Schrödinger's phenotypes: images of herbarium specimens are both good and (not so) bad sources of morphological data

Authors Leonardo M. Borges[1], Victor Candido Reis[2], Rafael Izbicki[2]

[1] Departamento de Botânica, Universidade Federal de São Carlos, São Carlos, SP, Brazil
[2] Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

Correspondence Leonardo M. Borges
Departamento de Botânica, Universidade Federal de São Carlos, Rodovia Washington Luís, Km 235, São Carlos, SP, 13565-905, Brazil
aquitemcaqui@gmail.com

Author contribution LMB conceived the idea. LMB and RI designed the research. LMB collected original data and compiled legacy datasets. VCR and RI designed and performed the analyses. LMB and RI wrote the manuscript. All authors approved final publication. LMB and RI contributed equally to this work.

Running headline Sourcing digitized plant specimens for trait data

Schrödinger's phenotypes: images of herbarium specimens are both good and (not so) bad sources of morphological data

Abstract 1. Museum specimens are the main source of knowledge on organisms' morphological features. Although this knowledge was commonly restricted to a small number of researchers, at least for plants, it is now freely available thanks to the digitization of millions of museum specimens. With these images, we will be able to collectively build large-scale morphological datasets, but they will only be useful if the limits to this approach are well-known. In order to establish these limits, we tested the precision and accuracy of image-based data and analyses.

2. To test measurement precision and accuracy, we compared leaf measurements taken from specimens and from images of the same specimens. Then we used legacy morphometric datasets to establish differences in quality of datasets and multivariate analyses between specimens and images. To do so, we compared the multivariate space based on original legacy data to spaces built with datasets that simulate image-based data.

3. We found that trait measurements in images is as precise as those obtained directly from specimens. On the other hand, accuracy of image measurements drops for smaller traits. This decrease in accuracy, however, has very low impact on dataset quality and derived analyses. The main problem with image-based datasets comes from missing observations due to image resolution or organ overlapping. Missing data lowers accuracy of datasets and similarity between multivariate analyses based on image and specimen data. Although not too strong, this decrease in accuracy asks for caution when designing morphological research that will rely on digitized specimens.

4. Images of herbarium specimens are reliable sources of measurements, even though accuracy of smaller traits lowers due to resolution problems. On the other hand, the impossibility of observing particular traits affects the quality of image-based datasets and, thus, of derived analyses. Despite these issues, gathering phenotypic data from images is overall trustworthy and may support large-scale studies on morphological diversity and evolution.

Keywords digitization, plant morphology, trait, crowdsourcing

1 Introduction

Specimens held in natural history collections give us most information on the diversity, morphological variety, and spacial and temporal distribution of living creatures. Based on this information, we obtain knowledge on many areas, such as environmental changes, public health, and evolution (Law & Salick 2005, Suarez & Tsutsui 2004, Davis et al. 2015, Babin-Fenske et al. 2008). Until recently, however, these specimens were available only to people with personal access to museum collections. Now, almost all institutions are digitizing their specimens, and plant museums (herbaria) are leading the way due to the usually flat, regular in size, and easy to image nature of plant samples. Millions of plant specimens' images are already available online (Soltis 2017, Le Bras et al. 2017), and ready to be used as part of scientific research.

Taxonomy is one area in which herbaria digitization was very fruitful. For example, taxonomic practice and herbarium curation can now be sped up with automatic species identification software (Remagnino et al. 2016, Carranza-Rojas et al. 2017). These tools, however, usually classify specimens by their overall image patterns, without making assumptions about organ identity (Remagnino et al. 2016). Trait delimitation with images, a much more complex but still underdeveloped task, is promising. Previously developed methods are able to correctly identify particular organs in images and to extract data from them (Corney, Tang, Clark, Hu & Jin 2012, Corney, Clark, Tang & Wilkin 2012, Gehan et al. 2017). Even if we still have to wait for automatic data extraction to become highly efficient, specimens imaging already facilitates data acquisition by researchers themselves, particularly if used in collective efforts to build comprehensive morphological datasets (e.g. O'Leary et al. 2013).

Such phenotipically more diverse datasets are becoming fundamental to support advances in comparative and evolutionary biology (Laing et al. 2018). However, in comparison to genomic data, they are slower and more expensive to produce (Burleigh et al. 2013). To overcome these challenges, a number of tools are being developed, and many of them rely on images (Burleigh et al. 2013). However, there are limits to this approach, particularly for herbarium specimens, commonly digitized as scans or photographs (Tulig et al. 2012, Takano et al. 2019, Sweeney et al. 2018). First, image resolution may limit observations, specially of small traits, which are more likely to be closer to resolution limits. Second, the nature of plants themselves, or of plant specimen preparation, hides some organs. For example, stamens in bean's flowers are naturally hidden by part of the petals, and, after pressing, the whole flower can be overlapped by leaves. Although this is not a problem when one has personal access to specimens, it may become an issue for images. Thus, such variations in the preparation and nature of herbarium specimens may impact image-based data collection.

To evaluate the influence of these restrictions on acquisition of morphological data from digitized herbarium specimens, we asked (1) if image measurements significantly differ from specimen measurements, and (2) if image-based datasets differ from the ones built with specimens to the point of affecting morphological analyses. To answer our first question, we compared measurements taken from specimens and their digital version. To answer our second question, we compared the results of multivariate analyses using specimens with analyses simulating image usage.

2 Materials and Methods

2.1 A word on terminology

Before explaining our analytic approach, we make some of the terms we use here explicit, in particular accuracy and precision (for details, see Streiner & Norman 2006). We consider precision to be an estimate of the variation between multiple measurements of the same feature. If values for these measurements do not differ widely, precision is high. Accuracy is treated as the difference between a measurement and a reference value, and also applied here to judge results of multivariate analyses. Close similarity between the measurement (or analyses' results) and the reference value indicates high accuracy.

Because our goal is to evaluate the confidence of images as a data source in comparison to that of specimens', we treated specimen data as the reference values against which we judged accuracy of image data.

2.2 Measurement precision and accuracy

In order to test if measurements made in images and specimens differ, we measured the size of different leaf parts of specimens from the Rio de Janeiro Botanical Garden herbarium (RB; acronym according to Thiers 2020, continuously updated) belonging to five species (hereafter, `measurement dataset`). These species (*Inga vera* Willd., *Ocotea divaricata* Mez, *Piper anisum* (Spreng.) Angely, *Smilax fluminensis* Steud. and *Sphagneticola trilobata* (L.) Pruski) were selected to encompass different leaf morphologies, such as simple, compound, and lobed leaves with different base and apex shapes. Besides, by using these taxa, we could avoid issues with sampling size, as all of them are well represented in RB (see the supplementary material for information on measurements, sampling, and vouchers. Metric data is also available from MorphoBank [project reference to be made available after review]). We used a digital caliper to measure specimens and the FSI Viewer v. 5.6.6 software (available from RB's website; `http://jabot.jbrj.gov.br`) to measure images of the same leaves from the same specimens. All measurements, both in images and specimens, were made twice in order to allow following analyses.

First we tested how precise are measurements made either in the specimens or in the images. To do so, we evaluated how each of the two repeated measurements are close to each other with the intraclass correlation coefficient (ICC) (Bartko 1966). Then, we tested the hypothesis that the average precision is the same for images and specimens. For that, we compared the difference between the two repeated measurements made on images with those made in specimens with a paired t-test. The last test in this context evaluated if precision varies according to measurement scale. For that, we first divided the absolute difference between the two repeated measurements by their average to obtain measurements' relative precision. Then, we plot the relative precision of each observation against a scale reflecting overall trait size.

Having tested precision, we computed the average between the two measurements of each variable. We used these averages to avoid bias in following measurement accuracy analyses. For each specimen and variable, we evaluated accuracy between measurements in the images with respect to the specimens'

by testing their similarity with ICC (Bartko 1966). We further evaluated accuracy testing the hypothesis that the median difference between image and specimen measurements is zero with a Wilcoxon signed rank two-sample test (Wilcoxon 1945). Finally, in order to quantify the relationship between measurement error and the magnitude of each variable, we plot the relative error of the measurements (i.e., the absolute difference between image and specimen values divided by image mensuration) against the image measurement itself.

2.3 Accuracy of image-based datasets

To test if image-based data impacts analyses of morphological data, we compiled datasets of morphometric studies of different plant groups (Alcantara et al. 2013, Ames et al. 2008, Andres-Sanchez et al. 2009, Bello et al. 2018, Bünger et al. 2015, Egan 2015, Kučera et al. 2006, Marhold 1992, Poulsen & Nordal 2005, Rose & Freudenstein 2014, Slovák et al. 2012, Trovó et al. 2008). These datasets (hereafter "`original datasets`") were used to generate three "`synthetic datasets`", which where modified to include variation as if they had been compiled from images: the first included bias observed in the measurement accuracy analysis described above; the second included missing data; and the third included both measurement bias and missing data. We generated `synthetic datasets` with the following procedures.

First, we added noise to the `original dataset` using a zero-mean Gaussian distribution with varying values of variance. To do so, we modeled how the standard deviation of the measurements' error varies as a function of each variable's magnitude (See supplementary material). We then used this model to simulate different values of standard deviations for the Gaussian distribution of each variable, as to reproduce the same level of measurement error expected for that variable.

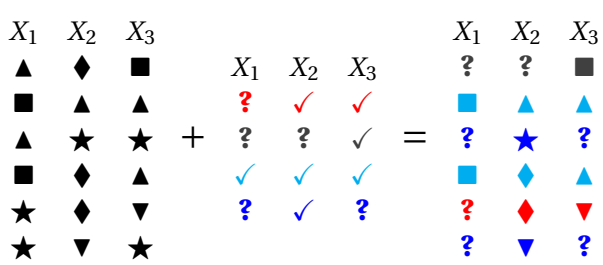

Figure 1: Illustration of the masking procedure. Left: `original` data set. Center: missing patterns obtained using digitized specimens; question marks indicate missing information and check marks indicate variables that are available. Right: masked data used for constructing the `synthetic dataset`, obtained by applying a randomly chosen missing pattern (from the middle table) to each row of the original data set.

Second, to simulate missing observations, we first checked which variables could be obtained from images on a set of up to ten digitized specimens from the same taxa of each morphometric study (see supplementary material for information on variables and digital vouchers). Using this information, we modeled patterns of missing data, which we used to mask observations in the `original dataset`, pro-

ducing synthetic data including missing observations (Fig. 1). Finally, we combined the approaches just described to simulate presence of both noise and missing data.

With `original` and `synthetic datasets` at hand, we compared agreement between results of Principal Component analyses (PCA) of all of them. Besides being one of the most common methods in morphometric analyses, differences in PCA results summarize the overall variation of datasets being compared. Thus, PCA comparison captures differences between data acquisition methods (images vs. specimens).

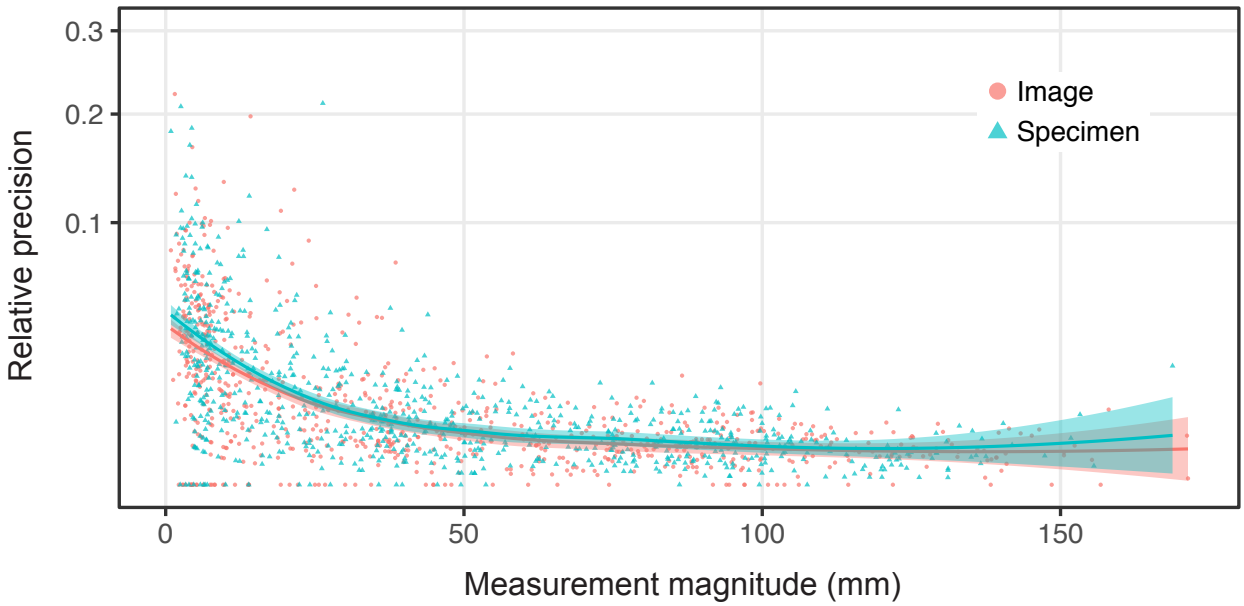

Figure 2: Relationship between the relative precision of measurements (i.e., the absolute difference between two repeated measurements divided by their average) and magnitude of each variable. Dots are different variable observations from the `measurement dataset`, either made on images (red) or on specimens (blue). The curves indicate the mean trend, along with their standard errors. Y-axis displayed in square root scale.

Our protocol first used `pcaMethods` (Stacklies et al. 2007) to run a Probabilistic PCA (Roweis 1998) for each dataset, which automatically handles missing data with an expectation-maximization algorithm. We then measured the agreement between the first four principal components (PC) inferred for `original` and `synthetic data` by computing Pearson's correlation coefficient (Pearson 1920) between the scores obtained for each observation. High correlation indicates agreement between analyses, and consequently between acquisition methods. However, as PC correlations may not capture overall dataset variation (Yang & Shahabi 2004, Melo et al. 2015), we also used `EvolQG` (Melo et al. 2015) to measure PCA similarity (Yang & Shahabi 2004) between `original` and `synthetic data`. PCA similarity weights PC correlation by their eigenvalues (Yang & Shahabi 2004, Melo et al. 2015), and, thus, takes into account

how much data variability is expressed in each PC.

The process described above was repeated 100 times (that is, 100 `synthetic datasets` were created), so we could evaluate accuracy of image-based analyses to specimen-based PCAs with mean values and 95% confidence intervals.

3 Results

Measurements taken both from images and specimens have high internal precision, as seen in intraclass correlation coefficient values of 100.0%. Moreover, the hypothesis that precision is the same between images and specimens was not rejected (p=0.678). Indeed, Fig. 2 indicates that measurements made on images and on specimens have similar precision, and that, independently of the source, precision decreases together with trait size.

Similarly, high ICC values indicate that measurements taken from images are accurate overall (Fig. 3). This lack of bias is reinforced by non-rejection of the hypothesis that they would not differ from specimens measurements (Fig. 4). On the other hand, although measurements of larger traits are quite similar between specimens and images, measuring smaller organs with images is not as accurate. For example, blade length measurements of *Ocotea divaricata* or *Piper anisum*, which vary between 50–160 mm, are extremely similar between specimens and images, as shown by the almost perfect fit of point to the identity line (curve $y = x$) in Fig. 3. On the other hand, points comparing petiole length measurements of the same species (2–15 mm long) are more dispersed around the identity line. This effect is clearly seen in the plot between relative error and specimen measurement values (Fig. 5), which shows that differences between image and specimen measurements are higher for smaller traits.

All but two `synthetic datasets` (E2015, T2008) include missing data. Total proportion of missing observations, percentage of variables with at least 50% of observations, and variables that were impossible to observe in images (Table 1) vary between `synthetic datasets`, but at least one of these classes lacks more than 25% of observations for most datasets. `Synthetic datasets` also differ in proportions of missing data by type of variable (continuous, categorical, discrete, or ratio; Table 2), with continuous variables being more commonly absent. Even though patterns of missing data may differ between datasets, correlation coefficients indicate that drops in accuracy are more related to absences spread in the dataset than to complete lack of a particular variable (Table 3).

The analyses of `original` and `synthetic datasets` (Fig. 6) show that the results of principal component analyses based on images are affected by noise and missing data. Nonetheless, noise effect is almost non-existent, as shown by values near 100% for both PCA similarity and PC correlation for the first and second principal components. On the other hand, accuracy drops in face of missing data, and particularly due to joint effect of noise and missing data.

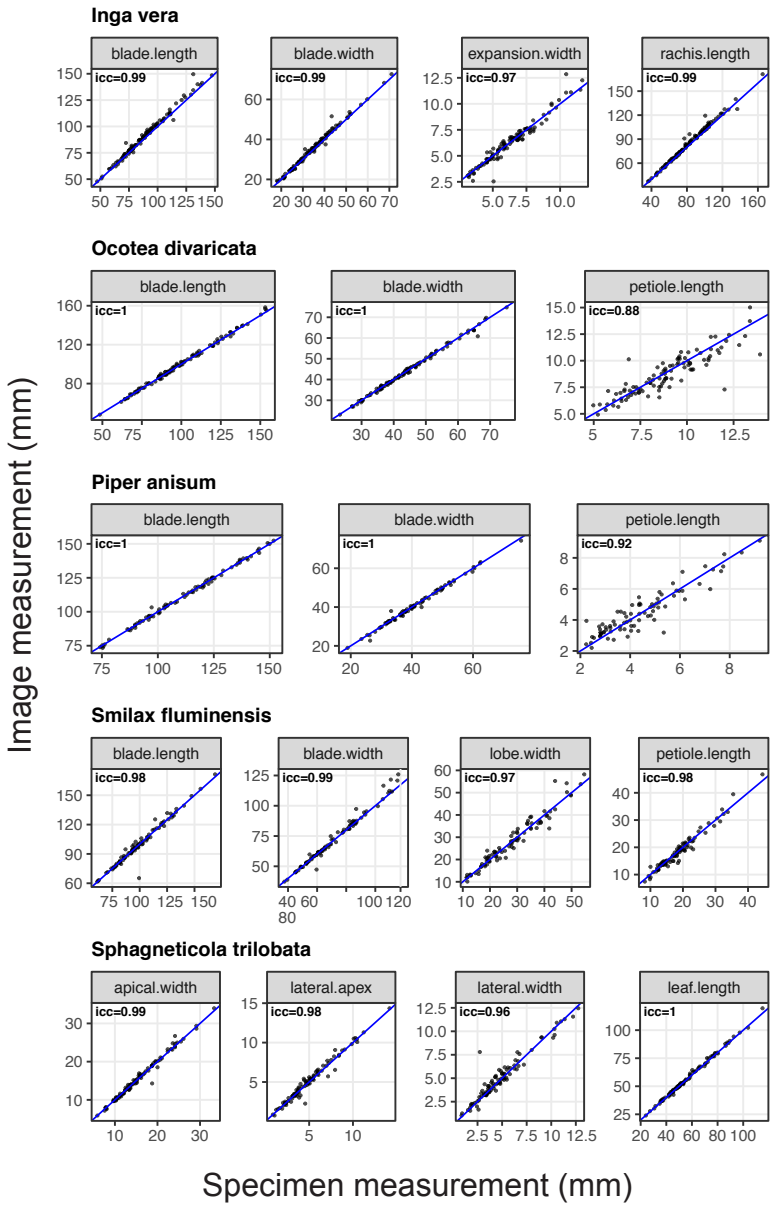

Figure 3: Scatter plots with measurements made in specimen versus measurements made in the images on the `measurement dataset`. Blue lines indicate the line $y = x$, in which both measurements are the same. Each panel also shows the Pearson's correlation coefficient between the measurements.

Table 1: Proportion of missing data in each dataset according to three criteria: total percentage of missing data (total); percentage of variables with at least 50% of missing observations (at least 50%); percentage of variables impossible to score from images (100%).

Dataset	total %	at least 50%	100%
A2013	50%	56.25%	0%
A2008	44.39%	45.12%	19.51%
A2009	69.17%	100%	8.33%
B2018	37.5%	35.71%	35.71%
B2015	19.44%	8.33%	0%
E2015	0%	0%	0%
K2006i	30.72%	41.18%	5.88%
K2006p	30.72%	41.18%	5.88%
M1992	10%	10%	10%
P2005	42.97%	43.75%	12.5%
R2014	60.61%	63.64%	45.45%
S2012	43.56%	44%	24%
T2008	0%	0%	0%

4 Discussion

Here we used a set of leaf measurements and legacy morphometric studies to evaluate the precision and accuracy of morphological data gathered from images of herbarium specimens. We found image measurements to be highly precise and accurate, even though there is a small drop in accuracy for smaller traits. On the other hand, it is not so accurate to compile morphological datasets from images, particularly due to problems related to missing observations. Below we discuss the problems, advantages, and consequences of collecting morphological data from digitized plant specimens.

4.1 Measurement precision and accuracy

Images of herbarium specimens are a good source of plant organ measurements. Measurements taken from images are as precise as the ones made on actual specimens. In fact, image measurements are in the confidence limits of being more precise (Fig. 2). Such an increase in precision has been seen for other imaging systems, as micro-computed tomography (Simon & Marroig 2015). Thus, as we have focused only on leaves of particular species, it is possible analyses including a wider diversity of morphological variables to find the same for images of herbarium specimens.

Image measurements are also accurate, as seen by their high similarity to specimen measurements (Fig. 3) and overall lack of bias (Fig. 4). Agreement between data gathered from images and specimens was also seen for extraction of morphological features using artificial intelligence, which compared measurements taken by computers and researchers (Corney, Clark, Tang & Wilkin 2012). Similar results for other organisms and different imaging methods (Simon & Marroig 2015, Aldridge et al. 2005), indicate that high-confidence measurements can be obtained from digitized specimens. Nonetheless, we have also evidenced particular issues in the accuracy of image-based measurements and overall data acquisi-

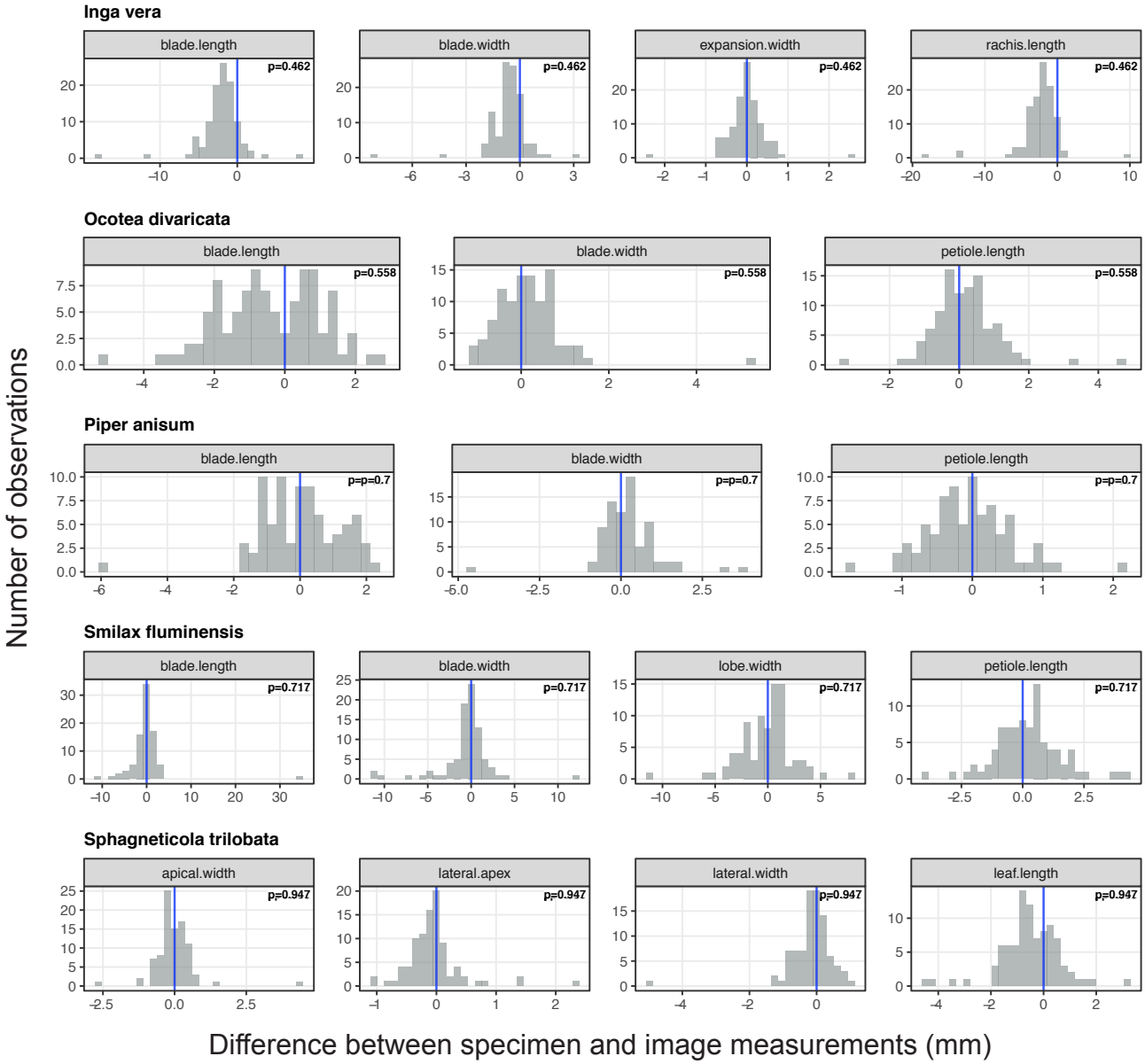

Figure 4: Histogram of the difference between measurements made in specimen and in the images on the `measurement dataset`. Vertical lines indicate the line $x = 0$, in which the difference between the measurements is zero. Each panel also shows the p-values for testing the hypothesis that the median difference between image and specimen measurements is zero.

Table 2: Percentage of missing data by type of variable for each data set used to build the masks with missing data pattern.

Dataset	Continuous (mm)	Categorical	Discrete	Ratio
A2013	50%	-	-	-
A2008	39.39%	53.33%	52.73%	37.65%
A2009	70%	-	60%	-
B2018	50%	16.67%	0%	-
B2015	19.44%	-	-	-
E2015	0%	-	-	-
K2006i	46.67%	-	13.89%	0%
K2006p	46.67%	-	13.89%	0%
M1992	20%	0%	0%	-
P2005	32.81%	56.25%	43.75%	-
R2014	51.04%	-	58.33%	100%
S2012	41.11%	56.94%	31.75%	-
T2008	0%	-	0%	-

Table 3: Spearman correlation coefficients and their respective p-values for comparing the relationship between the accuracy in the estimation of each principal component/the PCA similarity (according to the "M" scheme) and the percentage of missing observations. total% - total percentage of missing data in the dataset; at least 50% - variables with at least 50% of missing observations; 100% - variables impossible to score from images.

PCA Component	total %	at least 50%	100%
1	-0.45 (0.12)	-0.48 (0.10)	-0.15 (0.63)
2	-0.78 (<0.01)	-0.79 (<0.01)	-0.51 (0.08)
3	-0.65 (0.02)	-0.66 (0.01)	-0.41 (0.16)
4	-0.67 (0.01)	-0.68 (0.01)	-0.30 (0.32)
PCA Sim	-0.73 (<0.01)	-0.73 (<0.01)	-0.78 (<0.01)

tion.

4.2 Noise

The first problem with image use is a drop in measurement accuracy as variables scale down (Fig. 4). Measurements of smaller features, particularly those below 30–40 mm (Fig. 5), differ more between images and specimens than do larger structures. This association between accuracy error and scale was to be expected, as measurement precision lowers for smaller traits (Fig. 2). Such a decrease in precision is not limited to images and is seen in our specimen measurements (Fig. 2), as well as for bird bones (Yezerinac et al. 1992) and *Drosera* (Droseraceae) leaves (Hoyo & Tsuyuzaki 2013), for example. In this context and also considering accuracy tests rely on subjective selection of reference values, our results do not mean images are worse measurement sources. However, while we can increase precision and accuracy by studying specimens with different magnification tools, zooming images is restricted by resolution. Image resolution is not a major issue for some databases, as the ones in JStor Global Plants

(https://plants.jstor.org/). Nonetheless, as resolution may vary between different imaging methods (Tulig et al. 2012, Takano et al. 2019, Sweeney et al. 2018), databases will differ in how they are prone to measurement errors.

Even though errors will always be present, their effect over multivariate analysis is not strong. Analyses of synthetic data including noise accurately estimate the first two components (Fig. 6). The low correlation values for the third and fourth components are likely due to interaction between measurement error and inherent decreasing explanatory power for each component (Hair et al. 1998). Moreover, as the first and second PCA axes usually concentrate most part of data variation in biological morphometric studies, slightly poorer estimation of remaining components likely will not have big impact on the results of imaged-based analyses. This lack of impact is clearly seen in high values of PCA similarity for presence of noise (Fig. 6). High similiarity between PCAs indicates that datasets compiled from images and from specimens are essentialy the same (Yang & Shahabi 2004, Melo et al. 2015). It is possible our results could be biased, as most variables in our legacy datasets fall outside the range of sizes we found to be problematic. Thus, researchers working on smaller organs or with small plants should bear this potential bias in mind. Nonetheless, as seen for other studies of leaf image measurements (Corney, Tang, Clark, Hu & Jin 2012, Corney, Clark, Tang & Wilkin 2012), our results show that digitized herbarium specimens are good sources of plant metric data. Of course this will only hold true for traits that can be observed in the images.

4.3 Missing data

Another problem in our synthetic databases was the pervasive presence of missing data (Table 1), particularly of continuous variables (Table 2). As discussed above, these absences are likely due to image resolution issues, which hamper observation of small organs. Nonetheless, the extent of this problem is particular to each dataset. For example, even though they focused on larger organs, the only two datasets with no missing data (T2008 and E2015; 1) also include small scale variables (less than 30 mm). At the same time, larger organs that were concealed during specimen preparation or due to their overlapping nature (e.g. petals usually hide both androecium and gynoecium) can not be observed and, thus, bias data collection. Feature-related bias is also seen in automatic trait recognition in which organ identity affects accuracy (Younis et al. 2018). Similar examples come from our datasets. If Trovó et al. 2008 had investigated the morphology of flowers (usually smaller than 2mm in Eriocaulaceae), it would not be possible to reproduce their data solely with images. This is the case with the T2013 dataset (Alcantara et al. 2013), in which the main source of missing data are organs concealed within the Bignoniaceae tubular flowers. Even though we can manipulate specimens to expose such hidden features, the same can not be done with herbarium images.

Indeed, missing data lower PC correlation and PCA similarity, as seen in Fig. 6. This trend is reinforced by the negative correlation between proportion of missing data `synthetic dataset` and PCA similarity or PC estimation (Table 3). Accuracy drops are stronger for the second, and especially for the third and fourth PCs. As discussed above, poor estimation of third and fourth PCs may not be a

big problem. However, inaccuracy in estimation of the first two components and of overall multivarite space, expressed in PCA similarity, indicates significant differences between `original` and `synthetic datasets`. These differences could bias biological interpretations (Yezerinac et al. 1992) of studies relying on acquisition of morphological data, such as phenotypic evolution, disparity, modularity, phylogenetic inference, and others (Wiens 2003, Zaragüeta-Bagils & Bourdon 2007, Catalano et al. 2010, Klingenberg & Gidaszewski 2010, Vasconcelos et al. 2018, Castiglione et al. 2019, Dellinger et al. 2019, Gallaher et al. 2019, Guillerme & Cooper 2018). Our approach shows that missing data are particularly problematic for cases in which phenotypic variation is modeled with multivariate analyses, an increasingly common method in morphological studies.

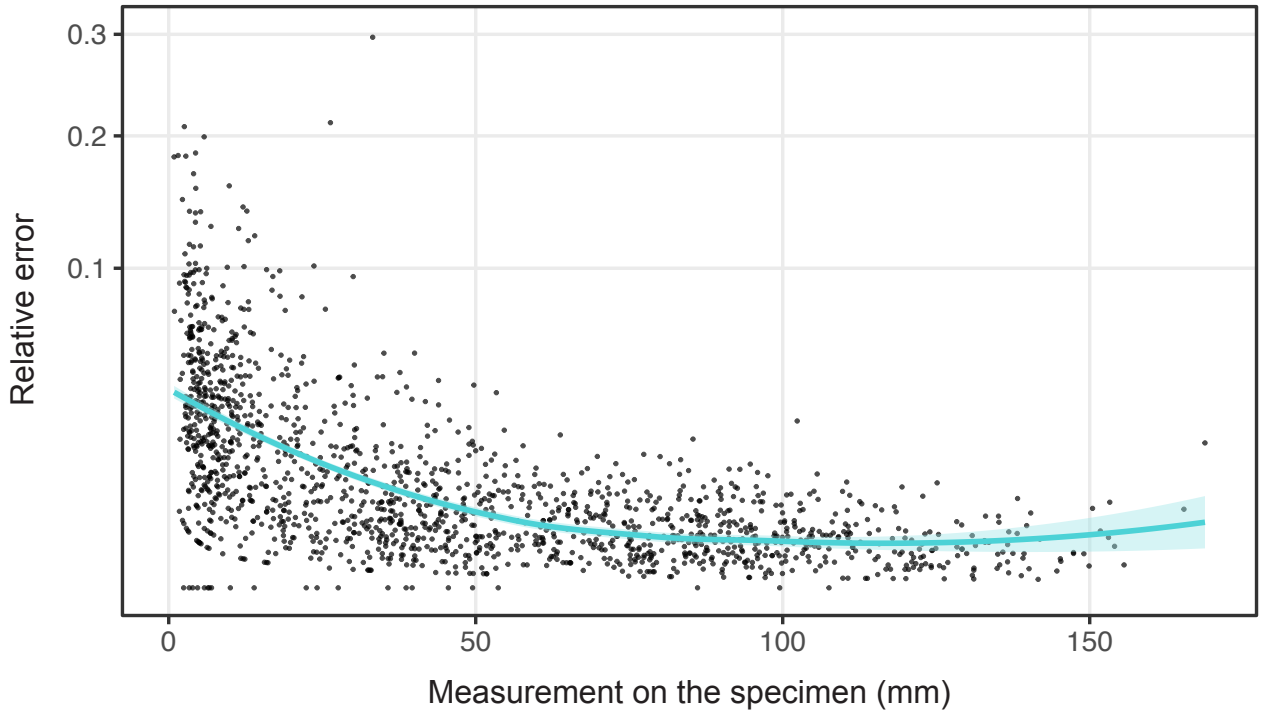

Figure 5: Relationship between the relative error of the measurements (i.e., the absolute difference between image and specimen measurements divided by specimen measurement) and the magnitude of each variable. Each dot represents a different observation from a variable in the `measurement dataset`. The curve indicates the mean trend, along with its standard error. Y-axis is displayed in a square root scale.

4.4 Joint effect of measurement noise and missing data

Above, we analysed the individual effects of measurement noise and missing data over PCA accuracy. However, most real datasets likely will include both issues. Our results indicate they interact, further

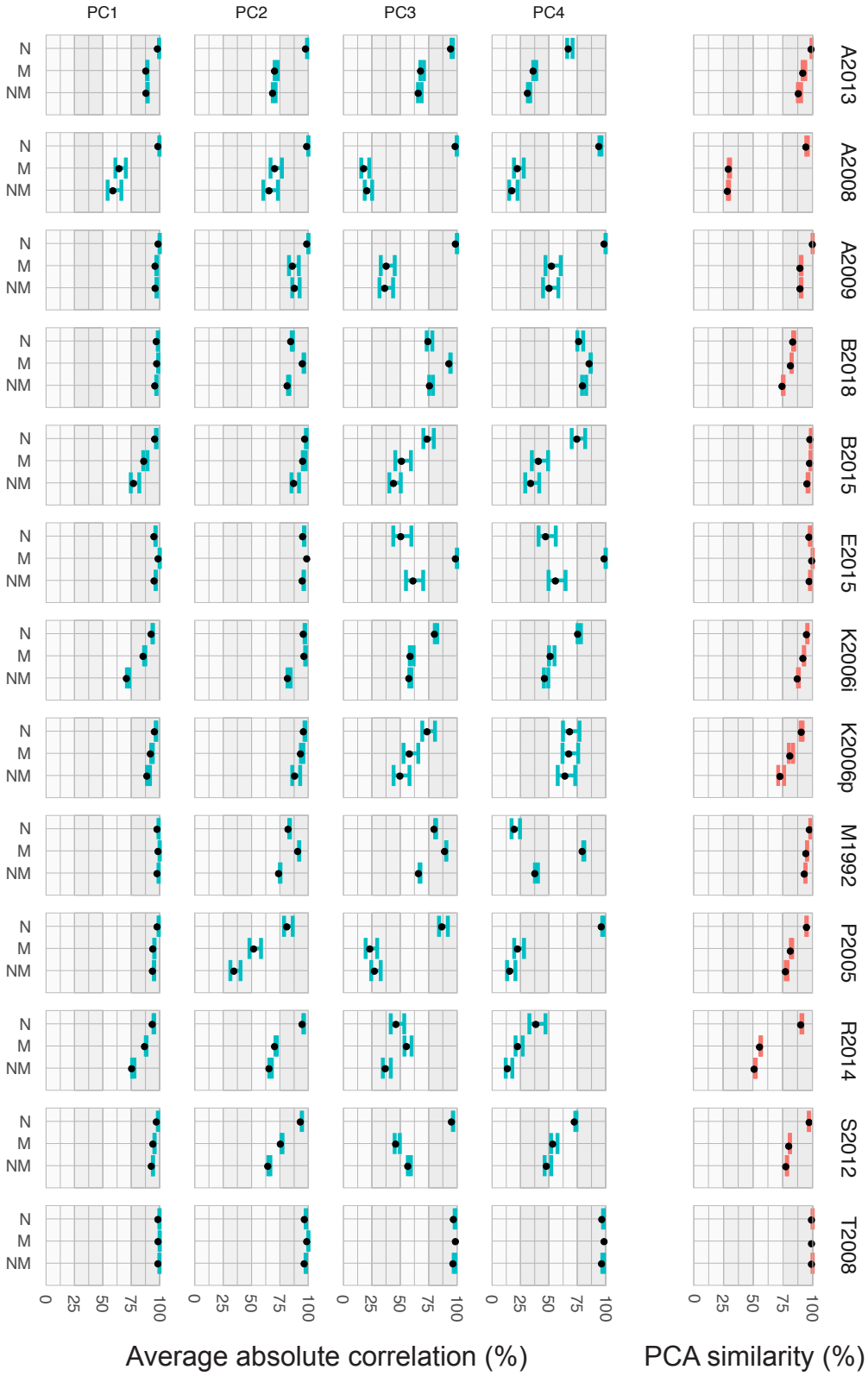

Figure 6: (Left) Accuracy (and 95% confidence intervals) in estimating the first four principal components using images (`synthetic datasets`). (Right) PCA Similarity between `synthetic datasets` versus `original`. 100% values equal maximum accuracy for both analyses. N - `synthetic dataset` with noise on the same level as those observed in Figure 3. M - `synthetic dataset` without variables that can not be acquired from images. NM - `synthetic dataset` with both noise and missing data.

dropping PC correlation and PCA similatiry (Fig. 6). Nonetheless, this impact over analysis tends to be closer to values seen in missing data alone, which is commonly the strongest source of PC estimation inaccuracy in our analyses. For the few cases in which noise is stronger than missing data (e.g. third and/or fourth PCs in B2018, and M1991), the joint effect may improve component estimation in synthetic analyses. Probably this happens because we removed noise when simulating missing data, improving PCA estimation for these particular cases.

Apart from those exceptions, when both noise and missing data are present, accuracy of third and fourth PCs is very low, usually less than 50%. More important, for most datasets values for the first component are around 85%, and fall below 75% for the second. It is hard to make general assumptions on how such levels of accuracy affect biological interpretations, as studies vary on their goals and analysis resilience. However, if the study system has low internal variation, common in comparisons of closely related taxa, rates of false positive and false negative conclusions may increase (Yezerinac et al. 1992), particularly if the second, third and fourth components are taken into account.

Although joint effect of noise and missing data may be serious, its influence will vary as datasets differ in the amount of missing data and presence of noise, which are linked to traits of interest size, image resolution, and organ overlap. Hence, before relying solely on morphological data compiled from digitized herbarium specimens, researchers should evaluate with trial tests how image use affects their particular studies. If changes in analyses' results are detected, it is important to investigate how they affect biological interpretations and then decide if and how to source morphological data from specimen images.

In summary, researchers sourcing data from images have to be aware that missing observations may compromise their study. Despite this issue, image-based data collection is overall accurate. Our results show that both PCA similarity and PC correlation values for the first and second PC are higher than 75% for most datasets 6.

4.5 Consequences for image-based morphological data

When compiling morphological data from images of herbarium specimens, measurement noise is not a problem, particularly if variables are outside the size range affected by resolution. This problematic range usually includes organs or structures that are commonly studied with magnification tools. Because of that, we expect studies focusing on such structures would not consider measuring digitized specimens. Hence, even though resolution may vary among digitization methods (Tulig et al. 2012, Takano et al. 2019, Sweeney et al. 2018), measuring them instead of specimens is an effective, less expensive (no need for traveling or loans), and, more important, reliable method for data acquisition.

On the other hand, missing data has higher impact over data collection. This impact, however, will vary between different plant groups, particular morphologies, and specific research questions. When missing data is low (and noise problems do not exist; see above), datasets compiled solely from images are trustworthy. For the cases in which missing data quantity is relevant, researchers can improve data quality with particular strategies, used individually or combined with each other: 1) complete data with

imputation methods; 2) exclude problematic variables or observations; 3) use specimens to fill missing data.

Data imputation methods are a good option to estimate missing values in a dataset. These methods usually estimate missing values based on variance of the variable at hand, or on its associations to other variables in the dataset (Frane 1976, Rubin 1976, Little & Rubin 2019). Here, for example, our PCA analyses used a relatively simple method (Roweis 1998). Even if this method is appropriate for some cases, it is important to consider particularities of each dataset and assumptions made by different imputation methods in order to improve performance (Little & Rubin 2019). Although not free of issues, data estimation methods are powerful tools to improve results without compromising biological meaning (König et al. 2019).

Data imputation, however, will be inaccurate in cases where almost all observations of a given variable are missing. In this situation, it is sometimes better to exclude the variables from the analysis, particularly if it is highly correlated to at least another variable (Hemel et al. 1987). This way, variable exclusion is less likely to impact analyses, or at least will have lower impact, as seen for the smaller negative correlation of fully absent variables to PC estimation and PCA similarity (Table 3). On the other hand, depending on the reason why a feature is missing, variable exclusion may bias results (Little & Rubin 2019). Thus, it is important to understand the performance of data imputation and data exclusion when dealing with missing observations (Musil et al. 2002, Saunders et al. 2006, Horton & Kleinman 2007, Enders 2008).

If variable exclusion and data imputation are not feasible for a given case, such as complete lack of observations on an important variable, researchers can use specimens to fill gaps in the dataset. Although this is not as practical as the other options, it can at least speed up work inside museum collections. If access to specimens is limited, an interesting option would be to combine examination of a minimum number of specimens and data imputation. This approach exemplifies how there is no single solution to improve an image-based dataset. As images reflect the taxonomic diversity of plant morphology, problems and their solutions will be particular to each case.

Despite particularities discussed above, herbaria digitization opened new frontiers for image-based phenotyping. Up to now, the bulk of research using plant images focused on species identification (e.g. MacLeod & Steart 2015, Unger et al. 2016, Remagnino et al. 2016, Carranza-Rojas et al. 2017, Cope et al. 2012), or on shape analysis of particular organs, as leaves (e.g. Corney, Clark, Tang & Wilkin 2012, Klingenberg et al. 2012), or floral structures (e.g. Smith & Kriebel 2018, Reginato & Michelangeli 2016). For such studies on particular organs, datasets usually included only images of the focal traits. Now, the millions of images of herbarium specimens available make a wider diversity of phenotypes freely accessible. Access to these images by anyone with an internet connection, by its turn, supports data collection with crowdsourcing. These methods are fast, cheap, and able to deliver high-quality data (Chang & Alfaro 2016, Zhou et al. 2018, O'Leary et al. 2018). Crowdsourced datasets may also support improvement of automatic extraction methods (Zhou et al. 2018, Burleigh et al. 2013), or be used in parallel with them to improve data collection. For instance, automatic tools still can not clearly detect organs if they are

overlapped (Gehan & Kellogg 2017), which is common in herbarium specimens. On the other hand, overlapping may not be a problem for humans, who are able to identify and bypass the issue. Independently of particular data collection strategies and of the need to be careful with missing observations, we reinforced here that digitized herbarium specimens are good sources of morphological data.

4.6 A note on data sharing

One of the main advantages of collecting data from digitized herbarium specimens is their free availability in digital repositories. It is only fair that sharing extends to collected data as well. However, at least for morphometric studies, this is not frequently done. For example, our `original dataset` came from just a few authors, who shared raw observations at publication time, or kindly replied to our data requests. A much larger number of studies, however, presented only tables with mean and standard deviation values, which do not allow proper verification of data treatments.

If we aim to build comprehensive datasets, we have to shift this trend and increase use of tools such as MorphoBank (`https://morphobank.org`; O'Leary & Kaufman 2011, 2012) to share morphological data. Data sharing improves scientific work, and also reduces unnecessary recollection of the same data by different people, particularly if data is strongly integrated with ontologies (Deans et al. 2015) and linked to vouchers and scientific names in curated databases (Bruneau et al. 2019).

5 Conclusion

Relying on images of herbarium specimens for data collection is effective, as demonstrated by high correlation between measurement in specimens and in images, even though image resolution introduces measurement errors on small scales. At the same time, as it is impossible to manipulate images, particular features can not be observed, and will be missing from image-based datasets. Even though there are a number of options to reduce the effect of missing data on data analyses, such as data imputation, the impact of this more serious problem has to be evaluated for each case.

Overall, we showed that digitized specimens are a trustworthy source of morphological data, which can be collectively gathered, and support a wide array of research questions on plant diversity and evolution. Particularly, despite the low number for other groups, as animals (Chang & Alfaro 2016), millions of specimen images (plus data on their distribution, ecology, and phenology) are already available for plants (Le Bras et al. 2017, Soltis 2017), unleashing the potential for morphological data acquisition (Soltis 2017, Willis et al. 2017). This makes Botany already able to merge big-data on geographical distribution, genomes, and morphology in order to unveil new knowledge on patterns and processes of phenotypic diversity and evolution.

Acknowledgments

We are grateful to Abubakar Bello, Axel Poulsen and Inger Nordal, David Spooner and Mercedes Ames, Jeffrey Rose, Karol Marhold, Marcelo Trovó, Marek Slovák, Mariana Bünger, M. Montserrat Martínez-Ortega, Suzana Alcantara, and Vanessa Staggemeier for sharing their datasets, either at publication time or directly to us. We also thank Sara Mortara for initial help on research design and data collection, as well as Rafaela C. Forzza, Paula Leitman, Fabiana L. R. Filard and the RB staff for support during work at the herbarium. Finally, Anna Penna, Diogo Melo, Matheus F. Santos, and Monique N. Simon's bright comments and discussions on methods helped us improve early versions of this work. RI was supported in part by FAPESP (2019/11321-9) and CNPq (306943/2017-4). Authors declare no conflict of interest.

Data accessibility statement

All data and code used here are available from [GitHub link to be made available after review]. Leaf measurements are also individually available from MorphoBank [project to be made available after review].

References

Alcantara, S., de Oliveira, F. & Lohmann, L. (2013), 'Phenotypic integration in flowers of neotropical lianas: diversification of form with stasis of underlying patterns', Journal of Evolutionary Biology **26**(10), 2283–2296. 4, 11

Aldridge, K., Boyadjiev, S. A., Capone, G. T., DeLeon, V. B. & Richtsmeier, J. T. (2005), 'Precision and error of three-dimensional phenotypic measures acquired from 3dMD photogrammetric images', American Journal of Medical Genetics Part A **138**(3), 247–253. 8

Ames, M., Salas, A. & Spooner, D. M. (2008), 'A morphometric study of species boundaries of the wild potato *Solanum* series *Piurana* (Solanaceae) and putatively related species from seven other series in *Solanum* sect. *Petota*', Systematic Botany **33**(3), 566–578. 4

Andres-Sanchez, S., Rico, E., Herrero, A., Santos-Vicente, M. & MartINez-Ortega, M. M. (2009), 'Combining traditional morphometrics and molecular markers in cryptic taxa: towards an updated integrative taxonomic treatment for *Veronica* subgenus *Pentasepalae* (Plantaginaceae sensu APG II) in the western Mediterranean', Botanical Journal of the Linnean Society **159**(1), 68–87. 4

Babin-Fenske, J., Anand, M. & Alarie, Y. (2008), 'Rapid morphological change in stream beetle museum specimens correlates with climate change', Ecological Entomology **33**(5), 646–651. 2

Bartko, J. J. (1966), 'The intraclass correlation coefficient as a measure of reliability', Psychological reports **19**(1), 3–11. 3, 4

Bello, A., Stirton, C., Chimphango, S. & Muasya, A. (2018), 'Morphological evidence for introgressive hybridization in the genus *Psoralea* L. (Psoraleeae, Fabaceae)', South African Journal of Botany **118**, 321–328. 4

Bruneau, A., Borges, L. M., Allkin, R., Egan, A. N., Estrella, M. d. l., Javadi, F., Klitgård, B., Miller, J. T., Murphy, D. J., Sinou, C., Vatanparast, M. & Zhang, R. (2019), 'Towards a new online species-information system for legumes', Australian Systematic Botany **32**(6), 495–518. 16

Bünger, M. d. O., Einsehlor, P., Figueiredo, M. L. N. & Stehmann, J. R. (2015), 'Resolving Species Delimitations in the *Eugenia involucrata* Group (*Eugenia* sect. *Phyllocalyx* - Myrtaceae) with Morphometric Analysis', Systematic Botany **40**(4), 995–1002. 4

Burleigh, J. G., Alphonse, K., Alverson, A. J., Bik, H. M., Blank, C., Cirranello, A. L., Cui, H., Daly, M., Dietterich, T. G., Gasparich, G. et al. (2013), 'Next-generation phenomics for the Tree of Life', PLoS Currents **5**. 2, 15

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E. & Joly, A. (2017), 'Going deeper in the automated identification of Herbarium specimens', BMC Evolutionary Biology **17**(1), 181. 2, 15

Castiglione, S., Serio, C., Tamagnini, D., Melchionna, M., Mondanaro, A., Di Febbraro, M., Profico, A., Piras, P., Barattolo, F. & Raia, P. (2019), 'A new, fast method to search for morphological convergence with shape data', PloS one **14**(12). 12

Catalano, S. A., Goloboff, P. A. & Giannini, N. P. (2010), 'Phylogenetic morphometrics (I): the use of landmark data in a phylogenetic framework', Cladistics **26**(5), 539–549. 12

Chang, J. & Alfaro, M. E. (2016), 'Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data', Methods in Ecology and Evolution **7**(4), 472—-482. 15, 16

Cope, J. S., Corney, D., Clark, J. Y., Remagnino, P. & Wilkin, P. (2012), 'Plant species identification using digital morphometrics: A review', Expert Systems with Applications **39**(8), 7562–7573. 15

Corney, D., Clark, J. Y., Tang, H. L. & Wilkin, P. (2012), 'Automatic extraction of leaf characters from herbarium specimens', Taxon **61**(1), 231–244. 2, 8, 11, 15

Corney, D. P., Tang, H. L., Clark, J. Y., Hu, Y. & Jin, J. (2012), 'Automating digital leaf measurement: the tooth, the whole tooth, and nothing but the tooth', PloS one **7**(8), e42112. 2, 11

Davis, C. C., Willis, C. G., Connolly, B., Kelly, C. & Ellison, A. M. (2015), 'Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms', American Journal of Botany **102**(10), 1599–1609. 2

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B. et al. (2015), 'Finding our way through phenotypes', PLoS Biology **13**(1). 16

Dellinger, A. S., Artuso, S., Pamperl, S., Michelangeli, F. A., Penneys, D. S., Fernández-Fernández, D. M., Alvear, M., Almeda, F., Armbruster, W. S., Staeder, Y. et al. (2019), 'Modularity increases rate of floral evolution and adaptive success for functionally specialized pollination systems', Communications Biology **2**(1), 1–11. 12

Egan, A. N. (2015), 'Species delimitation and recognition in the *Pediomelum megalanthum* complex (Fabaceae) via multivariate morphometrics', PhytoKeys **44**, 65. 4

Enders, C. K. (2008), 'A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models', Structural Equation Modeling: A Multidisciplinary Journal **15**(3), 434–448. 15

Frane, J. W. (1976), 'Some simple procedures for handling missing data in multivariate analysis', Psychometrika **41**(3), 409–415. 15

Gallaher, T. J., Adams, D. C., Attigala, L., Burke, S. V., Craine, J. M., Duvall, M. R., Klahs, P. C., Sherratt, E., Wysocki, W. P. & Clark, L. G. (2019), 'Leaf shape and size track habitat transitions across forest–grassland boundaries in the grass family (Poaceae)', Evolution **73**(5), 927–946. 12

Gehan, M. A., Fahlgren, N., Abbasi, A., Berry, J. C., Callen, S. T., Chavez, L., Doust, A. N., Feldman, M. J., Gilbert, K. B., Hodge, J. G., Hoyer, J. S., Lin, A., Liu, S., Lizárraga, C., Lorence, A., Miller, M., Platon, E., Tessman, M. & Sax, T. (2017), 'PlantCV v2: Image analysis software for high-throughput plant phenotyping', PeerJ **5**, e4088. 2

Gehan, M. A. & Kellogg, E. A. (2017), 'High-throughput phenotyping', American Journal of Botany **104**(4), 505–508. 16

Guillerme, T. & Cooper, N. (2018), 'Time for a rethink: time sub-sampling methods in disparity-through-time analyses', Palaeontology **61**(4), 481–493. 12

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. et al. (1998), Multivariate data analysis, Prentice hall Upper Saddle River, NJ. 11

Hemel, J. B., van der Voet, H., Hindriks, F. R. & van der Slik, W. (1987), 'Stepwise deletion: a technique for missing-data handling in multivariate analysis', Analytica Chimica Acta **193**, 255–268. 15

Horton, N. J. & Kleinman, K. P. (2007), 'Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models', The American Statistician **61**(1), 79–90. 15

Hoyo, Y. & Tsuyuzaki, S. (2013), 'Characteristics of leaf shapes among two parental Drosera species and a hybrid examined by canonical discriminant analysis and a hierarchical Bayesian model', American Journal of Botany **100**(5), 817–823. 10

Klingenberg, C., Duttke, S., Whelan, S. & Kim, M. (2012), 'Developmental plasticity, morphological variation and evolvability: a multilevel analysis of morphometric integration in the shape of compound leaves', Journal of Evolutionary Biology **25**(1), 115–129. 15

Klingenberg, C. P. & Gidaszewski, N. A. (2010), 'Testing and quantifying phylogenetic signals and homoplasy in morphometric data', Systematic Biology **59**(3), 245–261. 12

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J. & Kreft, H. (2019), 'Biodiversity data integration – The significance of data resolution and domain', PLoS biology **17**(3), e3000183. 15

Kučera, J., Lihova, J. & Marhold, K. (2006), 'Taxonomy and phylogeography of *Cardamine impatiens* and *C. pectinata* (Brassicaceae)', Botanical Journal of the Linnean Society **152**(2), 169–195. 4

Laing, A. M., Doyle, S., Gold, M. E. L., Nesbitt, S. J., O'Leary, M. A., Turner, A. H., Wilberg, E. W. & Poole, K. E. (2018), 'Giant taxon-character matrices: the future of morphological systematics', Cladistics **34**(3), 333–335. 2

Law, W. & Salick, J. (2005), 'Human-induced dwarfing of Himalayan snow lotus, *Saussurea laniceps* (Asteraceae)', Proceedings of the National Academy of Sciences **102**(29), 10218–10220. 2

Le Bras, G., Pignal, M., Jeanson, M. L., Muller, S., Aupic, C., Carré, B., Flament, G., Gaudeul, M., Gonçalves, C., Invernón, V. R., Jabbour, F., Lerat, E., Lowry, P. P., Offroy, B., Pimparé, E. P., Poncy, O., Rouhan, G. & Haevermans, T. (2017), 'The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset', Scientific data **4**, 170016. 2, 16

Little, R. J. & Rubin, D. B. (2019), Statistical analysis with missing data, Vol. 793, John Wiley & Sons. 15

MacLeod, N. & Steart, D. (2015), 'Automated leaf physiognomic character identification from digital images', Paleobiology **41**(4), 528–553. 15

Marhold, K. (1992), 'A multivariate morphometric study of the *Cardamine amara* group (Cruciferae) in the Carpathian and Sudeten mountains', Botanical Journal of the Linnean Society **110**(2), 121–135. 4

Melo, D., Garcia, G., Hubbe, A., Assis, A. P. & Marroig, G. (2015), 'EvolQG - An R package for evolutionary quantitative genetics', F1000Research **4**. 5, 11

Musil, C. M., Warner, C. B., Yobas, P. K. & Jones, S. L. (2002), 'A comparison of imputation techniques for handling missing data', Western Journal of Nursing Research **24**(7), 815–829. 15

O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z.-X., Meng, J., Ni, X., Novacek, M., Perini, F., Randall, Z., Rougier, G., Sargis, E., Silcox, M., Simmons, N., Spaulding, M., Velazco, P., Weksler, M., Wible, J. & Cirranello, A. (2013), 'The placental mammal ancestor and the post–K-Pg radiation of placentals', Science **339**(6120), 662–667. 2

O'Leary, M. A., Alphonse, K., Mariangeles, A. H., Cavaliere, D., Cirranello, A., Dietterich, T. G., Julius, M., Kaufman, S., Law, E., Passarotti, M., Reft, A., Robalino, J., Simmons, N. B., Smith, S. Y., Stevenson, D. W., Theriot, E., Velazco, P. M., Walls, R. L., Yu, M. & Daly, M. (2018), 'Crowds replicate performance of scientific experts scoring phylogenetic matrices of phenotypes', Systematic Biology **67**(1), 49–60. 15

O'Leary, M. A. & Kaufman, S. G. (2011), 'MorphoBank: phylophenomics in the "cloud"', Cladistics **27**(5), 529–537. 16

O'Leary, M. A. & Kaufman, S. G. (2012), 'MorphoBank 3.0: Web application for morphological phylogenetics and taxonomy', See http://www. morphobank. org . 16

Pearson, K. (1920), 'Notes on the history of correlation', Biometrika **13**(1), 25–45. 5

Poulsen, A. D. & Nordal, I. (2005), 'A phenetic analysis and revision of Guineo-Congolean rain forest taxa of *Chlorophytum* (Anthericaceae)', Botanical Journal of the Linnean Society **148**(1), 1–20. 4

Reginato, M. & Michelangeli, F. A. (2016), 'Diversity and constraints in the floral morphological evolution of *Leandra* s. str. (Melastomataceae)', Annals of Botany **118**(3), 445–458. 15

Remagnino, P., Mayo, S., Wilkin, P., Cope, J. & Kirkup, D. (2016), Computational botany: methods for automated species identification, Springer. 2, 15

Rose, J. P. & Freudenstein, J. V. (2014), 'Cryptic and overlooked: species delimitation in the mycoheterotrophic *Monotropsis* (Ericaceae: Monotropoideae)', Systematic Botany **39**(2), 578–593. 4

Roweis, S. T. (1998), EM algorithms for PCA and SPCA, in M. I. Jordan, M. J. Kearns & S. A. Solla, eds, 'Advances in neural information processing systems', pp. 626–632. 5, 15

Rubin, D. B. (1976), 'Inference and missing data', Biometrika **63**(3), 581–592. 15

Saunders, J. A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E. K. & Pescarino, R. (2006), 'Imputing missing data: A comparison of methods for social work researchers', Social Work Research **30**(1), 19–31. 15

Simon, M. N. & Marroig, G. (2015), 'Landmark precision and reliability and accuracy of linear distances estimated by using 3D computed micro-tomography and the open-source TINA Manual Landmarking Tool software', Frontiers in Zoology **12**(1), 12. 8

Slovák, M., Kučera, J., Marhold, K. & Zozomová-Lihová, J. (2012), 'The morphological and genetic variation in the polymorphic species *Picris hieracioides* (Compositae, Lactuceae) in Europe strongly contrasts with traditional taxonomical concepts', Systematic Botany **37**(1), 258–278. 4

Smith, S. D. & Kriebel, R. (2018), 'Convergent evolution of floral shape tied to pollinator shifts in *Iochrominae* (Solanaceae)', Evolution **72**(3), 688–697. 15

Soltis, P. S. (2017), 'Digitization of herbaria enables novel research', American Journal of Botany **104**(9), 1281–1284. 2, 16

Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. (2007), 'pcamethods – a bioconductor package providing pca methods for incomplete data', Bioinformatics **23**, 1164–1167. 5

Streiner, D. L. & Norman, G. R. (2006), '"Precision" and "accuracy": two terms that are neither', Journal of Clinical Epidemiology **59**(4), 327–330. 3

Suarez, A. V. & Tsutsui, N. D. (2004), 'The value of museum collections for research and society', AIBS Bulletin **54**(1), 66–74. 2

Sweeney, P. W., Starly, B., Morris, P. J., Xu, Y., Jones, A., Radhakrishnan, S., Grassa, C. J. & Davis, C. C. (2018), 'Large–scale digitization of herbarium specimens: Development and usage of an automated, high–throughput conveyor system', Taxon **67**(1), 165–178. 2, 11, 14

Takano, A., Horiuchi, Y., Fujimoto, Y., Aoki, K., Mitsuhashi, H. & Takahashi, A. (2019), 'Simple but long-lasting: A specimen imaging method applicable for small-and medium-sized herbaria', PhytoKeys **118**, 1–14. 2, 11, 14

Thiers, B. (2020, continuously updated), 'Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium', http://sweetgum. nybg. org/ih . 3

Trovó, M., Sano, P. & Winkworth, R. (2008), 'Morphology and environment: geographic distribution, ecological disjunction, and morphological variation in *Actinocephalus polyanthus* (Bong.) Sano (Eriocaulaceae)', Feddes Repertorium **119**(7-8), 634–643. 4, 11

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, A. & Thiers, B. M. (2012), 'Increasing the efficiency of digitization workflows for herbarium specimens', ZooKeys **209**, 103–113. 2, 11, 14

Unger, J., Merhof, D. & Renner, S. (2016), 'Computer vision applied to herbarium specimens of German trees: testing the future utility of the millions of herbarium specimen images for automated identification', BMC Evolutionary Biology **16**(1), 248. 15

Vasconcelos, T. N., Chartier, M., Prenner, G., Martins, A. C., Schönenberger, J., Wingler, A. & Lucas, E. (2018), 'Floral uniformity through evolutionary time in a species-rich tree lineage', New Phytologist . 12

Wiens, J. J. (2003), 'Missing data, incomplete taxa, and phylogenetic accuracy', Systematic Biology **52**(4), 528–538. 12

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', Biometrics Bulletin **1**(6), 80–83. 4

Willis, C. G., Ellwood, E. R., Primack, R. B., Davis, C. C., Pearson, K. D., Gallinat, A. S., Yost, J. M., Nelson, G., Mazer, S. J., Rossington, N. L. et al. (2017), 'Old plants, new tricks: Phenological research using herbarium specimens', Trends in Ecology & Evolution **32**(7), 531–546. 16

Yang, K. & Shahabi, C. (2004), A PCA-based similarity measure for multivariate time series, in 'Proceedings of the 2nd ACM International Workshop on Multimedia Databases', pp. 65–74. 5, 11

Yezerinac, S. M., Lougheed, S. C. & Handford, P. (1992), 'Measurement error and morphometric studies: statistical power and observer experience', Systematic Biology **41**(4), 471–482. 10, 12, 14

Younis, S., Weiland, C., Hoehndorf, R., Dressler, S., Hickler, T., Seeger, B. & Schmidt, M. (2018), 'Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks', Botany Letters **165**(3-4), 377–383. 11

Zaragüeta-Bagils, R. & Bourdon, E. (2007), 'Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data', Comptes Rendus Palevol **6**(6-7), 527–534. 12

Zhou, N., Siegel, Z. D., Zarecor, S., Lee, N., Campbell, D. A., Andorf, C. M., Nettleton, D., Lawrence-Dill, C. J., Ganapathysubramanian, B., Kelly, J. W. et al. (2018), 'Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning', PLoS Computational Biology **14**(7), e1006337. 15

A Appendix

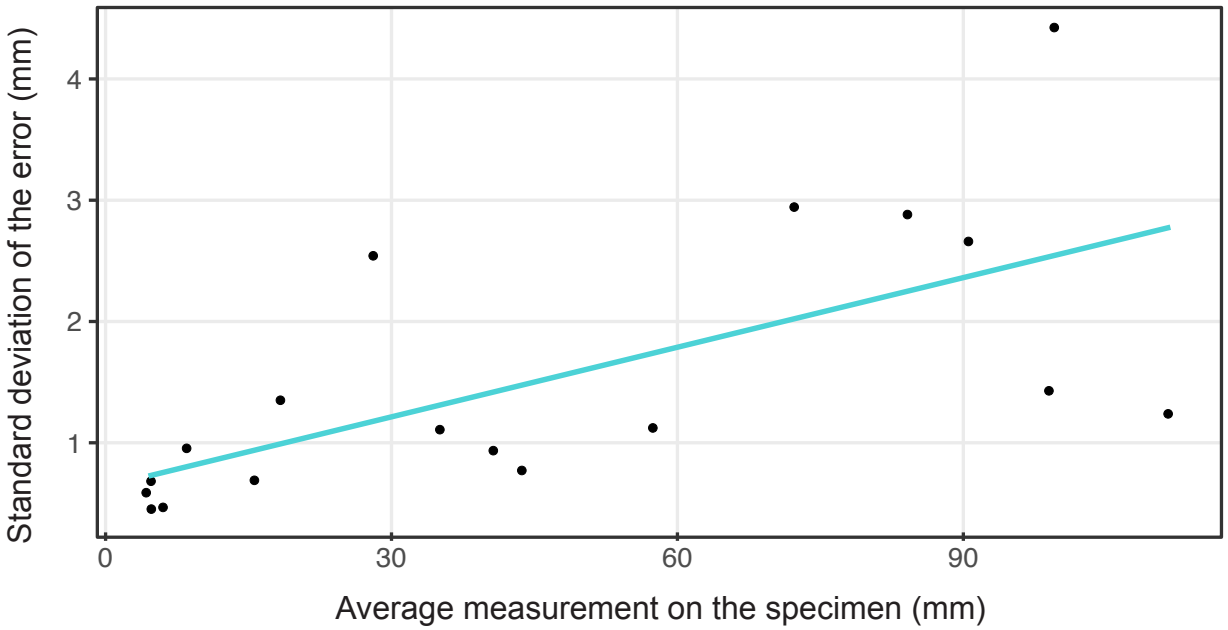

Figure 1: Relationship between the standard deviation of the error of the measurements (i.e., the difference between image and specimen measurements) and the magnitude of each variable. Each dot represents a different variable from the `measurement dataset`.

Table 1: Mean PCA similarity (Sim) and mean accuracy in estimating the first four components (PC1, ..., PC4) using images instead of specimen. This is the data plotted in Figure 6.

dataset	group	type	mean
A2013	PC1	NM	0.89
A2013	PC1	M	0.89
A2013	PC1	N	0.99
A2013	PC2	NM	0.69
A2013	PC2	M	0.72
A2013	PC2	N	0.99
A2013	PC3	NM	0.66
A2013	PC3	M	0.70
A2013	PC3	N	0.96
A2013	PC4	NM	0.33
A2013	PC4	M	0.37
A2013	PC4	N	0.69
A2013	Sim	NM	0.90
A2013	Sim	M	0.93
A2013	Sim	N	0.99
A2008	PC1	NM	0.60
A2008	PC1	M	0.69
A2008	PC1	N	1.00
A2008	PC2	NM	0.64
A2008	PC2	M	0.74
A2008	PC2	N	1.00
A2008	PC3	NM	0.22
A2008	PC3	M	0.23
A2008	PC3	N	1.00
A2008	PC4	NM	0.22
A2008	PC4	M	0.25
A2008	PC4	N	0.93
A2008	Sim	NM	0.29
A2008	Sim	M	0.30
A2008	Sim	N	0.95
A2009	PC1	NM	0.97
A2009	PC1	M	0.97
A2009	PC1	N	1.00
A2009	PC2	NM	0.89
A2009	PC2	M	0.90
A2009	PC2	N	1.00
A2009	PC3	NM	0.36
A2009	PC3	M	0.38
A2009	PC3	N	1.00
A2009	PC4	NM	0.55
A2009	PC4	M	0.60
A2009	PC4	N	1.00
A2009	Sim	NM	0.90
A2009	Sim	M	0.90
A2009	Sim	N	1.00
B2018	PC1	NM	0.97
B2018	PC1	M	0.99
B2018	PC1	N	0.98
B2018	PC2	NM	0.83
B2018	PC2	M	0.96
B2018	PC2	N	0.86
B2018	PC3	NM	0.75
B2018	PC3	M	0.94
B2018	PC3	N	0.76
B2018	PC4	NM	0.79
B2018	PC4	M	0.87
B2018	PC4	N	0.78
B2018	Sim	NM	0.75
B2018	Sim	M	0.82
B2018	Sim	N	0.85
B2015	PC1	NM	0.77
B2015	PC1	M	0.88
B2015	PC1	N	0.97
B2015	PC2	NM	0.86
B2015	PC2	M	0.97

dataset	group	type	mean
B2015	PC2	N	0.98
B2015	PC3	NM	0.45
B2015	PC3	M	0.67
B2015	PC3	N	0.74
B2015	PC4	NM	0.39
B2015	PC4	M	0.56
B2015	PC4	N	0.77
B2015	Sim	NM	0.96
B2015	Sim	M	0.98
B2015	Sim	N	0.99
E2015	PC1	NM	0.96
E2015	PC1	M	1.00
E2015	PC1	N	0.96
E2015	PC2	NM	0.96
E2015	PC2	M	1.00
E2015	PC2	N	0.96
E2015	PC3	NM	0.59
E2015	PC3	M	1.00
E2015	PC3	N	0.56
E2015	PC4	NM	0.52
E2015	PC4	M	1.00
E2015	PC4	N	0.50
E2015	Sim	NM	0.98
E2015	Sim	M	1.00
E2015	Sim	N	0.98
K2006i	PC1	NM	0.71
K2006i	PC1	M	0.87
K2006i	PC1	N	0.94
K2006i	PC2	NM	0.81
K2006i	PC2	M	0.98
K2006i	PC2	N	0.97
K2006i	PC3	NM	0.59
K2006i	PC3	M	0.60
K2006i	PC3	N	0.81
K2006i	PC4	NM	0.48
K2006i	PC4	M	0.53
K2006i	PC4	N	0.76
K2006i	Sim	NM	0.88
K2006i	Sim	M	0.93
K2006i	Sim	N	0.96
K2006p	PC1	NM	0.90
K2006p	PC1	M	0.92
K2006p	PC1	N	0.97
K2006p	PC2	NM	0.86
K2006p	PC2	M	0.93
K2006p	PC2	N	0.97
K2006p	PC3	NM	0.45
K2006p	PC3	M	0.53
K2006p	PC3	N	0.71
K2006p	PC4	NM	0.60
K2006p	PC4	M	0.66
K2006p	PC4	N	0.66
K2006p	Sim	NM	0.74
K2006p	Sim	M	0.81
K2006p	Sim	N	0.91
M1992	PC1	NM	0.99
M1992	PC1	M	1.00
M1992	PC1	N	0.99
M1992	PC2	NM	0.76
M1992	PC2	M	0.92
M1992	PC2	N	0.83
M1992	PC3	NM	0.68
M1992	PC3	M	0.90
M1992	PC3	N	0.81
M1992	PC4	NM	0.38

dataset	group	type	mean
M1992	PC4	M	0.81
M1992	PC4	N	0.19
M1992	Sim	NM	0.94
M1992	Sim	M	0.95
M1992	Sim	N	0.98
P2005	PC1	NM	0.95
P2005	PC1	M	0.95
P2005	PC1	N	0.99
P2005	PC2	NM	0.35
P2005	PC2	M	0.53
P2005	PC2	N	0.81
P2005	PC3	NM	0.23
P2005	PC3	M	0.25
P2005	PC3	N	0.86
P2005	PC4	NM	0.17
P2005	PC4	M	0.22
P2005	PC4	N	0.98
P2005	Sim	NM	0.78
P2005	Sim	M	0.83
P2005	Sim	N	0.95
R2014	PC1	NM	0.77
R2014	PC1	M	0.88
R2014	PC1	N	0.95
R2014	PC2	NM	0.67
R2014	PC2	M	0.71
R2014	PC2	N	0.96
R2014	PC3	NM	0.38
R2014	PC3	M	0.57
R2014	PC3	N	0.50
R2014	PC4	NM	0.14
R2014	PC4	M	0.24
R2014	PC4	N	0.46
R2014	Sim	NM	0.52
R2014	Sim	M	0.57
R2014	Sim	N	0.91
S2012	PC1	NM	0.94
S2012	PC1	M	0.95
S2012	PC1	N	0.98
S2012	PC2	NM	0.66
S2012	PC2	M	0.77
S2012	PC2	N	0.94
S2012	PC3	NM	0.57
S2012	PC3	M	0.49
S2012	PC3	N	0.97
S2012	PC4	NM	0.51
S2012	PC4	M	0.58
S2012	PC4	N	0.74
S2012	Sim	NM	0.79
S2012	Sim	M	0.81
S2012	Sim	N	0.97
T2008	PC1	NM	1.00
T2008	PC1	M	1.00
T2008	PC1	N	1.00
T2008	PC2	NM	0.98
T2008	PC2	M	1.00
T2008	PC2	N	0.98
T2008	PC3	NM	0.98
T2008	PC3	M	1.00
T2008	PC3	N	0.98
T2008	PC4	NM	0.98
T2008	PC4	M	1.00
T2008	PC4	N	0.98
T2008	Sim	NM	1.00
T2008	Sim	M	1.00
T2008	Sim	N	1.00