Zoological Systematics, 42(1): 46–58 (January 2017), DOI: 10.11865/zs.201704

ORIGINAL ARTICLE

The maximum likelihood identification method applied to insect morphometric data

Jean-Pierre Dujardin[1], Sebastien Dujardin[2], Dramane Kaba[3], Soledad Santillán-Guayasamín[4], Anita G. Villacís[4], Sitha Piyaselakul[5], Suchada Sumruayphol[6], Yudthana Samung[6], Ronald Morales Vargas[6]

[1]*IRD, UMR 177 IRD-CIRAD INTERTRYP, Campus international de Baillarguet, Montpellier, France; E-mail: dujjepi@gmail.com*
[2]*École des Technologies Numériques Appliquées, Paris, France; E-mail: sdujardi@gmail.com*
[3]*Institut Pierre Richet, Institut National de Santé Publique, Abidjan, Côte d'Ivoire; E-mail: kaba.dramane2@gmail.com*
[4]*Center for Research on Health in Latin America, School of Biological Sciences, Pontifical Catholic University of Ecuador, Quito, Ecuador, Av. 12 de Octubre 1076 y Roca; E-mail: solsg.25@gmail.com; agvillacis@puce.edu.ec*
[5]*Department of Anatomy, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand; E-mail: sitha.piy@mahidol.ac.th*
[6]*Department of Medical Entomology, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand; E-mail: suchada.sum@mahidol.ac.th; ronald.mor@mahidol.ac.th*

Abstract To distinguish species or populations using morphometric data is generally processed through multivariate analyses, in particular the discriminant analysis. We explored another approach based on the maximum likelihood method. Simple statistics based on the assumption of normal distribution at a single variable allows to compute the chance of observing a particular data (or sample) in a given reference group. When data are described by more than one variable, the maximum likelihood (MLi) approach allows to combine these chances to find the best fit for the data. Such approach assumes independence between variables. The assumptions of normal distribution of variables and independence between them are frequently not met in morphometrics, but improvements may be obtained after some mathematical transformations. Provided there is strict anatomical correspondence of variables between unknown and reference data, the MLi classification produces consistent classification. We explored this approach using various input data, and compared validated classification scores with the ones obtained after the Mahalanobis distance-based classification. The simplicity of the method, its fast computation, performance and versatility, make it an interesting complement to other classification techniques.

Key words Medical entomology, morphometrics, classification, probability, Mahalanobis distance.

1 Introduction

Morphometrics has a long tradition in systematics. In most taxonomic keys, including recent ones, one can find here a ratio, there a range of distances between anatomical points, or somewhere else more complex descriptions such as “its width less than two times the first segment of the antenna”, *etc*. With the increasing awareness of the intraspecific morphological variation, this application of morphometrics applied to taxonomy has developed as a discipline on its own, and was given a powerful push thanks to the integration of multivariate analyzes, especially the principal component analysis (Pimentel, 1992) and, for classification purpose, the canonical variate analysis (Albrecht, 1980; Sorensen, 1992).

The classification ability of the discriminant analysis was leaving unsolved the question whether groups were separate because of environmental influence on their metric properties, or because of genetic separation. With the idea that shape has

Special Issue: Geometric morphometrics: Current shape and future directions
Received 13 September 2016, accepted 6 Janyary 2017
Executive editor: Fuqiang Chen; guest editor: Ming Bai

less environmental influence than size, the multivariate morphometrics suggested various developments aiming at the separation of size and shape (for a review, see Klingenberg, 1996).

Shape as the contour of an organ was already providing interesting possibilities since the 1970s (Lestrel, 1989, 1997). It is only after the 1980s that the landmark-based approaches provided other statistical solutions to capture shape separately from size (Rohlf & Marcus, 1993).

Classifying groups on the base of their shape is generally a good, indirect method to reveal genetic differentiation between populations and to recognize species, even cryptic ones (Perrard *et al.*, 2014). Molecular methods are more accurate to separate species, but they are not always possible or they are simply too costly, or too laborious for routine activities. Because morphological determination might be compromised on damaged specimens, or problematic between cryptic species, complementary and objective techniques are necessary. Morphometrics, especially geometric morphometrics has imposed itself as the fastest and the most affordable technique to help morphological identification. Its quantitative nature and the statistical possibilities it offers make it a scientific, reliable method to distinguish taxa or geographic populations.

Today, the commonly used Mahalanobis distance (the metric of the discriminant space) is among the most powerful approaches to discriminate diverging shapes such as genera, species, subspecies, or even geographic or ecologic populations (Perrard *et al.*, 2014; for a review on medically important insects, see Dujardin, 2008).

Discriminating groups is a problem different than (tentatively) identifying them. The latter needs reference data to which to compare observed data, a classification method, and a decisional algorithm. The most frequently used classification method in morphometrics, either traditional or geometric morphometrics, is the distance-based classification, i.e. the observed data are assigned to the closest group. The metric distance commonly used is the Mahalanobis distance (Mahalanobis, 1936; Sorensen, 1992). Other techniques are gaining interest, such as neural networks (Ripley, 1994), increasingly applied to organismal forms (Marcondez & Borges, 2000; Vaňhara *et al.*, 2007; Fedor *et al.*, 2008; Hernandez-Serna & Jimenez-Segura, 2014; Lorenz *et al.*, 2015).

On the contrary, the maximum likelihood (MLi) classification, suggested for fossils in 2004 (Polly & Head, 2004), is poorly applied. We explored here the method on some insect groups.

The MLi estimation is a method for fitting models to data and comparing those models to find the best one (Fisher, 1950). It is a widely applied method to many fields as diverse as remote sensing (Foody *et al.*, 1992; Paola & Schowengerdt, 1995; Otukei & Blaschke, 2010), computer vision or machine learning methods (Sonnenschein *et al.*, 2015). Developments in morphometrics are discussed in topics related to shape theory or missing data estimation (Mitteroecker & Gunz, 2009). Our study just considers its ability to assign an observed data to a given reference distribution. Its simplicity and versatility could make it an interesting addition to the Mahalanobis classification, and in some situations could represent a valid alternative.

2 Materials and methods

2.1 Method

What is the chance of observing particular data (or sample) in a given reference group? Comparing this chance among various reference groups allows one to propose a classification, hence to suggest an identification by selecting the reference group having the highest chance to fit the data.

The probability P_X that an observation X (an unknown data) belong to a given class R (a putative source, or reference data) was computed from the standard normal deviate z_X, where

$z_X = [X – (\text{mean R})] / (\text{stdev R})$

with (mean R) and (stdev R) the mean and standard deviation values of the reference class, respectively.

The normal deviate allows to compute the cumulative distribution function for the standard normal distribution as:

$P_X = f(z_X)$

The P_X is calculated from the density function of a normal curve fit to the mean and variance of the reference variable. Even with small samples, the normal distribution rather than the t-distribution is appropriate for MLi fitting (Polly & Head, 2004). Our computation of P_X followed the *pnorm* algorithm (see Supplement 2) as implemented in the Tcllib statistics library (http://wiki.tcl.tk/1246) of the Tcl language (https://www.tcl.tk/software/tcltk/license.html).

This simple calculation of P_X allowed to suggest a classification based on a single variable. For multiple characters based classification, or for shape based classification where one individual is described by many variables, the probability P_X was first computed for each variable, then the geometric mean of the probabilities was computed as the sum of the log transformed probabilities. This procedure is valid if the probabilities are independent ones. The log likelihood 'L' of the

unknown X was computed as

$L(X) = \text{sum } (\log P_{Xi})$

The process was repeated for each reference distribution. The reference data which was the best fit for the observed data corresponded to the putative identification of the unknown.

A "leave-one-out" cross-validation procedure (Manly, 2004) was used to assess the effectiveness of the algorithm. Each specimen was iteratively extracted from its sample and treated as though its identity were unknown. Its identity was then estimated using the reference samples (including its own n-1 parent sample).

The MLi identification process is detailed in the Supplement 1. For the global size estimators like the centroid size or the perimeter of a contour, i.e. for a single variable, the algorithm is straightforward: each individual is described by one value, and its best fit is looked for among reference samples. For shape descriptors, or for traditional measurements performed on each individual, i.e. for multivariate data, each individual is represented by a vector of variables. We prepared the data creating five files, as follows: (1) *PC_ur,* the principal components (PC) on the total sample gathering unknown ("u") and reference ("r") data; (2) *PC_u* and *PC_r*, the *PC_ur* splitted into a file for the unknown and a separate file for the reference specimens; (3) *firstPC_r,* the set of first PC providing the best validated reclassification of "r" (thus excluding "u"); (4) *firstPC_u,* the file *PC_u* reduced to this subset of first PC; (5) *firstPC_ur*, the concatenation of *firstPC_u* and *firstPC_r* (which may be obtained also by slicing accordingly the *PC_ur* initial file).

The final step was the classification step, either a validated classification procedure performed on the *firstPC_ur* file, or an identification test. In the validated classification, the step (1) above computes the PC on all individuals, and step (3) includes as "r" all the individuals except the one which was removed. In the identification procedure, the five steps above are repeated adding the unknown individuals one at a time, reducing their possible influence on PC computations.

We show various examples of validated classification on real data. Tests were performed on Diptera, like the *Glossina* spp. ("tsetse" flies), the *Aedes* spp. (mosquitoes), and the *Sergentomyia* spp. and *Phlebotomus stantoni* (sandflies), on Hemiptera like the Triatominae ("kissing bugs"). Wing venation patterns of these different insects were used to perform classical landmark-based analyses (see an example Fig. 1B). We also used egg contour comparing four species belonging to three genera of Triatominae, namely *Triatoma*, *Panstrongylus* and *Rhodnius*, as well as comparing two geographic populations of *R. ecuadoriensis* (Fig. 1A). The same eggs were classified also on the basis of three measurements (Fig. 1C), as in traditional morphometrics.

We finally give an example of identification procedure performed on a set of unidentified male sandflies, using as reference the data of three female species which were captured in the same trap the same night.

The MLi classification scores were compared with the Mahalanobis ones. The identification criterion was based on the shortest Mahalanobis distance between unknown and reference groups.

2.2 Software

The scripts were written in Tcl, or Tool Command Language (https://www.tcl.tk/software/tcltk/license.html) and will be made available online through the TOM development (http://littletom.io/). They are described in Supplement.

2.3 Materials

2.3.1 *Glossina* spp.

Glossina flies are known as tsetse flies, they transmit the parasite responsible for sleeping sickness to the human in Africa, and its corresponding animal disease, the Nagana. We analyzed wings of wild tsetse flies belonging to the *Glossina palpalis* species (Kaba *et al.*, 2016), itself subdivided into the *G. p. palpalis* and *G. p. gambiensis* subspecies (Challier, 1982).

2.3.2 *Aedes* spp.

The *Aedes* mosquitoes contain important vectors of arboviruses such as dengue, chikungunya or zika viruses. Close species *Aedes aegypti*, *Ae. albopictus* and *Ae. scutellaris* may be difficult to identify on damaged specimens (Sumruayphol *et al.*, 2016a). Mosquito larvae were collected in Chachoengsao Province, Thailand, at a variety of breeding places in February 2013. Collected mosquitoes were reared under laboratory until adult emergence.

2.3.3 Triatominae

The members of this subfamily transmit the parasite responsible for the American trypanosomiasis in the New World,

and some of them might be delicate to identify on a morphological basis (Bargues *et al.*, 2010). We used three different approaches, one based on the wing geometry, the second one on the egg contour, and the third one on traditional measurements of egg dimensions (Fig. 1).

Wing landmarks were used to compare two North American subspecies of *T. protracta: T. p. protracta* and *T. p. woodi* (Dujardin *et al.*, 2007). They also were used to compare two cryptic species of the South American genus *Rhodnius*: *R. prolixus* and *R. robustus* (Dujardin *et al.*, 2014) previously identified by molecular tools (Monteiro *et al.*, 2003).

Eggs from four species of Triatominae were examined: *Triatoma carrioni, Panstrongylus chinai* and *P. howardi* and *Rhodnius ecuadoriensis*. Except for *P. howardi*, they were collected in the Loja province in the southern Andean region of Ecuador. The *P. howardi* were collected from the Manabi province, located along the Central Coast of Ecuador. The 78 *P. chinai* eggs came from 48 females (6 localities), the 75 *P. howardi* eggs came from 8 females (one locality), the *T. carrioni* from 7 females (4 localities). The 73 and 76 *R. ecuadoriensis* eggs came from 48 females (5 localities) in Loja and 24 females (3 localities) in Manabí, respectively.

2.3.4 Sergentomyia spp. and Phlebotomus stantoni

Sandflies contain the vectors of various *Leishmania* spp., the causative agents of human and animal leishmaniases (Apiwathnasorn *et al.*, 1989; Maroli *et al.*, 2012; Sumruayphol *et al.*, 2016b). The specimens used in this study were collected accidentally by one of us (RMV) in Thailand when using BG traps for mosquito captures in the city of Bangkok. Ten males and 51 females were found in the traps. Females were relatively easy to identify on morphological ground: 17 *S. bailyi*, 28 *S. barraudi* and 6 *P. stantoni*. Males however could not be assigned to species. These unidentified males were considered here as "unknown" specimens. Assuming that wing shape of males and females are more similar within a same species than between species, these male specimens were tentatively identified using female species as reference.

3 Results

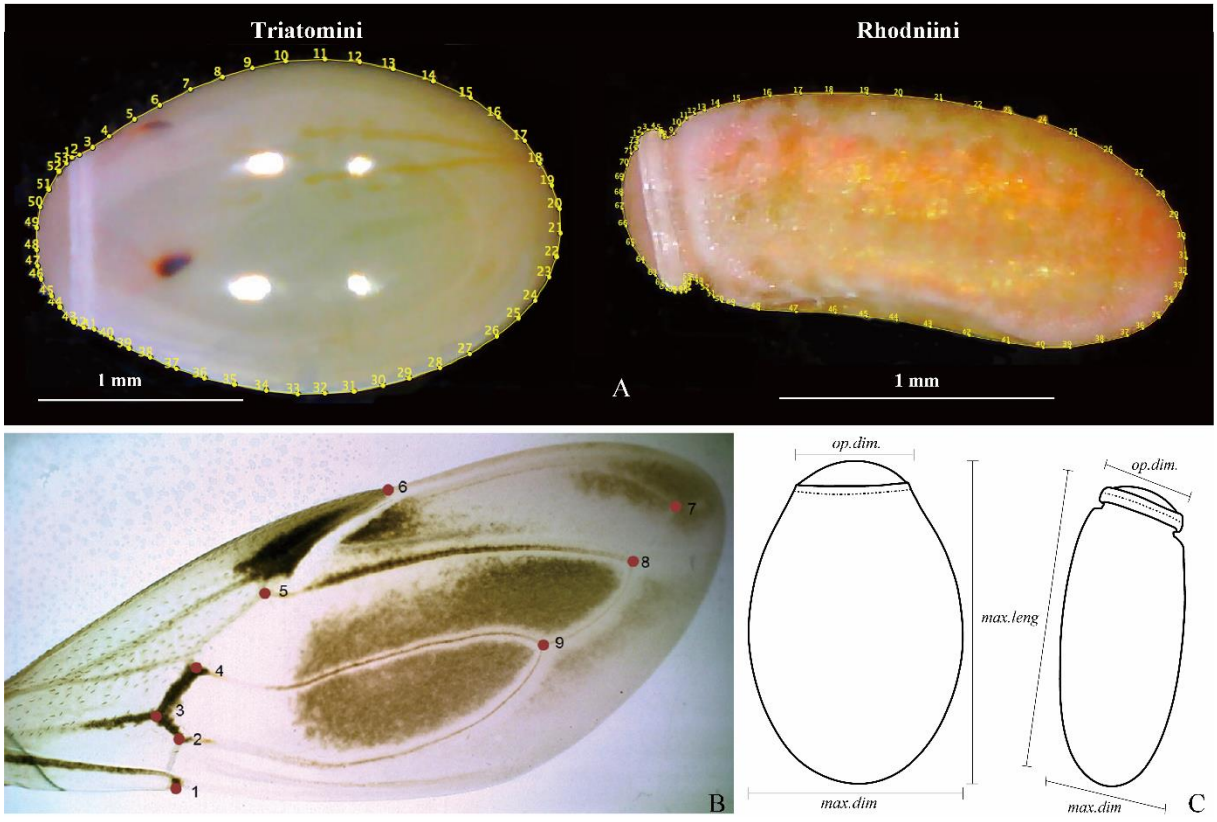

Figure 1. Measurements on Triatominae. A. Digitization of the contour of the eggs of Triatomini and Rhodniini. B. Landmarks as digitized on the wing of *Panstrongylus chinai*. C. Traditional measurement of eggs of the tribe Triatomini (*Panstrongylus* sp. and *Triatoma* sp., left) and the tribe Rhodniini (*Rhodnius* sp., right). Abbreviation: *op.dim*—dimensions of the operculum; *max.leng*—maximum length; *max.dim*—maximum diameter.

The details of the classification results are shown in Tables 1, 2, and 3. Table 4 details the various classifications of 10 unidentified male sandflies using the reference females of three species collected in the same traps the same night. Main conclusions are: (1) size was less informative (and more variable) than shape; (2) shape provided better classifications than size, but a consistent improvement could be obtained using a few first principal components of shape as input; (3) there was no apparent difference in the quality of assignment scores using shape variables based on either landmarks, landmarks and semi-landmarks, or pseudo-landmarks. Figures 2 and 3 illustrate in part these trends, and allow a visual comparison of MLi and Mahalanobis performances.

Table 1. Cross-validation assessment on wing venation patterns. Scores are compared between the maximum likelihood (Mli) and the Mahalanobis methods, and different input variables are considered. *

a. *Triatoma protracta* (Hemiptera: Triatominae) subspecies		
Method, input variables	*T. protracta protracta* (44 M)	*T. p. woodi* (29 M)
MLi, centroid size	73%	62%
MLi, aligned (13 LM)	100%	79%
MLi, shape (4 first PC)	100%	86%
Mahalanobis (4 first PC)	93%	100%
b. *Rhodnius* (Hemiptera: Triatominae) species		
Method, input variables	*R. prolixus* (16 F, 8M)	*R. robustus* (7F, 10M)
MLi, centroid size	88%	88%
MLi, aligned (8 LM)	92%	88%
MLi, shape (5 first PC)	96%	88%
Mahalanobis (5 first PC)	95%	85%
c. *Aedes* (Diptera) species		
Method, input variables	*Ae. albopictus* (80 F)	*Ae. scutellaris* (65 F)
MLi, centroid size	18%	57%
MLi, aligned (16 LM)	58%	83%
MLi, shape (12 first PC)	80%	71%
Mahalanobis (12 first PC)	75%	76%
d. *Glossina* (Diptera) species		
Method, input variables	*G. palpalis palpalis* (138 F)	*G. p. gambiensis* (94 F)
MLi, centroid sizes (LM)	80%	77%
MLi, aligned (11 LM)	93%	90%
MLi, shape (10 first PC)	98%	85%
Mahalanobis (10 first PC)	92%	93%
MLi, centroid sizes (LMSL)	83%	80%
MLi, aligned (LMSL)	90%	89%
MLi, shape (24 first PC)	99%	94%
Mahalanobis (24 first PC)	98%	100%

*Abbreviation: M—males; F—females; LM—landmarks; LMSL—landmarks and semilandmarks; PC —principal components; n—number of specimens; MLi—maximum likelihood classification, based on either size, residual coordinates ("aligned" specimens) or the PC of shape variables; Mahalanobis—classification is based on the Mahalanobis distance computed from the PC of shape variables. For *G. p. palpalis* and *G. p. gambiensis*, "centroid size (LM)" is the centroid size computed from landmarks only, and "centroid size (LMSL)" from landmarks and semi-landmarks.

3.1 Global estimator of size

Cross-validation assessments on size generally produced unsatisfactory (Table 1c) or relatively low (Table 1a, d) scores of validated classification, except for the two *Rhodnius* species, *R. prolixus* and *R. robustus* (Table 1b).

The identification of male sandflies using centroid size (Table 4) produced a very different classification, and less likely to be correct, than based on shape.

3.2 Shape

True shape variables, at least their principal components (PC), provided slightly better scores than simply the aligned variables (Table 1). They allowed to reach scores close to (Tables 1a, c–d, 2a–b) or sometimes higher (Table 1b) than the Mahalanobis scores.

Table 2. Triatominae (Hemiptera): cross-validation assessment on egg contours. *

a. Eggs of *Rhodnius*, a genus of Triatominae (Hemiptera)			
Method, input variables	*R. ecuadoriensis*, Loja (76)	*R. ecuadoriensis*, Manabí (73)	
Mli, perimeter	63.00%	66.00%	
Mli, 20 harmonics	84.00%	82.00%	
Mli, 3 first PC	82.00%	92.00%	
Mahalanobis, 3 first PC	94.00%	86.00%	
b. Eggs of *Panstrongylus* sp. and *Triatoma* sp., two genera of the Triatominae (Hemiptera)			
Method, input variables	*P. chinai* (78)	*P. howardi* (75)	*T. carrioni* (76)
Mli, perimeter	74.00%	33.00%	47.00%
Mli, 20 harmonics	94.00%	87.00%	89.00%
Mli, 11 first PC	87.00%	91.00%	100.00%
Mahalanobis, 11 fist PC	93.00%	96.00%	96.00%

*Scores are compared between the maximum likelihood (MLi) and the Mahalanobis methods. Comparisons are made also according to the input variables. Abbreviation: MLi—maximum likelihood classification, based on either size or shape. Size is the perimeter of the external egg contour. Shape is the set of harmonics containing the normalized Fourier coefficients (NEF), or their principal components (PC). Mahalanobis—the classification based on the Mahalanobis distances computed from PC of NEF.

3.3 Traditional measurements

In our sample, the validated classification of various species gave satisfactory results, even for geographic populations of the same species (Table 3a). They reached the same level as the one obtained with shape variables (see scores of Table 2a versus 3a, 2b versus 3b). According to the species or population, the Mahalanobis classification could provide better (Table 3a) or lower scores (Table 3b), but on average performed slightly better.

Table 3. Triatominae (Hemiptera): cross-validation assessment on egg traditional measurements. *

a. Eggs of *Rhodnius*, a genus of Triatominae (Hemiptera)			
Method, variables	*R. ecuadoriensis*, Loja (78)	*R. ecuadoriensis*, Manabí (57)	
MLi, 3 measurements	82%	82%	
MLi, *max.leng* and *op.dim*	86%	84%	
MLi, 3 PC	87%	84%	
Mahalanobis, 3 PC	89%	85%	
b. Eggs of *Panstrongylus* sp. and *Triatoma* sp., two genera of the Triatominae (Hemiptera)			
Method, variables	*P. chinai* (78)	*P. howardi* (75)	*T. carrioni* (76)
MLi, 3 measurements	83%	88%	96%
MLi, 3 PC	91%	93%	99%
Mahalanobis, 3 PC	91%	93%	100%

*Abbreviation: MLi—maximum likelihood classification; measurements—the three dimensions of the egg (see Fig. 1); PC—their principal components; Mahalanobis—the classification based on the Mahalanobis distances computed from the PC of measurements; "max.leng" and "op.dim"—2 measurements shown in Fig. 1.

3.4 The sandfly example

The morphospace of sandflies (Fig. 4) strongly suggested the unidentified 10 males were probably a mixed sample of *S. bailyi* and *S. barraudi*, leaving however two specimens unclearly positioned. The MLi identification of these males, based on a few first PC of geometric shape, agreed with the morphospace, and suggested an identification for the two intermediate specimens. One was assigned to the *S. baily* and the other to *S. barraudi* (Table 4). The use of global size did not match the

shape classification.

Table 4. Sandflies data. *

Code	Order	Morphospace	MLi on shape	MLi on size	Discrepancies
26193717	1	*bailyi ?*	*bailyi*	*barraudi*	+
26193756	2	*barraudi ?*	*barraudi*	*barraudi*	
28165356	3	*barraudi*	*barraudi*	*barraudi*	
26203702	4	*bailyi*	*bailyi*	*barraudi*	+
26203738	5	*bailyi*	*bailyi*	*barraudi*	+
28153848	6	*bailyi*	*bailyi*	*barraudi*	+
28155127	7	*bailyi*	*bailyi*	*bailyi*	
28155206	8	*bailyi*	*bailyi*	*bailyi*	
28162140	9	*bailyi*	*bailyi*	*bailyi*	
26194353	10	*barraudi*	*barraudi*	*barraudi*	

*Classification of the male sandflies according to their Euclidean distances with the females (column "Morphospace", see Fig. 4), and according to the maximum likelihood classification, either on PC of shape variables (column "Mli on shape") or on size (column "MLi on size"). Plus signs indicate discrepancies between size and shape (column "Discrepancies").

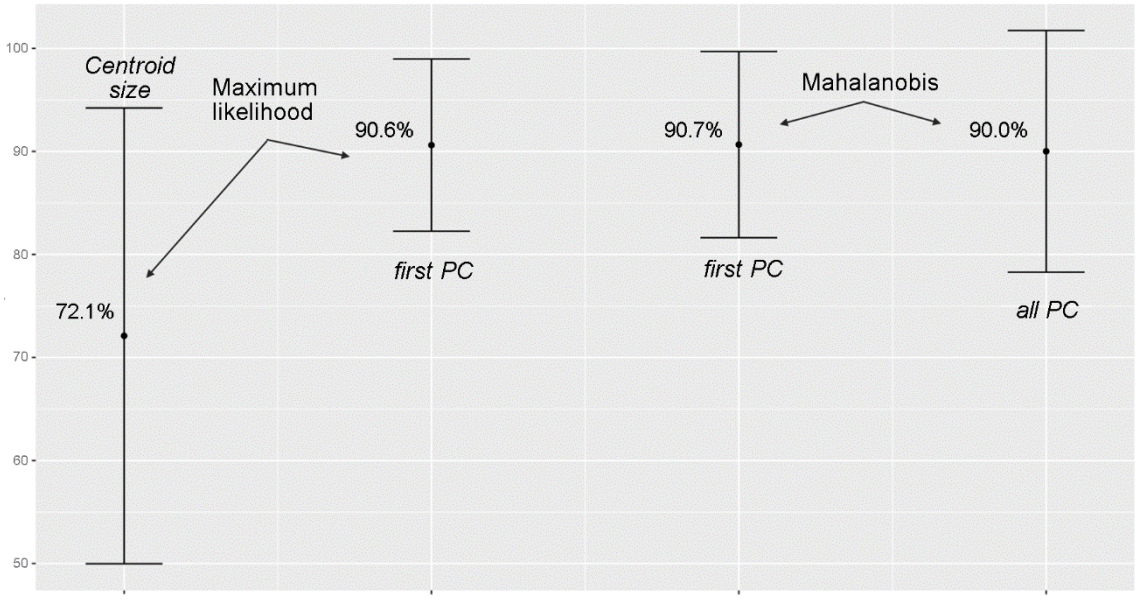

Figure 2. Classification based on landmarks (and semilandmarks). Values in percent are the proportions of correct assignments averaged over the complete set of data, mixing the various species, but restricted to the landmark-based method (including here the combination of landmarks and semilandmarks). Each value is shown with an error bar which is its standard deviation. The left part shows the results from the maximum likelihood method applied to the centroid size and to the shape variables. The right part shows the average scores obtained with the Mahalanobis classification. Abbreviation: *first PC*—the shape variables used were the few first PC of shape variables that the MLi classification selected as the most discriminant set of PC; *all PC*—the maximum number of PC that could be used by the Mahalanobis classification.

4 Discussion

In spite of its versatility and attractive simplicity, the MLi classification of organisms described by morphometrics did not inspire a lot of published works. Such kind of classification has been applied to mammal species or reptiles pieces of skeleton, based on landmark data (Polly & Head, 2004). The method was recommended to deal with small samples of unknown origin or identification, but where the reference sample almost certainly contained the group origin of the unknown specimens. This is the situation of the sandfly sample that we used as an example of MLi identification.

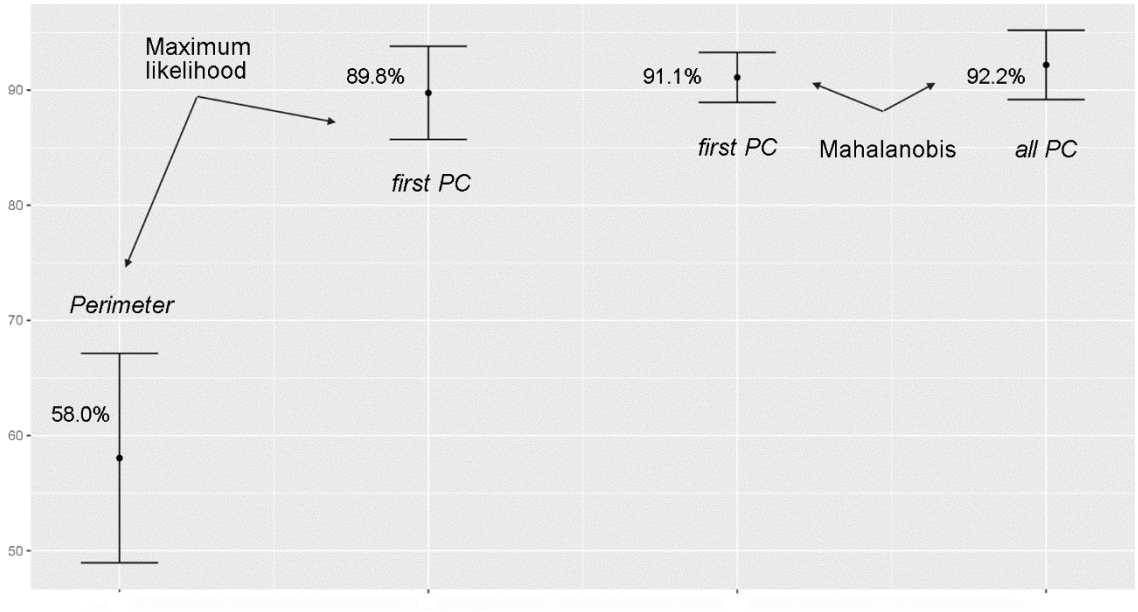

Figure 3. Classification based on harmonics. Values in percent are the proportions of correct assignments averaged over the complete set of data, mixing the various species, but restricted to the outline-based method. Each value is shown with an error bar which is its standard deviation. The left part shows the results from the maximum likelihood method applied to the perimeter of the contour and to the shape variables. The right part shows the average scores obtained with the Mahalanobis classification. Abbreviation: *first PC*—the shape variables used were the few first PC of shape variables that the MLi classification selected as the most discriminant set of PC; *all PC*—the maximum number of PC that could be used by the Mahalanobis classification.

Except for the sandfly data, our analyses did not involve truly unknown specimens, they were MLi-based validated assessment exploring the classification power of the method according to the input variables, and relative to the Mahalanobis method.

There are three theoretical conditions to perform an MLi classification: (1) the normal distribution of each reference variable; (2) the independence between them; (3) the anatomical correspondence between unknown and reference variables.

The first condition assumes that classes in the reference data have a Gaussian (also called "normal") distribution. For the classification based on the size estimator of an organism, like the centroid size or the perimeter of a contour, this condition is the only one to exist for a clean application of the method. But the Gaussian distribution is not always a safe assumption, and if necessary, one can adopt a transformation to a more normal distribution such as log-transformation of size, for instance. In our data, log-transformation of global size estimators (Tables 1, 2 and 4), or of measurements between landmarks (Table 3), did not significantly improve the reclassification scores (details not shown). The MLi method applied to the global size estimator did not provide satisfactory reclassification scores, except in the comparison of the two cryptic species *Rhodnius prolixus* and *R. robustus* (Table 1b): these two taxa are well known to present large differences in size, although sometimes overlapping ones.

The second condition to perform an MLi classification, the independence between characters, is probably never satisfied in morphometrics of organismal forms. In living organisms, dimensions are generally correlated, even highly correlated. The correlation between the sizes of different parts of an organism is studied under the name of "allometry", not to be confounded with the other kind of "allometry" which is the relationship between size and shape. To check for the importance of this condition, we compared the organisms using either raw variables or their principal components. For shape, "raw" variables were either the residual coordinates (the "aligned" specimens), their orthogonal projections onto the tangent space ("tangent space variables", or simply "shape" variables), or the normalized elliptic Fourier coefficients (NEF), according to the digitization technique.

Aligned specimens and their tangent space projections (shape variables) present two drawbacks to be suitable input for a MLi classification (1) they are not independent, due to the overall Procrustes fit, and (2) there is collinearity (loss of 4 degrees of freedom for 2D data). Harmonics are not sample dependent as landmarks are, and they do not suffer from some redundancy, so that they could appear as better candidates for MLi classification. However, at least in our data (Table 2), they did not show better performances than those based on traditional variables (Table 3).

Some transformation of input data, either measurement between landmarks (Table 3) or their coordinates (Tables 1–2),

could be desirable to remove their correlation, at least mathematically. We thus used as input the principal components (PC) of true shape variables, either the tangent space variables or the normalized elliptic Fourier coefficients. As can be computed from the tables (Tables 1–3), the use of the PC of shape could improve the score by an average of up to 5%. This improvement is supposed to be due to more independence between the PC than between raw variables. It could be due also to the rearrangement of the variance because only a few first PC were used.

The third condition is a common sense one: there must be anatomical correspondence between the variable and the reference distributions. Likelihood values are only relevant in comparison to other groups with the same data: it is not meaningful to test the likelihood of leg lengths using the reference distributions of wing lengths. The idea of anatomical correspondence may be put in doubt about landmarks: after Procrustes superimposition, because of the minimum squared optimum criterion, the LM displacement does not necessarily represent the anatomical change it suggests. The anatomical correspondence between harmonics from one individual to another could also be questioned, since harmonics represent just a step to the final shape reconstitution. The anatomical correspondence is easily verified however when it is about traditional measurements, or about the global estimator of size. Traditional measurements do not cast any doubt about anatomical correspondence: as such, they could appear as good candidates for the MLi identification of organisms. Unfortunately, they are also the less powerful variable to describe shape changes between species.

There is an additional condition, or more exactly a warning, also intuitively understandable: the best candidate reference group for a given observation could be completely wrong if no one of the reference groups actually corresponded to the species or the geographic origin of the observed data. If the unknown is not likely to literally be a member of sampled

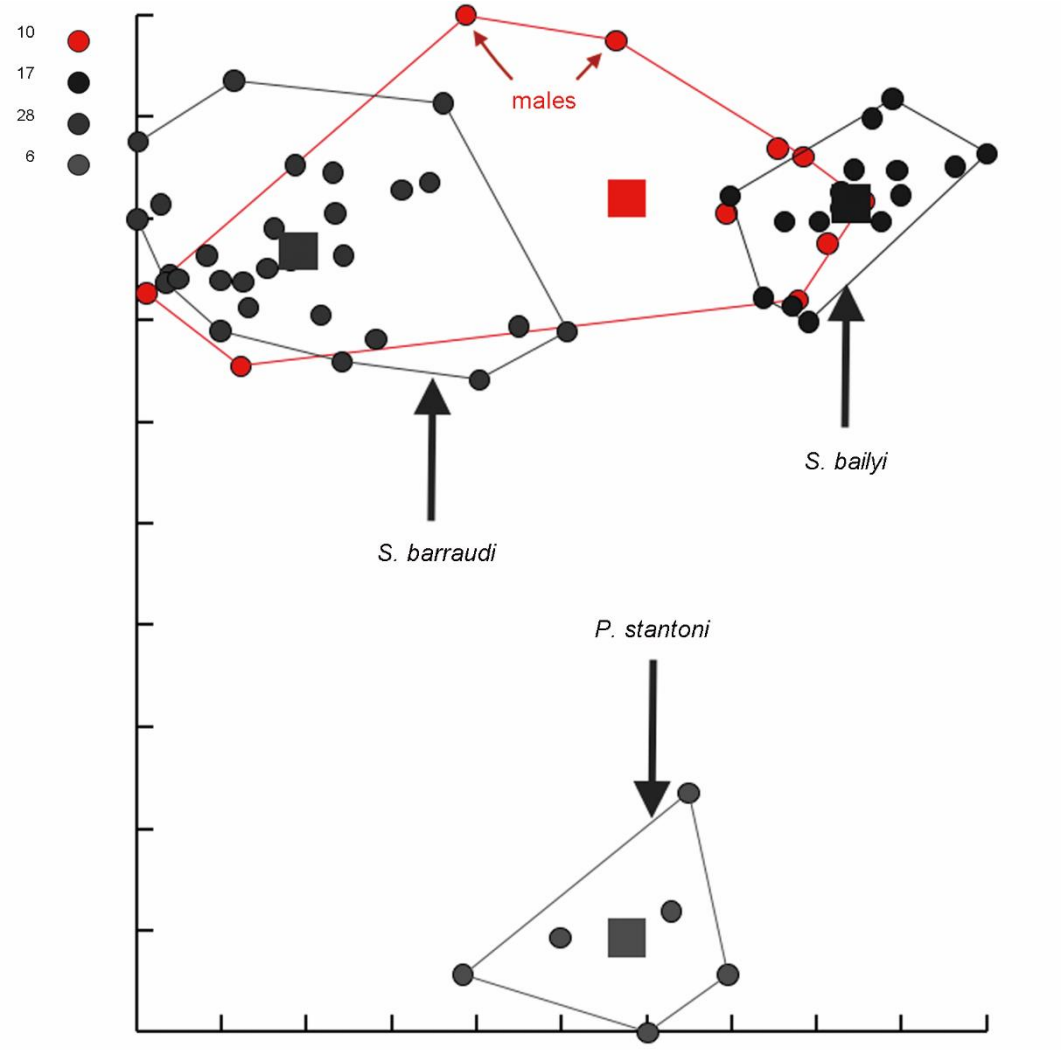

Figure 4. Principal component analysis of sandfly data. The factor map of the two first principal components (PC) describing the morphospace of female sandfly data: *S. bailyi, S. barraudi* and *P. stantoni*. Unidentified males show a convex hull covering both female *S. bailyi* and *S. barraudi*, leaving two specimens outside the female hulls.

reference groups, the MLi classification could just be meaningless. In these case, Polly & Head (2004) recommend a tree-based MLi approach. Another way to deal with this question could be the following: to compare the tentative identification of the unknown using the external reference data, with the validated classification where they are, iteratively, their own reference. If they do not belong to one of the reference groups, they should recognize only themselves.

The MLi approach presents some attributes that make it an interesting complement to the multivariate approach when it is about classifying unknown individuals using reference groups.

1. It can be easily applied to a single variable. The global estimator of size, the centroid size, or the perimeter, can be that variable. For systematics use, size is generally not considered since it is a very plastic character. Most studies show overlapping of size among taxa, even where statistical difference exist, making size poorly informative as a taxonomic marker. The centroid size of male sandflies produced a very different classification than shape, and probably a wrong one when considering the nice correspondence between morphospace (Fig. 4) and shape-based MLi classification (Table 4). We show that, in spite of overlapping values, and provided that statistical differences exist between means, MLi classification based on global size could give consistent results (see for instance the *R. prolixus* and *R. robustus* comparisons, Table 1b). There is a clear example of its interest also in the analysis of sexual dimorphism of the distal femur in humans (Piyaselakul *et al.*, 2017).

2. This singular feature, the ability to classify an individual based on a single variable, makes it possible to select some of the variables among the most discriminant ones. The MLi classification of eggs of Triatomini based on three measurements was shown to be less resolutive than when using the two most discriminant of them (Table 3a). Thus, it could help taxonomists to select species diagnostic morphological traits. It also has the interesting consequence that dimensionality can be reduced.

3. The method has been advocated in previous studies in case of small samples. Small reference samples are possible with the MLi classification, a situation which might compromise a Mahalanobis approach. The sandflies data give an example of a sample hardly tractable by discriminant analyses, since the *P. stantoni* counted 6 specimens only. If one would apply the Mahalanobis distance classification, he would not include such a small sample. It was possible to include it in the MLi classification (Table 4). Polly *et al.* (2004) suggested 10 reference specimens as a minimum. It however may depend on the amount of geometric shape (or size) difference between groups. A common trend visible in the tables is that the best MLi classification scores were obtained for the groups having the largest “n”.

4. Finally, although this point of interest may exit from pure morphometrics, it is worth noting that the MLi method can gather continuous variables from different nature. One can combine for instance morphometric and meristic traits, or variables obtained from different techniques (traditional or geometric). Applied to the problem of identifying the geographic origin of invading pests after local treatment, it has been based on gene frequencies (Dujardin *et al.*, 1996).

5 Conclusions

The MLi method assumes both normality and independence. Some transformation of the data may be necessary to tentatively conform with these assumptions. In our data, the real benefit was obtained by improving independence between variables (using their principal components as input).

On average, the direct comparison between MLi and Mahalanobis classifications gave the latter a slight advantage (Figs 2–3). Looking at the details, the MLi classification could sometimes provide better scores than the Mahalanobis one, which suggests the interest of using more than one approach when it is about classification.

The MLi method performed very well on traditional measurements. The validated classification using one variable at a time allowed to select a combination of traits among the most discriminant ones, improving the classification. The use of 2 variables (*max.leng* and op.*dim*, Fig. 1) instead of 3 could improve the scores of eggs classification by 3% (see Table 3a). Such procedure is easy to perform with the Mli method, and could help taxonomists to select species diagnostic morphological traits.

For shape variables as well as for traditional measurements, the best scores were obtained on their principal components (PC) instead of on the raw variables.

Funding This study was financed by the Chaires Merieux foundation (Paris, France) and Pontifical Catholic University of Ecuador (M 13480).

Acknowledgments We are indebted to the many people having collaborate to the collection of the specimens of this study.

References

Abdi, H., Williams, L.J. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4): 433–459. doi:10.1002/wics.101.

Albrecht, G.H. 1980. Multivariate analysis and the study of form, with special reference to canonical variate analysis. *American Zoologist*, 20: 679–693.

Apiwathnasorn, C., Sucharit, S., Rongsriyam, Y., Leemingsawat, S., Kerdpibule, V., Deesin, T., Surathin, K., Vutikes, S., Punavuthi, N. 1989. A brief survey of Phlebotomine sandflies in Thailand. *The Southeast Asian Journal of Tropical Medicine and Public Health*, 20(3): 429–432.

Bargues, M.D., Schofield, C.J., Dujardin, J.P. 2010. Classification and phylogeny of the Triatominae. *In*: Telleria, J., Tibayrenc, M. (eds.). *American Trypanosomiasis Chagas Disease: One Hundred Years of Research*. Elsevier insights, London. pp. 117–148.

Challier, A. 1982. The ecology of tsetse (*Glossina spp.*) (Diptera, Glossinidae): A review (1970–1981). *Insect Science Applications*, 3(2/3): 97–143.

Dujardin, J.P., Cardozo, L., Schofield, C. 1996. Genetic analysis of *Triatoma infestans* following insecticidal control interventions in Central Bolivia. *Acta Tropica*, 61(3): 263–266.

Dujardin, J.P., Beard, B., Rykman, R. 2007. The relevance of wing geometry in entomological surveillance of Triatominae. *Infection, Genetics and Evolution*, 7(2): 161–167.

Dujardin, J.P. 2008. Morphometrics applied to Medical Entomology. *Infection, Genetics and Evolution*, 8: 875–890.

Dujardin, J.P., Kaba, D., Henry, A.B. 2010. The exchangeability of shape. *BMC Research Notes*, 3: 266. doi: 10.1186/1756-0500-3-266.

Dujardin, J.P., Kaba, D., Solano, P., Dupraz, M., McCoy, K.D., Jaramillo-O, N., 2014. Outline-based Morphometrics, an overlooked method in arthropod studies? *Infection, Genetics and Evolution*, 28: 704–714. doi: 10.1016/j.meegid.2014.07.035.

Fedor, P., Malenovsky, I., Vanhara, J., Havel, J. 2008. Thrips (Thysanoptera) identification using artificial neural networks. *Bulletin of Entomological Research*, 98(5): 437–447.

Fisher, R.A. 1950. *Statistical Methods for Research Workers (11th edition)*. Oliver & Buyd, Edinburgh. 354pp.

Foody, G.M., Campbell, N.A., Trodd, N.M., Wood, T.F. 1992. Derivation and applications of probabilistic measures of class membership from maximum likelihood classification. *Photogrammetric Engineering and Remote Sensing*, 58(9): 1335–1341.

Hernández-Serna, A., Jiménez-Segura, L.F. 2014. Automatic identification of species with neural networks. *PeerJ*, 2: e563. doi: 10.7717/peerj.563.

Kaba, D., Berté, D., Ta, B.T.D., Tellería, J., Solano, P., Dujardin, J.P. 2016. The wing venation patterns to identify single tsetse flies. *Infection, Genetics and Evolution*, 47: 132–139.

Kitthawee, S., Dujardin, J.P. 2016. The *Diachasmimorpha longicaudata* complex in Thailand discriminated by its wing venation. *Zoomorphology*, 135: 323–332. doi: 10.1007/s00435-016-0307-x.

Klingenberg, C.P. 1996. Multivariate allometry. *In*: Marcus, L.F., Corti, M., Loy, A., Naylor, G.J.P., Slice, D.E. (eds.). *Advances in Morphometrics*. Plenum Press, New York. pp. 23–49.

Lestrel, P.E. 1989. Methods for analyzing complex two-dimensional forms: elliptic Fourier functions. *American Journal of Human Biology*, 1: 149–164.

Lestrel, P.E. 1997. Introduction and overview of Fourier descriptors. *In*: Lestrel, P.E. (ed.). *Fourier Descriptors and their Applications in Biology*. Cambridge University Press, Cambridge. pp. 22–44.

Lorenz, C., Ferraudo, A.S., Suesdek, L. 2015. Artificial neural network applied as a methodology of mosquito species identification. *Acta Tropica*, 152: 165–169. doi: 10.1016/j.actatropica.2015.09.011.

Mahalanobis, P.C. 1936. On the generalised distance in statistics. 1936. *Proceedings of the National Institute of Sciences of India*, 2(1): 49–55.

Manly, B.F.J. 2004. *Multivariate Statistical Methods: A Primer, Third Edition*. Chapman & Hall/CRC Press, Boca Raton, FL, US. 214pp.

Marcondes, C.B., Borges, P.S.S. 2000. Distinction of males of the *Lutzomyia intermedia* (Lutz & Neiva, 1912) species complex by ratios between dimensions and by an artificial neural network (Diptera: Psychodidae, Phlebotominae). *Memórias do Instituto Oswaldo Cruz*, 95(5): 685–688.

Maroli, M., Feliciangeli, M., Bichaud, L., Charrel, R.N., Gradoni, L. 2012. Phlebotomine sandflies and the spreading of leishmaniases and other diseases of public health concern. *Medical and Veterinary Entomology*, 27(2): 123–147. doi: 10.1111/j.1365-2915.2012.01034.x.

Mitteroecker, P., Gunz, P. 2009. Advances in geometric morphometrics. *Evolutionary Biology*, 36(2): 235–247. doi: 10.1007/s11692-009-9055-x.

Monteiro, F.A., Barrett, T.V., Fitzpatrick, S., Cordon-Rosales, C., Feliciangeli, D., Beard, C.B. 2003. Molecular phylogeography of the Amazonian Chagas disease vectors *Rhodnius prolixus* and *R. robustus*. *Molecular Ecology*, 12(4): 997–1006.

Mosquera, K.D., Villacís, A.G, Grijalva, M.J. 2016. Life cycle, feeding, and defecation patterns of *Panstrongylus chinai* (Hemiptera: Reduviidae: Triatominae) under laboratory conditions. *Journal of Medical Entomology*, 53(4): 776–781.

Otukei J.R., Blaschke, T. 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation*, 12(Suppl.): S27–S31.

Paola, J.D., Schowengerdt, R.A. 1995. A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. *IEEE Transactions on Geoscience and Remote Sensing*, 33(4): 981–996.

Perrard, A., Baylac, M., Carpenter, J.M., Villemant, C. 2014. Evolution of wing shape in hornets: why is the wing venation efficient for species identification? *Journal of Evolutionary Biology*, 27(12): 2665–2675. doi: 10.1111/jeb.12523.

Piyaselakula, S., Sanannama, B., Dujardin, J.P. 2017. Multiple 2D approaches to human sexual dimorphism of the distal end of femur. *Zoological Systematics*, 42(1): 107–121.

Pimentel, R.A. 1992. An introduction to ordination, principal components analysis and discriminant analysis. *In*: Foottit, R.G., Sorensen, J.T. (eds.). *Ordination in the Study of Morphology, Evolution and Systematics of Insects: Applications and Quantitative Genetic Rationales*. Elsevier, New York. pp. 11–28.

Polly, P.D., Head, J.J. 2004. Maximum-likelihood identification of fossils: taxonomic identification of Quaternary marmots (Rodentia, Mammalia) and identification of vertebral position in the pipesnake Cylindrophis (Serpentes, Reptilia). *In*: Elewa, A.M.T. (ed.). *Morphometrics, Applications in Biology and Paleontology*. Springer Verlag Berlin Heidelberg, New York. pp.197–221.

Ripley, B.D. 1994. Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B: Methodological*, 56(3): 409–456.

Rohlf, F.J., Marcus, L.F. 1993. A revolution in morphometrics. *Trends in Ecology & Evolution*, 8(4): 129–132.

Sorensen, J.T. 1992. The use of discriminant function analysis for estimation of phylogeny: partitioning, perspective and problems. *In*: Foottit, R.G., Sorensen, J.T. (eds.). *Ordination in the Study of Morphology, Evolution and Systematics of Insects: Applications and Quantitative Genetic Rationales*. Elsevier, New York. pp. 65–93.

Sumruayphol, S., Apiwathnasorn, C., Ruangsittichai, J., Sriwichai, P., Attrapadung, S., Samung, Y., Dujardin, J.P. 2016a. DNA barcoding and wing morphometrics to distinguish three Aedes vectors in Thailand. *Acta Tropica*, 159: 1–10.

Sumruayphol, S., Chittsamart, B., Polseela, R., Sriwichai, P., Samung, Y., Apiwathnasorn, C., Dujardin, J.P. 2016b. Wing geometry of *Phlebotomus stantoni* and *Sergentomyia hodgsoni* from different geographical locations in Thailand. *Comptes Rendus Biologies*, doi: 10.1016/j.crvi.2016.10.002.

Sonnenschein, A., VanderZee, D., Pitchers, W.R., Chari, S., Dworkin, I. 2015. An image database of *Drosophila melanogaster* wings for phenomic and biometric analysis. *GigaScience*, 4: 25. doi: 10.1186/s13742-015-0065-6.

Vaňhara, J., Muráriková, N., Malenovský, I., Havel, J. 2007. Artificial neural networks for fly identification: A case study from the genera *Tachina* and *Ectophasia* (Diptera, Tachinidae). *Biologia*, 62(4): 462–469. doi: 10.2478/s11756-007-0089-1.

Villacís, A.G., Grijalva, M.J., Catalá, S.S. 2010. Phenotypic variability of *Rhodnius ecuadoriensis* populations at the Ecuadorian central and southern Andean region. *Journal of Medical Entomology*, 47(6): 1034–1043.

Villacís, A.G., Ocaña-Mayorga, S., Lascano, M.S., Yumiseva, C.A., Baus, E.G., Grijalva, M.J. 2015. The Abundance, natural infection with Trypanosomes, and food source of an endemic species of Triatomine, Panstrongylus howardi (Neiva 1911), on the Ecuadorian Central Coast. *The American Journal of Tropical Medicine and Hygiene*, 92(1): 187–192. doi:10.4269/ajtmh.14-0250.

Supplement 1. The MLi classification, with TOM.

The user will be asked only to hit one button, either the one for an MLi-based validated classification, or the one for MLi-based identification. If necessary, TOM will ask for more information about the nature of input data. However, behind the scene, TOM will use the following Tcl scripts:

Scripts	Input data
ccc_ml-One	One variable
ccc_ml-X	Multiple variables
ccc_ml_PC	Data are principal components
unkn_ml	Any kind of the above data, but preferably the *firstPC_ur*.txt file (see below).

The first part of the name refers to its function, the next part to the input data. Thus "*ccc_*" refers to cross checked classification, i.e. validated assessment, and "*unkn_*" refers to unknown, i.e. identifying an unknown individual or a group of them according to reference groups.

An important point is that the steps described in the Materials & Methods require a file of the total sample, gathering unknown and reference data, with the unknown as the very first group, followed by the reference ones. For a validated classification, 6 files will be produced and stored in the user's space, they are described in the Materials & Methods section: *PC_ur*.txt, *PC_u*.txt, *PC_r*.txt, *firstPC_r*.txt, *firstPC_u*.txt and *firstPC_ur*.txt

To perform a validated MLi-based classification, the user must enter the input file gathering unknown and reference specimens. If data are described by more than one variable, the six files mentioned above will be produced, as well as a final report. If data are described by a single variable (see above "One variable"), only a final report file is produced.

To perform an identification is a different task. In case of multivariate data, the identification procedure starts with the validated classification one performed on the *PC_ur*.txt file, allowing to select a number of first PC with which the identification will actually be performed. To avoid excessive influence of the unknown individuals, the *PC_ur.txt* file will be computed separately for each unknown specimen, in the same way the "one-by-one" process was suggested for the Mahalanobis distance based classification (Dujardin *et al.*, 2010; Kitthawee & Dujardin, 2016; Kaba *et al.*, 2016).

It is not recommended to perform the identification on the total number of PC. Indeed, the quality of the prediction does not always increase with the number of PC of the model: it generally first increases, then decreases (Abdi & Williams, 2010).

In case of univariate data, one can enter the "One variable" file "as is".

Final report files are saved in the user's space with the name of the input file ending by "..._MLi.txt".

Supplement 2. Source code of the *pnorm* function of the Tcllib library.

```
#
proc ::math::statistics::pnorm {x} {
   #
   # cumulative distribution function (cdf)
   # for the standard normal distribution like in the statistical software 'R'
   # (mean=0 and sd=1)
   #
   # x -> value for which the cdf should be calculated
   #
   set sum [expr {double($x)}]
   set oldSum 0.0
   set i 1
   set denom 1.0
   while {$sum != $oldSum} {
         set oldSum $sum
         incr i 2
         set denom [expr {$denom*$i}]
         set sum [expr {$oldSum + pow($x,$i)/$denom}]
   }
   return [expr {0.5 + $sum * exp(-0.5 * $x*$x - 0.91893853320467274178)}]
}
```