fst_Ssk*$_{SC}$) is the same as a) The discrimination is neither poorer nor better than before. f) Biplot with the two best untransformed parameters. *fst_Rmax*$_{SC}$ has lower *p*-values than *ent_Rmax*$_{SC}$ however the results with *ent_Rmax*$_{SC}$ are better. It can be concluded that the Box-Cox transformations are really useful. The ellipses are qualitative representations of the groups.

Results without ANOVA F-test. According the literature, *post hoc* tests should only be performed after a significant difference between groups has been shown. However it can also be found that the rejection of H_0 is not a prerequisite, *e.g.* for HSD [30] p1570. The LSD *post hoc* test is then carried out without prior *F*-test; this test is kept instead of HSD because, according Fig.S6 results, the largest p-values are 1.E-7. The *post hoc* test leads to the same Top9 parameter set. As a consequence, the one-way ANOVA test can be skipped.

Results without Box-Cox transformations.

parameter	physical meaning	type	AB-AA	AB-CS	AA-CS
**fst_Rmax$_{SC}$*	*acf* ellipsis major axis 5-percentile value of the resampled *SC* surface	spatial	5.40E-12	1.06E-08	2.17E-02
**ent_Rmax$_{SC}$*	*acf* ellipsis major axis value calculated on the entire *SC* surface	spatial	4.93E-10	2.38E-08	~~2.41E-01~~
**med_Rmax$_{SC}$*	*acf* ellipsis major axis median value of the resampled *SC* surface	spatial	1.18E-09	5.55E-08	~~2.47E-01~~
**fst_Ssk$_{SC}$*	skewness 5-percentile value of the resampled *SC* surface	height	8.82E-11	~~2.29E-01~~	4.47E-09
mea_Ssk$_{SC}$	skewness mean value of the resampled *SC* surface	height	2.76E-09	~~4.61E-01~~	3.18E-08
ent_Ssk$_{SC}$	skewness value calculated on the entire *SC* surface	height	5.76E-09	~~3.63E-01~~	1.19E-07
mea_Rmax$_{SC}$	*acf* ellipsis major axis mean value of the resampled *SC* surface	spatial	4.77E-10	1.43E-07	~~8.90E-02~~
MIN_Rmax$_{SC}$	*acf* ellipsis major axis lowest value of the resampled *SC* surface	spatial	6.81E-10	7.53E-07	3.89E-02
min_Rmax$_{SC}$	*acf* ellipsis major axis ten lowest value mean of the resampled *SC* surface	spatial	7.94E-10	1.58E-06	2.59E-02

Table S2 – *post hoc* results when the parameters are unchanged.
A crossed *p*-value (above 5%) means non-significant differences.
An asterisk prefixes the common parameters with the previous Top9

Despite similar *p*-values, the separation is not as clear as with transformed parameters, Fig.7.f). It can be concluded that increasing the normality and variance homogeneity of the parameters makes the *post hoc* tests more efficient, even if the gain remains slight.

4. Analysis of dimensionless surfaces

4.1. With the full conservative procedure

There are fewer parameters (305) because normalizing the surfaces makes some parameters irrelevant, *e.g. ent_Sq*. After the whole ANOVA procedure (assumption checking and ANOVA test), 165 parameters remain, from which the Top9 set is presented in Table S3.

Top9 dimensionless	*ent_Rmax$_{SC}$*	*fst_Str$_{SC}$*	*fst_Rmax$_{SC}$*	*fst_Ssk$_{SC}$*	*ent_Ssk$_{SC}$*	*min_Ssk$_{SC}$*	*mea_Sku$_{SC}$*	*med_Rmax$_{SC}$*	*MIN_Sv$_{SB}$*
Top9 dimensioned	x	x	x	x	x				
Height (h) or spatial (s)	s	s	s	h	h	h	h	s	h

Table S3 – dimensionless Top9 parameters

The fact that the Top9 dimensionless set is close to the one detailed in Table S1 (dimensioned parameters) was foreseeable. Indeed, the parameter built on the autocorrelation function – *Rmax* – depends only on the lateral scale, and the statistics *Ssk* and *Sku* don't depend on the height mean, nor on the height standard deviation. Consequently the same observations hold: the surface *SC* seems to bring much more information (eight parameters) than *SB* and the numbers of height and spatial parameters are balanced.

The three selected parameters ,with the lowest correlation within group filter, are presented in Table S4.

parameter	physical meaning	type	AB-AA	AB-CS	AA-CS
fst_Rmax_{SC}	*acf* ellipsis major axis 5-percentile value of the resampled *SC* surface	spatial	8.28E-12	4.41E-08	1.06E-02
ent_Ssk_{SC}	skewness value calculated on the entire *SC* surface	height	5.82E-09	~~3.38E-01~~	1.43E-07
mea_Sku_{SC}	kurtosis mean value of the resampled *SC* surface	height	6.78E-11	2.33E-03	2.68E-06

Table S4 – the three less correlated parameters that best identify the groups (*p*-values crossed for non -H_0-rejection)

The biplots are not presented here because the separation power reached by the Top3 set is lower than the one obtained without the within group correlation step as detailed hereafter.

4.2. With the full procedure, skipping the last correlation step

The results are as satisfactory as for non normalized surfaces. The Table S5 details the Top3 set.

parameter	physical meaning	type	AB-AA	AB-CS	AA-CS
fst_Ssk_{SC}	skewness 5-percentile value of the resampled *SC* surface	height	2.39E-12	~~6.16E-02~~	9.74E-10
ent_Rmax_{SC}	*acf* ellipsis major axis value calculated on the entire *SC* surface	spatial	3.77E-11	3.81E-08	3.80E-02
fst_Rmax_{SC}	*acf* ellipsis major axis 5-percentile value of the resampled *SC* surface	spatial	8.28E-12	4.41E-08	1.06E-02

Table S5 – Top3 parameter set for normalized surfaces, skipping the last correlation step – (*p*-values crossed for non -H_0-rejection)

As for skipping the transformations, it is the same Top9 as for dimensioned surfaces, the conclusion is therefore the same: the Box-Cox transformations must be kept.

References

1. In press. Optical profiler Leica DCM8. *Leica Microsyst.*

2. In press. ISO 25178-2:2012 - Geometrical product specifications (GPS) -- Surface texture: Areal -- Part 2: Terms, definitions and surface texture parameters. *ISO*.

3. Grubbs, F. E. 1969 Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11**, 1–21. (doi:10.1080/00401706.1969.10490657)

4. Eisenhart, C. 1947 The Assumptions Underlying the Analysis of Variance. *Biometrics* **3**, 1–21. (doi:10.2307/3001534)

5. Hallahan, M. & Rosenthal, R. 1996 Statistical power: concepts, procedures, and applications. *Behav. Res. Ther.* **34**, 489–499.

6. Cohen, J. 1962 The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* **65**, 145–153.

7. In press. Universität Düsseldorf: G*Power.

8. McDonald, J. H. 2014 *Handbook of Biological Statistics*. 3rd edn. Baltimore, Maryland: Sparky House Publishing.

9. Osborne, J. W. 2007 *Best Practices in Quantitative Methods*. Thousand Oaks, Calif: SAGE Publications, Inc.

10. Osborne, J. W. 2010 Improving Your Data Transformations: Applying the Box-Cox Transformation. *Pract. Assess. Res. Eval.* **15**.

11. Johnson, N. L. 1949 Systems of Frequency Curves Generated by Methods of Translation. *Biometrika* **36**, 149–176. (doi:10.2307/2332539)

12. Fernandez, E. S. 2014 *Johnson: Johnson Transformation*. [cited 2016 Jul. 26].

13. George, F. & Ramachandran, K. 2011 Estimation of Parameters of Johnson's System of Distributions. *J. Mod. Appl. Stat. Methods* **10**.

14. Razali, N. M., Wah, Y. B. & others 2011 Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *J. Stat. Model. Anal.* **2**, 21–33.

15. Yap, B. W. & Sim, C. H. 2011 Comparisons of various types of normality tests. *J. Stat. Comput. Simul.* **81**, 2141–2155. (doi:10.1080/00949655.2010.520163)

16. Ahmad, F. & Khan, R. A. 2015 A power comparison of various normality tests. *Pak. J. Stat. Oper. Res.* **11**, 331–345. (doi:10.18187/pjsor.v11i3.845)

17. In press. 1.3.5.7. Bartlett's Test.

18. Hawkins, D. M. 1980 *Identification of Outliers*. Dordrecht: Springer Netherlands. [cited 2016 Jul. 26].

19. Cornillon, P.-A. & Matzner-Lober, E. 2010 *Régression avec R*. Paris; New York: Springer Editions.

20. Schulz, E., Calandra, I. & Kaiser, T. M. 2010 Applying tribology to teeth of hoofed mammals. *Scanning* **32**, 162–182. (doi:http://dx.doi.org/10.1002/sca.20181)

21. Quinn, G. P. & Keough, M. J. 2002 *Experimental Design and Data Analysis for Biologists*. 1 edition. Cambridge, UK ; New York: Cambridge University Press.

22. Pearson, R. 2005 *Mining Imperfect Data*. Society for Industrial and Applied Mathematics. [cited 2016 Jul. 26].

23. Shiffler, R. E. 1988 Maximum Z Scores and Outliers. *Am. Stat.* **42**, 79–80. (doi:10.2307/2685269)

24. Iglewicz, B. & Hoaglin, D. C. 1997 *How to Detect and Handle Outliers*. Milwaukee, Wis: ASQC/Quality Press.

25. Rosner, B. 1983 Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* **25**, 165–172. (doi:10.2307/1268549)

26. Dixon, W. J. 1950 Analysis of Extreme Values. *Ann. Math. Stat.* **21**, 488–506.

27. Dixon, W. J. 1951 Ratios Involving Extreme Values. *Ann. Math. Stat.* **22**, 68–78.

28. Rorabacher, D. B. 1991 Statistical treatment for rejection of deviant values: critical values of Dixon's 'Q' parameter and related subrange ratios at the 95% confidence level. *Anal. Chem.* **63**, 139–146. (doi:10.1021/ac00002a010)

29. Welch, B. L. 1951 On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika* **38**, 330–336. (doi:10.2307/2332579)

30. Salkind, N. J., editor 2010 *Encyclopedia of research design*. Thousand Oaks, Calif: SAGE Publications.