Myrmecological News	19	1-15	Vienna, January 2014

Application of Exploratory Data Analyses opens a new perspective in morphology-based alpha-taxonomy of eusocial organisms

Bernhard SEIFERT, Markus RITZ & Sándor CSŐSZ

Abstract

This article introduces a new application of the Exploratory Data Analysis (EDA) algorithms Ward's method, Unweighted Pair Group Method with Arithmetic Mean (UPGMA), *K*-Means clustering, and a combination of Non-Metric Multidimensional Scaling and *K*-Means clustering (NMDS-*K*-Means) for hypothesis formation in morphology-based alpha-taxonomy of ants. The script is written in R and freely available at: http://sourceforge.net/projects/agnesclustering/. The characteristic feature of the new approach is an unconventional application of linear discriminant analysis (LDA): No species hypothesis is imposed. Instead each nest sample, composed of individual ant workers, is treated as a separate class. This creates a multidimensional distance matrix between group centroids of nest samples as input data for the clustering methods. We mark the new method with the prefix "NC" (Nest Centroid). The performance of NC-Ward, NC-UPGMA, NC-*K*-Means clustering, and a combination of Non-Metric Multidimensional Scaling and *K*-Means clustering (NC-NMDS-*K*-Means) was comparatively tested in 48 examples with multiple morphological character sets of 74 cryptic species of 13 ant genera. Data sets were selected specifically on the criteria that the EDA methods are likely to lead to errors – i.e., for the condition that any character under consideration overlapped interspecifically in bivariate plots against body size. Morphospecies hypotheses were formed through interaction between EDA and a confirmative linear discriminant analysis (LDA) in which samples with disagreements between the primary species hypotheses and EDA classification were set as wild-cards. Subsequent Advanced Species Hypotheses were formed by aligning Morphospecies Hypotheses with biological and genetic data. Over all 48 cases and all four methods using nest centroid data generated by a hypothesis-free LDA, the mean deviation of clustering from Advanced Species Hypotheses was 5.25% in NC-UPGMA, 2.58% in NC-NMDS-*K*-Means, 2.40% in NC-Ward and 2.09% in NC-*K*-Means. A dramatically larger mean error of 21.50% was observed if *K*-Means used nest-sample means of morphological characters instead of centroid data. This indicates that having first run a hypothesis-free LDA was a deciding factor for the unexpectedly high performance of the new clustering algorithms. Advantages and disadvantages of the EDA methods are discussed. A combination of NC-Ward, NC-UPGMA and NC-*K*-Means clustering is recommended as the most conclusive and most rapidly working routine for the exploration of cryptic species. The method is applicable to any group of eusocial organisms such as ants, bees, wasps, termites, gall-making aphids, thrips, weevils, pistol shrimps, and mole rats. In general, NC-Clustering can be applied for all cohesive systems providing repeats of definitely conspecific elements – e.g., leaves and flowers of the same plant, a coral "head" of genetically identical polyps, an aphid colony produced by a single fundatrix. It can also be used to monitor intraspecific zoogeographical structures. However, the clustering methods presented did not appear to be good tools for the investigation of hybrid scenarios, for which we recommend alternative methods.

Key words: Taxonomy, cryptic species, eusociality, hierarchical cluster analysis, agglomerative nesting, non-hierarchical cluster analysis, multi-dimensional scaling, automated determination.

Myrmecol. News 19: 1-15 (online 12 April 2013)
ISSN 1994-4136 (print), ISSN 1997-3500 (online)

Received 15 November 2012; revision received 28 January 2013; accepted 11 February 2013
Subject Editor: Florian M. Steiner

Bernhard Seifert & Markus Ritz, Senckenberg Museum für Naturkunde Görlitz, Am Museum 1, 02826 Görlitz, Germany. E-mail: bernhard.seifert@senckenberg.de; markus.ritz@senckenberg.de

Sándor Csősz (contact author), MTA-ELTE-MTM, Ecology Research Group, Pázmány Péter sétány 1C, H-1117 Budapest, Hungary. E-mail: sandorcsosz2@gmail.com

Introduction – morphology-based alpha-taxonomy, nomenclature and species delimitation problems

The basic function and challenge of alpha taxonomy is to recognize relevant entities of biodiversity – i.e., to determine boundaries between species or subspecific populations and improve their delimitation. Integrative taxonomy – as a combined view of morphology, genetics and surface biochemistry – undoubtedly offers the most powerful solution to difficult problems of ant taxonomy (CREMER & al. 2008, ROSS & al. 2010, SCHLICK-STEINER & al. 2010, SEPPÄ & al. 2011, GOTZEK & al. 2012, WARD & al. 2012). However, since an integrative taxonomy combining these disciplines is costly in money and manpower and requires the best conservation status of the material under investigation, it is of practical importance constantly to improve the methodology of the backbone-discipline of biodiversity research: morphology-based alpha-taxonomy (MOBAT). SCHLICK-STEINER & al. (2007) emphasized the leading function of MOBAT in biodiversity research, because many practical situations (degrading storage conditions, historical, subfossil or fossil vouchers) exclude access to genetic or biochemical information and because only this discipline (not molecular genetics) can link entomological species delimitation with Zoological Nomenclature (STEINER & al. 2009, BAGHERIAN & al. 2012). We present here a new tool that improves the accuracy and reliability of the results produced by MOBAT data analysis considerably.

In conventional MOBAT approaches, taxonomists use observable morphological characters intuitively to form hypothetical groups of objects with similar characteristics. These groups are assumed to represent either different species or well-differentiated populations within a species' metapopulation. This subjective grouping works well when the differences between the entities are obvious but it is likely to contain errors or be completely wrong if there is data overlap in any of the characters under consideration. Such cases, in which a single fully discriminative morphological key character is missing, are common when cryptic species are involved. The frequency of these was estimated in three ant genera subject to a thorough analysis as approximately 46% of about 94 Palaearctic *Lasius* species, 43% of about 67 Palaearctic *Formica* species, and approximately 52% of about 77 *Cardiocondyla* species worldwide (SEIFERT 2009). As comparably high ratios are predicted for other genera too, the cryptic species problem is expected to occur everywhere in ant taxonomy (FERREIRA & al. 2010). The problem is particularly serious in the main field of alpha-taxonomy: determination of the weakly differentiated worker caste. One manner of treating this identification problem is numeric recording of multiple characters followed by a confirmatory (= hypothesis-driven) data analysis – usually in the form of a linear discriminant analysis (LDA). However, these supervised techniques always require a-priori decisions on the part of the taxonomist, and if the morphological separation of the entities is diffuse, several quite different hypotheses (groupings) might be imposed on the same data set each of which achieves a high "probability of confirmation" in the LDA. Unless there is some independent information available, such as genetic and biochemical data or complementary morphology of sexuals, the taxonomist is entirely unable to determine which hypothesis is more probable. The introduction of exploratory data analyses (EDA) that generate independent grouping hypotheses which then can be compared with the hypotheses generated by the confirmatory data analysis offers one way out of this dilemma.

As a heuristic approach, EDA algorithms are typically designed for recognition of hidden patterns (TUKEY 1977) in a mass of weakly differentiated objects and to yield new insights into multivariate data structures without one or more a priori hypotheses. Many tools have been developed for EDA, and many of them are also used to discern complex patterns in morphological data (GOLOBOFF & al. 2006, BAUR & LEUENBERGER 2011, KLINGENBERG 2011). Several types of visualization technique can be used in various EDA approaches, such as histograms, scatter plots or dendrograms. Histograms and scatter plots are sufficient in order to display results within low-dimensional datasets, but with increasing dimensionality (i.e., the number of objects and axes) their usefulness as a means of producing a visual depiction of structures in data sets decreases rapidly. Dendrograms, in comparison, can depict the structure of a complex dataset in two dimensions only, and they offer the most rapid and transparent means of extracting grouping hypotheses from multiple data sets (ROUSSEEUW 1986). These algorithms typically require little computer capacity and many statistical packages are available to run such analyses. Moreover, hierarchical clustering consistently performs well on biological datasets for many of the validation measures (BROCK 2008). These types of analyses are therefore very popular in scientific research.

Hierarchical clustering is a widely-used method for exploratory learning in various biological domains such as for example genetic diversity in crop species (ODONG & al. 2011), morphological traits and molecular markers in grass species (TABACCHI & al. 2006), allozyme data in various vertebrates and insects (WIENS 2000), venom molecules in scorpions (NASCIMENTO 2006), or protein patterns in *Drosophila* (FELTENS 2010). Hierarchical clustering algorithms construct dendrograms from a distance matrix as input.

Agglomerative Nesting (AGNES) is one variant of hierarchical clustering. It was introduced by KAUFMAN & ROUSSEEUW (1990) and was later improved for increased robustness (STRUYF & al. 1997). Using an objective function that finds the optimal pair of clusters to merge, AGNES successively forms nodes that have the least dissimilarity, until each smallest unit is clustered (DAY & EDELSBRUNNER 1984, R DEVELOPMENT CORE TEAM 2012). We consider in this paper the application of this method in morphological pattern recognition. Two of the six different agglomeration methods, Ward's method (Ward's minimum variance method) (WARD 1963) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA), implemented by the R package cluster (ROUSSEEUW & al. 2012), clearly showed the highest performance in our tests with ant data. This selection is further supported in the case of Ward's method which is used in many standard clustering procedures and in UPGMA, because of its robustness, monotonicity and consistency (STRUYF & al. 1997). We tested these two methods on exclusively morphological data sets of 48 pairs of 74 cryptic ant species and compared the performance with the highly flexible ordination method Non-Metric Multidimensional Scaling (NMDS) (KRUSKAL 1964, LEGENDRE & LEGENDRE 1998) and the non-hierarchical clustering method *K*-Means (LLOYD 1982) which do not require the estimation of distances to nodes.

Material and Methods

Our script is written in R and freely available under the GNU ⁄ GPL licence from the following website: http://sourceforge.net/projects/agnesclustering/.

Preparation of the data matrix by a hypothesis-free linear discriminant analysis

In order to compare realistically the performance of Ward's method, UPGMA, NMDS, and *K*-Means, the input variables were prepared in all four cases in the same way: A cumulative Linear Discriminant Analysis (LDA) was run which considers each nest sample (composed of n individual ant workers) as a different class and determines the centroid position of these sample in the Euclidean space. This procedure builds on the performance of LDA without

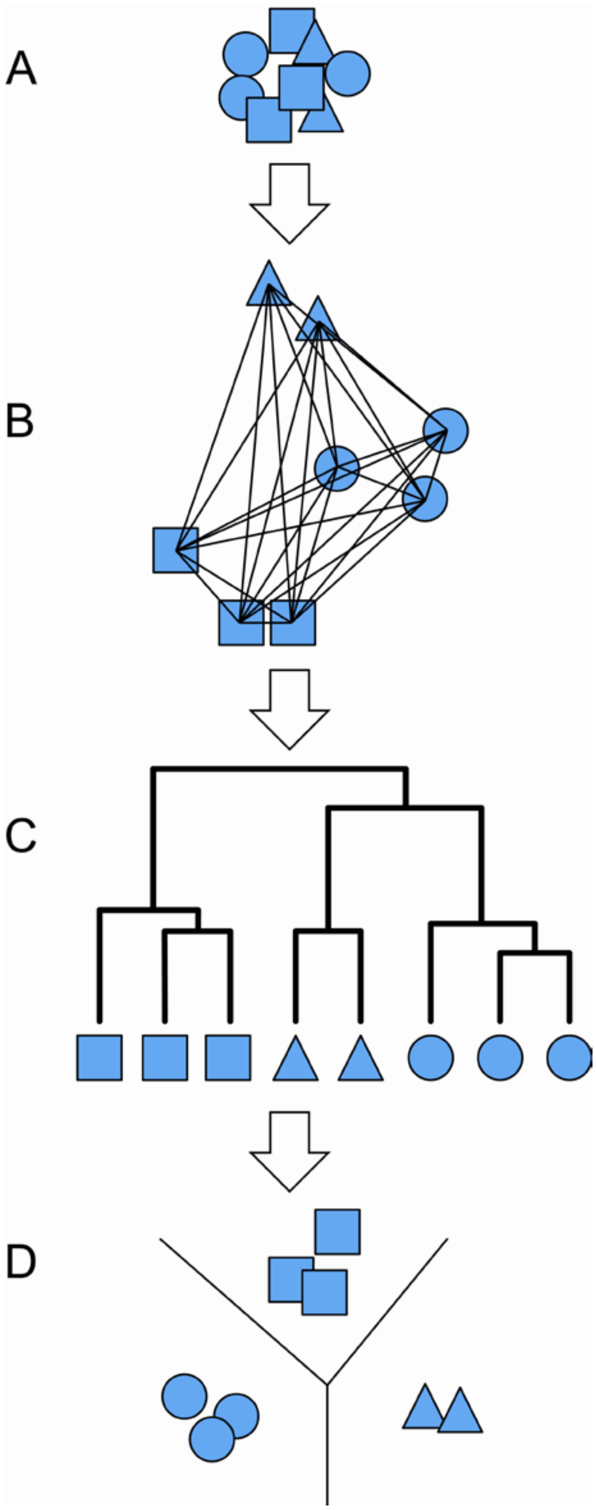

Fig. 1: Steps of proposed exploratory decision making protocol are illustrated. No boundaries are shown (A) by the broadly overlapping raw data of nest samples. Cumulative LDA (B) determines group centroids and creates a distance matrix for the set of samples. Distance-based hierarchical clustering decomposes data into several clusters (C); the less dissimilar objects form one cluster. Samples that nested in a cluster are considered members of newly hypothesized morphological entities. Finally, the accuracy of hypothesized morphological entities is tested (D) according to the supervised confirmative method LDA.

imposing a species hypothesis – i.e., the whole process remains explorative, the LDA is used here as an exploratory tool, not as a confirmation. All four methods then operate with a dissimilarity (distance) matrix based on these centroid coordinates with each nest sample forming one object (Fig. 1). The ways of computing these distances, however, did differ between algorithms and / or distance methods.

Considering that a nest sample is composed of one to many individual workers, it seems problematic to consider them as a single object. However, this assumption biologically makes sense as all ants found within a nest can be assumed to belong to the same species and to be genetically more or less closely related. Exceptions to this rule are socially parasitic and parabiotic ant communities in which different species are found within the same ant nest. In these cases, however, the heterospecific ants, as a rule, are easily separable by simple visual inspection of their phenotype.

Agglomerative Nesting (AGNES)

Agglomerative Nesting (AGNES) is a well-known group of robust hierarchical clustering methods (STRUYF & al. 1997). AGNES proceeds by a series of fusions. It starts with the situation (at step 0), in which each object forms a separate cluster of its own, and then the algorithm merges a pair of minimally dissimilar objects into one cluster. The method successively builds new clusters or merges clusters optimizing least dissimilarity, until each object is part of a cluster (DAY & EDELSBRUNNER 1984, LEGENDRE & LEGENDRE 1998).

Here, we used the AGNES software as introduced by KAUFMAN & ROUSSEEUW (1990) to construct rootless phenetic trees from distance measurements. It is implemented in the R-environment (MAECHLER & al. 2012), where six different agglomeration methods are available: "average" (Unweighted Pair Group Method with Arithmetic Mean, UPGMA), "single" (single linkage), "complete" (complete linkage), "ward" (Ward's method), "weighted" (weighted average linkage) and its generalization "flexible" which uses (a constant version of) the Lance-Williams formula and the par.method argument. Default is "average" (MAECHLER & al. 2012).

In our analyses we compared the two agglomeration methods that clearly showed the highest performance in initial tests with ant data: method = "ward" (Ward's method) and method = "average" (UPGMA).

Agglomeration methods: UPGMA and Ward's method: UPGMA uses the distance matrix of the centroids of nest samples as simple Euclidean distance and Ward's method utilizes squared Euclidean distance. UPGMA first calculates the arithmetic mean of the Euclidean distance between the closest units and forms the first node (pair of objects) for the condition that the mean within-pair distance is smallest. Assume that this node combines samples A and B. The next node is then formed under the same rule – either by merging two other smallest units C, D, ... Z or by merging the cluster AB with one of the other smallest units C, D, ... Z. Rather than presenting further details, we refer here to the excellent and illustrative explanation of the UPGMA procedure (EDWARDS & PARKER 2011). Ward's method essentially uses the same procedure but optimizes within cluster-variance of distances rather than mean of distances. A side effect of this is that Ward's method is constrained to Euclidean distances that are squared during the process. When the fusion is completed, dendrograms are generated that display distance among objects and fusion nodes along their branches. The analysis was run using the AGNES function implemented in the cluster package of R (R DEVELOPMENT CORE TEAM 2012) using the methods "average" (UPGMA) and "ward" with default settings.

K-Means clustering

K-Means clustering is a non-hierarchical / non-nested cluster analysis that aims to partition *n* objects into *K* clusters in which each object belongs to the cluster with the nearest mean (LLOYD 1982). The algorithm uses an iterative refinement technique. Its first step typically includes a subjective decision on the part of the researcher regarding how many *K* clusters have to be separated. *K* starting means are generated and the data space is divided into *K* segments. The data points within these segments form the initial clusters. The centroid of all data points within a segment then becomes the new means and the segments are adjusted accordingly. This process is repeated iteratively until all centroids remain stationary. The approach resembles the Ward-approach described above, and also has conceptual similarities to analysis of variance. Being iterative, *K*-Means-clustering can be trapped in local minima, which is usually addressed by comparing several starting configurations. The analysis was run using the *K*-Means function implemented in the stats package of R using the default algorithm of HARTIGAN & WONG (1979) and ten different (random) starting conditions. In contrast to AGNES, *K*-Means clustering does not require the estimation of distances to nodes, thus avoiding a weakness of agglomerative techniques.

NMDS ordination

Non-Metric Multidimensional Scaling is a highly flexible ordination method that transforms information from an initially provided dissimilarity matrix into a ranking of pairwise dissimilarities and depicts them within a low-dimensional space. We used the NMDS algorithm implemented in the ecodist package of R with default settings and ten iterations. In order to keep the data comparable and to have reasonable performance in each of the 48 cases under investigation, we fixed the number of dimensions to six. This was, in our data sets, a reasonable number of dimensions which resulted in Stress values between 5 and 12. Configurations with Stress values in this range are considered to produce plots that are easily interpretable and have little risk of misinterpretation (Clarke's rules of thumb).

However, the graphical display – 15 bivariate plots of the six first dimensions – was difficult to translate into a species hypothesis. Some plots provided a clear separation of cryptic species while others did not show any structure. In order to make use of the six-dimensional NMDS data, we evaluated these by *K*-Means clustering with Euclidean distances of objects in NMDS space being used as a distance matrix.

The selection of data sets for performance testing

We aimed at rigorous performance testing and only included data sets in which all four exploratory data analysis methods are likely to produce errors. Sufficiently large data sets with numeric recording of multiple phenotypic characters of 209 ant species belonging to 15 genera are available in the Senckenberg Museum of Natural History Görlitz. These data would theoretically allow one to test the discrimination of > 3800 intrageneric species pairs. However, the vast majority of these data sets had to be excluded from the tests – primarily for the following reason: If there is only a single character in a multiple character combination showing no interspecific overlap on the individual level, both UPGMA and Ward's method almost always provide dendrograms showing 100% congruence with preestablished species hypotheses. Therefore such data are not appropriate for test purposes. We established the following rules for the selection of test systems.

(a) The species separation must be very difficult. On the individual level and for each character considered, there must be interspecific overlap demonstrable in bivariate plots against a body size measure (here head size).

(b) Data included both continuous and meristic data types such as linear morphometrics or seta counts. These data were normally distributed in the vast majority of cases. Seta counts of weakly-haired species may be positively skewed in the cases of a few characters but we did not find significant differences in the final clustering whether we normalized the data distributions or not.

(c) A total of at least 40 nest samples must be available in each species pair.

(d) Only two species were considered in each run.

(e) The frequency of single-worker nest samples should be < 50% in the most difficult cases. However, successful analyses can also be run with frequencies > 70%!

(f) Cases obviously including hybrid samples were excluded from the performance test (but see below).

As a result, we selected 42 species pairs represented by 74 cryptic or very closely related species. In case of six species pairs with larger clustering errors, accessory runs with reduced character number were performed. The reduced character set was determined following the stepwise elimination procedure of the linear discriminant function implemented in the software package SPSS 15.0. These data sets and their references are listed in Table 1. In addition to this performance test, we also checked the behavior of the four classification methods in five published hybrid cases of *Formica*, *Lasius*, *Temnothorax* and *Messor*.

Formation of Improved Species Hypotheses in the test data sets

The leading sources of hypothesis formation in these weakly differentiated workers of cryptic ant species were numerically recorded morphological characters but also subjective assessment of the overall phenotypic picture. Subjective morphospecies hypotheses were tested and modified with these data by linear discriminant analyses. They were changed to Improved Species Hypotheses if there was very strong counter-evidence by data sources independent from worker morphology. These were:

(a) Information from the associated sexual castes, which in some species show more obvious interspecific differences than workers. Such data could be used (or were useful) in < 10% of all ant samples.

(b) Chorological data (zoogeography, altitudinal zonation, sympatry, parapatry, syntopic occurrence). These data were constantly available.

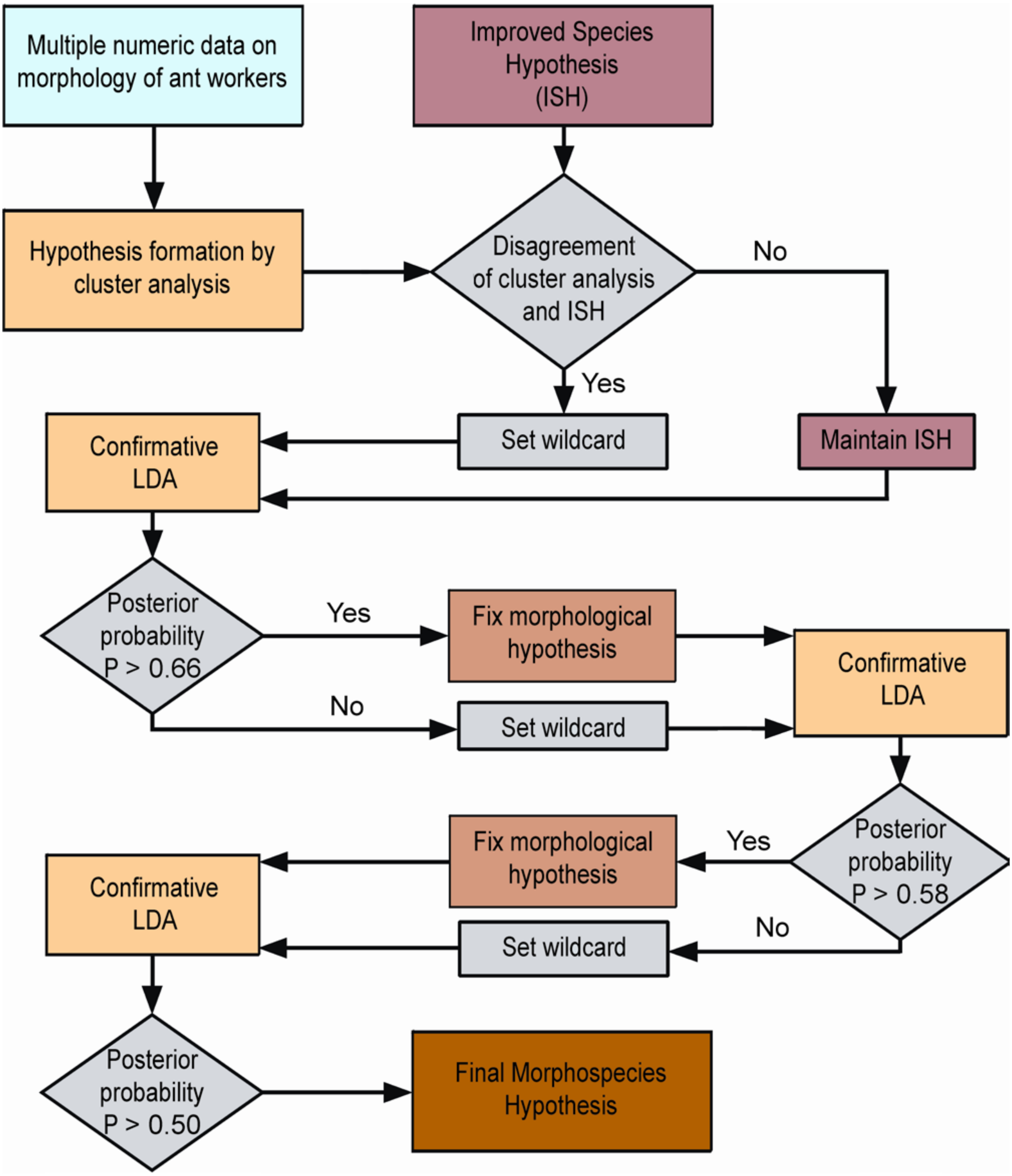

Fig. 2: Flowchart of forming the Final Morphospecies Hypothesis.

(c) Data on habitat selection or nest sites. These data were available in < 50% of all investigated samples.

(d) Seasonal or diurnal timing of swarming. Such data were available in < 10% of samples.

(e) DNA data. These were mainly mtDNA data. The latter were considered to have a rather low weight in decision making due to the high level of mtDNA paraphyly in some ant groups (SEIFERT 2009). They were available in only 20% of the studied species and in no case for complete data sets.

Despite the incomplete availability of the information sources (a) to (e) on the per-sample level, their cumulative information was most useful for assessing if the morphologically delimited entities might represent biospecies.

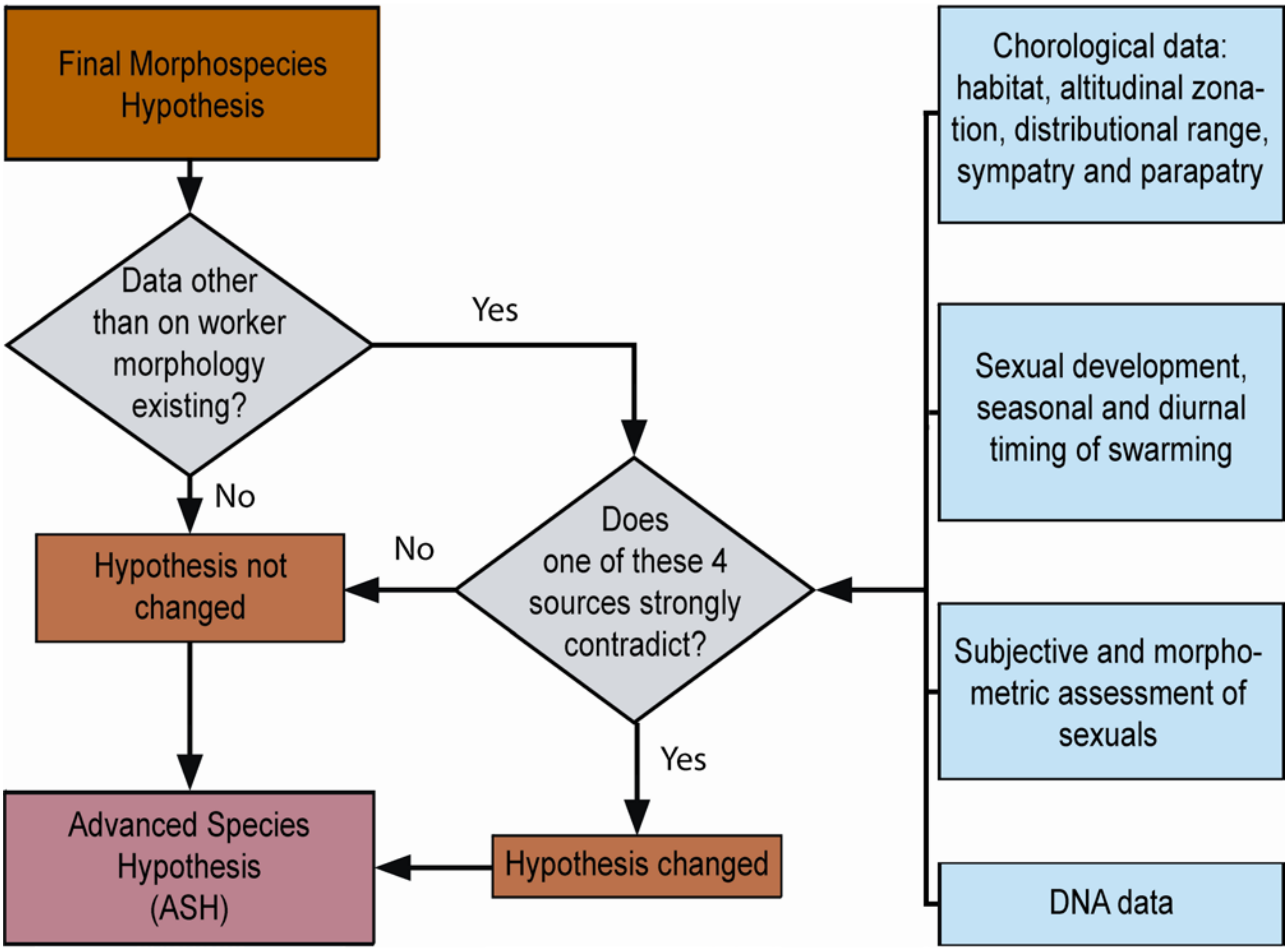

Fig. 3: Flowchart of forming the Advanced Species Hypothesis.

Formation of Advanced Species Hypothesis as interaction between clustering and an iterative LDA operating on the level of individual workers

Advanced Species Hypotheses were determined by combining exploratory and supervised analyses. After running the four clustering methods, the two with the lowest deviation from Improved Species Hypotheses were considered as indicative. Explicitly: Samples not placed in the intersection set being in congruence with Improved Species Hypotheses were marked as wild-cards in a confirmative linear discriminant function – i.e., no species hypotheses were imposed on them. The change from nest sample means to individuals has the advantage of considerably increasing the number of objects. This fulfilled the condition of minimizing the supervisor-induced bias of LDA classifications – i.e., the requirement that the number of objects N in the smallest class should be more than threefold larger than the number of considered morphological characters *C*. We checked the LDA results for all cases with *N* / C ratios 2.9 - 4.0 by a "Leave-One-Out Cross-Validation" LDA (LACHENBRUCH & MICKEY 1968, LESAFFRE & al. 1989) using the software package SPSS 15.0 and found no result to differ by more than 4%.

After the first run of the confirmative LDA, species hypotheses were allocated to samples with the mean posterior probability of all nest members of $P > 0.66$. Samples with $P \leq 0.66$ retained their wild-cards or, exceptionally, even got a wild-card if they had not had it before. In the second run of LDA the decision threshold was reduced to 0.58 and in the last run, if necessary, the most doubtful samples were allocated according to the $p > 0.5$ threshold. No undecided samples remained. We refer to this result of the iterative sample assignment as the Final Morphospecies Hypothesis (Fig. 2). The sample mean of posterior probability P of a species A was calculated as the square root of the product of probabilities for individual workers to belong to species A divided by the sum of the square roots of individual probability products of species A and B. The sample means of posterior probabilities P showed a perfect correlation to the arithmetic mean of individual LDA scores following a logistic sigmoid function. The Final Morphospecies Hypothesis was only accepted if there were no strong counter-arguments from the information sources (a) to (e) as discussed in the previous section. We refer to the classification achieved by this procedure as Advanced Species Hypothesis (Fig. 3). The reader should be aware that an iterative LDA generates contingent posterior probabilities. This does not present a problem within the methodological frame described in this paper but applying an iterative LDA in other contexts should be considered with care.

Considering the substantial problems of delimitation of cryptic species, it is reasonable to expect some deviations

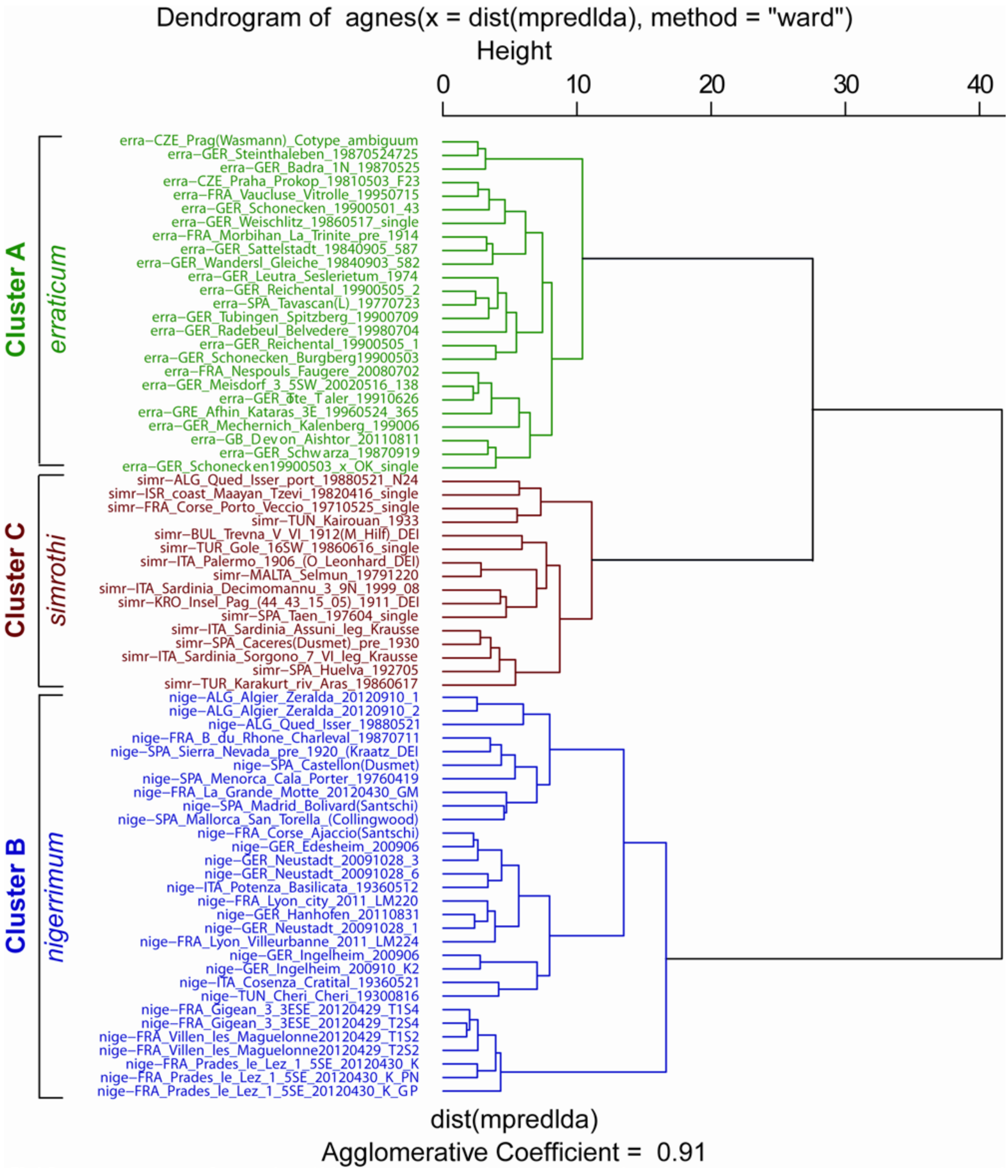

Fig. 4: Ward's method dendrogram of three cryptic species *Tapinoma erraticum*, *T. nigerrimum* and *T. subboreale*. Sequence of information in the string designating the samples: Advanced Species Hypothesis – country – locality – date – sample number. There is a full agreement of AGNES clustering with advanced species hypotheses.

of the Advanced Species Hypotheses from the biologically meaningful entities – at least in the most difficult cases. However, it is not the aim of this methodological paper to discuss taxonomic details, such as substructures within the branches of hypothesized species and their possible biological or evolutionary meaning, or reasons for single misplaced samples, nor is it our goal to discuss complete or incomplete species divergence and intraspecific polymorphism. The important point in this context is to use a defined procedure of hypothesis formation leading to reasonable results and to allow a neutral, comparative performance testing of the four exploratory data analysis methods.

Results and Discussion

Agreement of Advanced Species Hypotheses with published findings

Table 1 shows the data of 48 species discriminations with five clustering methods: 42 used unreduced character combinations and 6 are repetitions of less clearly separated species pairs using reduced character sets. These reduced sets were selected by the stepwise elimination procedure of the linear discrimiant function implemented in SPSS 15.0. In the case of each of the five methods tested, character reduction resulted in a significant improvement of clustering in four species pairs while the situation remained more or less unchanged in two pairs. This indicates that character reduction frequently is a useful step in species discrimination (see also MODER & al. 2007). Applying the more effective character reduction method of MODER & al. is recommended if one has access to the high computing capacity needed.

The Advanced Species Hypotheses presented here are congruent in 35 tested species pairs with species hypotheses published in 25 earlier taxonomic papers (see Tab. 1). Slight deviations (up to 6%) from some of the published hypotheses did occur because the original data files were extended by material investigated later, because new discriminative characters have been introduced, and because the biological background information assisting species delimitation was improved. Seven of the tested species pairs represent unpublished evidence: Five would involve rank elevations of described taxa and two the discovery of new, unknown taxa. We emphasize in this context that any species hypothesis stated in this purely methodological paper has no significance in the sense of Zoological Nomenclature; the rank elevations take effect only after having been published in a formal taxonomic paper.

Comparing the performance of the tested clustering methods

As explained above, the test samples were selected for the condition that the clustering methods are likely to lead to errors – in particular we emphasize the requirement that there must be interspecific overlap in every character considered. In the case of *Tetramorium alpestre* / *impurum*, for example, as much as 68% of worker individuals were found in the 95% confidence interval of interspecific overlap of the most discriminative character and this increased to 78% in the second best character.

A full coincidence of Ward's method dendrograms with Advanced Species Hypotheses in the example of three cryptic *Tapinoma* species *T. erraticum*, *T. nigerrimum* and *T. simrothi* is shown in Figure 4.

Such perfect agreement was achieved in 46% of all classifications via Ward's method, 21% of all UPGMA, 31% of NMDS- *K*-Means and 35% of *K*-Means classifications. Figures 5 and 6, presenting the case of the cryptic Central European ant species *Formica cinerea* and *F. fuscocinerea*, give examples for an average situation. The arrows point to samples in which the AGNES clustering contradicted the Advanced Species Hypotheses (= errors): one sample or 1.3% in Ward's method and three samples or 3.8% in UPGMA. Furthermore the UPGMA dendrogram (Fig. 6) shows two outliers. The errors and outliers were set as wild-cards in the controlling LDA which rectified them in agreement with the Advanced Species Hypotheses.

The mean deviation in the 48 cases from Advanced Species Hypotheses was 5.25% in UPGMA, 2.40% in Ward's method, 2.58% in NMDS-*K*-Means, 2.09% in *K*-Means and 21.50% in *K*-Means without hypothesis-free LDA. The performance of the five methods was tested with a generalized linear model (GLM) with binomial errors and the species pairs as a random factor (grouping variable). Ward's method achieves a significantly higher ratio of coincident assignments than UPGMA and *K*-Means without LDA (both $p < 0.0001$) but performs similarly to NMDS-*K*-Means and *K*-Means with LDA. We direct attention to the dramatically worse classification success of *K*-Means clustering using nest means of morphological primary characters as input instead of the centroid coordinates generated by the hypothesis-free LDA. Accordingly, it appears that having first run a hypothesis-free LDA is one of the key factors of the surprising success of the four other clustering methods.

The use of squared instead of simple Euclidean distance in Ward's method apparently leads to clearer treeing with longer distances between the clusters but possibly also to a higher risk of showing sub-clusters of doubtful biological significance. In other words, there is some danger of over-interpretation. UPGMA tends to form unbalanced clusters (ODONG & al. 2011) and the algorithm often assigns outliers between separate clusters. We confirmed this finding. In general, the dendrograms of UPGMA are more difficult to translate into hypotheses, but an advantage of the method is that the outliers may direct attention to particular samples with deviating characters – outliers might signalize a different taxon, measurement errors, morphological abnormalities or typing errors while creating the input file. Ward's method more frequently tends to hide such samples within a cluster. Accordingly, for control and risk assessment, parallel running of Ward's method and UPGMA is advisable.

Although similar to Ward's method from the perspective of performance, NMDS-*K*-Means and *K*-Means none-

Tab. 1: Deviation in per cent of UPGMA, Ward's method, NMDS-*K*-Means, and *K*-Means clustering from advanced species hypotheses in cryptic species. "*K*-Means no LDA" is a *K*-Means clustering using nest sample means of primary morphological data – i.e., not using centroid data generated by a hypothesis-free LDA. NS = number of samples; N_C = number of characters taken into consideration (an arrow in this column indicates a repetition of the previous case but with a reduced number of characters). All deviations are given in per cent. Abbreviations, [BS] and [SCS] denote unpublished data of Bernhard Seifert [BS] and Sándor Csősz [SCS]. The references for publications indicated [1 - 23] are as follows: [1] SEIFERT (2012b), [2] SEIFERT (2007), [3] SEIFERT (2008), [4] SEIFERT (2003a), [5] SEIFERT (2003b), [6] SEIFERT & SCHULTZ (2009a), [7] SEIFERT & SCHULTZ (2009b), [8] SEIFERT (2000), [9] SEIFERT & GOROPASHNAYA (2004), [10] SEIFERT (2004), [11] SEIFERT (1992), [12] SEIFERT (1988), [13] SEIFERT (1991), [14] SEIFERT & al. (2009), [15] SEIFERT (2005), [16] SEIFERT (2011), [17] CSŐSZ & SEIFERT (2003), [18] SEIFERT (2012a), [19] SEIFERT (1984), [20] SEIFERT (2006a), [21] STEINER & al. (2010), [22] CSŐSZ & al. (2007), [23] CSŐSZ & SCHULZ (2010).

Example	**N_C**	**N_S**	**UPGMA**	**Ward**	**NMDS-*K*-Means**	***K*-Means**	***K*-Means no LDA**	**References**
Bothriomymex communistus / corsicus	16	111	0.00	0.00	0.00	0.00	1.80	[1]
Camponotus atricolor / piceus	11	94	4.25	0.00	3.19	2.12	14.89	[BS], [2]
Camponotus herculeanus / ligniperda	6	49	0.00	0.00	0.00	0.00	22.45	[3]
Cardiocondyla bulgarica/ sahlbergi	17	65	0.00	0.00	3.08	3.08	20.00	[BS], [4]
Cardiocondyla dalmatica / elegans	13	78	10.26	2.56	2.56	2.56	17.95	[BS]
Cardiocondyla mauritanica / kagutsuchi	14	140	0.71	0.71	0.71	0.71	27.14	[BS], [4]
Crematogaster schmidti / scutellaris	10	69	0.00	0.00	1.45	1.45	5.80	[BS], [2]
Formica cinerea / fuscocinerea	13	78	6.41	1.28	2.56	2.56	11.54	[BS], [5]
Formica clara / cunicularia	18	121	4.96	4.13	4.13	2.48	8.26	[6]
Formica clarissima / litoralis	17	143	5.59	0.00	2.80	1.40	35.66	[7]
Formica exsecta / fennica	14	126	0.79	0.00	3.17	3.17	50.00	[BS], [8]
Formica foreli / pressilabris	7	229	1.74	1.74	0.00	0.00	0.87	[BS], [8]
Formica litoralis / pamirica 1 + 2	17	120	3.33	3.33	3.33	3.33	44.17	[7]
Formica lugubris / pratensis Panpalaearctic	5	316	1.27	0.63	0.32	0.32	0.32	[9], [BS]
Hypoponera punctatissima / schauinslandi	10	54	1.85	1.85	0.00	0.00	37.04	[10], [BS]
Lasius barbarus / lasioides	13	100	25.00	6.00	6.00	6.00	17.00	[BS]
Lasius barbarus / lasioides	→8	100	26.00	9.00	7.00	3.00	28.00	[BS]
Lasius emarginatus / illyricus	15	85	4.71	3.53	0.00	1.18	7.06	[BS]
Lasius gebaueri / psammophilus	16	77	0.00	0.00	1.30	1.30	42.86	[BS], [11]
Lasius japonicus / platythorax	14	70	2.86	2.86	0.00	0.00	18.57	[BS], [11]
Lasius mixtus / sabularum	14	66	0.00	0.00	0.00	0.00	33.33	[BS], [12]
Lasius niger / platythorax	14	114	0.00	0.00	0.00	0.00	9.65	[BS], [13]
Lasius paralienus / paralienus 2	16	62	1.61	4.84	3.23	4.84	3.23	[BS]
Lasius paralienus / psammophilus	16	123	1.63	0.00	3.25	2.44	32.52	[BS], [11]
Lasius paralienus 2 */ psammophilus*	16	91	6.59	1.10	1.10	7.69	23.08	[BS]
Lasius piliferus / psammophilus	16	67	7.46	0.00	0.00	0.00	43.28	[BS], [11]
Lasius sabularum / umbratus	14	150	1.33	0.00	0.00	0.00	26.00	[BS], [12]
Myrmica constricta / hellenica	16	91	7.69	3.30	1.10	1.10	9.89	[14]
Myrmica lobicornis /lobulicornis	16	97	0.00	7.22	1.03	1.03	7.22	[15], [BS]
Myrmica lobicornis /lobulicornis	→10	97	2.06	0.00	2.06	2.06	7.22	[15], [BS]
Myrmica salina / specioides	16	161	9.32	9.94	2.48	3.73	30.13	[16], [BS]
Myrmica salina / specioides	→6	161	3.11	9.32	3.11	3.11	5.59	[16], [BS]
Ponera coarctata / testacea	8	141	1.42	0.71	0.00	0.00	10.64	[BS], [17]
Tapinoma erraticum / nigerrimum	14	55	0.00	0.00	0.00	0.00	3.64	[18]
Tapinoma erraticum / simrothi	14	41	2.44	0.00	0.00	0.00	2.44	[BS], [19]
Tapinoma erraticum / subboreale	14	52	0.00	0.00	0.00	0.00	3.84	[18]
Tapinoma nigerrimum / simrothi	14	48	0.00	0.00	0.00	0.00	18.75	[BS], [19]
Temnothorax crassispinus / nylanderi sp. 2	18	100	5.00	2.00	12.00	3.00	50.00	[BS]
Temnothorax lichtensteini / parvulus	18	113	6.19	0.88	7.96	6.19	39.82	[BS]
Temnothorax lichtensteini / parvulus	→ 8	113	1.77	1.77	1.77	1.77	4.42	[BS]
Temnothorax luteus / racovitzai	17	64	14.06	3.12	3.12	3.12	17.18	[BS]
Temnothorax nigriceps / tuberum	18	89	3.37	0.00	1.12	1.12	33.71	[BS]
Temnothorax saxonicus / sordidulus	18	96	17.71	17.71	2.08	1.04	43.75	[20]
Temnothorax saxonicus / sordidulus	→10	96	7.29	4.17	0.00	0.00	14.58	[20]
Tetramorium alpestre / impurum	26	103	33.98	8.74	33.98	20.39	48.54	[21]
Tetramorium alpestre / impurum	→10	103	13.59	2.91	2.91	2.91	38.83	[21]
Tetramorium chefketi / moravicum	17	62	3.23	0.00	0.00	0.00	29.03	[SCS], [22]
Tetramorium diomedeum / ferox	21	63	1.59	0.00	0.00	0.00	30.16	[23]
Mean percentage of 48 cases			5.25	2.40	2.58	2.09	21.50	

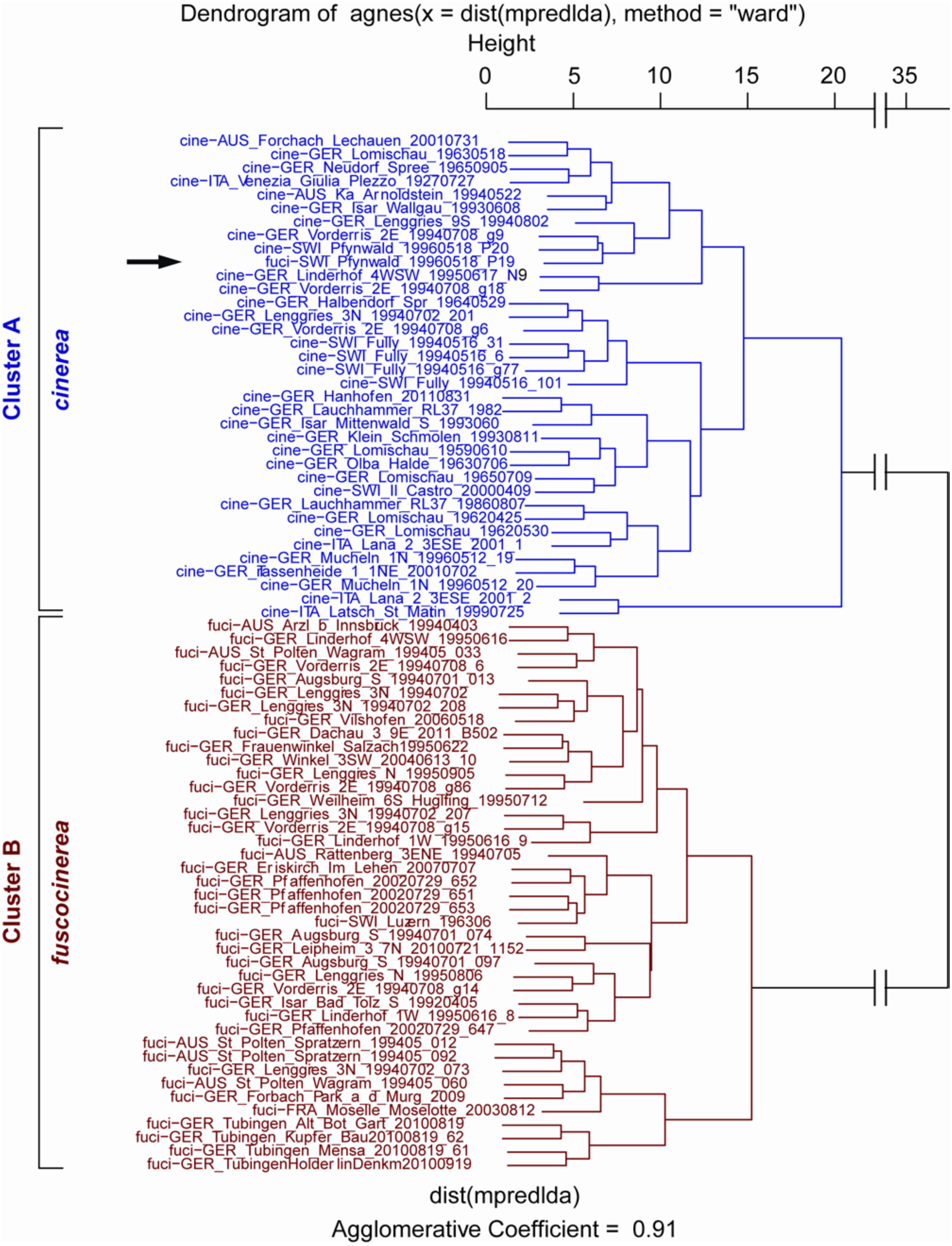

Fig. 5: Ward's method dendrogram of the cryptic species *Formica cinerea* and *F. fuscocinerea*. Sequence of information in the string designating the samples: Advanced Species Hypothesis – country – locality – date – sample number. Deviations from advanced species hypotheses: 1.3% error in Ward's method, 3.8% error and 2.5% outliers in UPGMA. Arrows mark the erroneously placed nest samples.

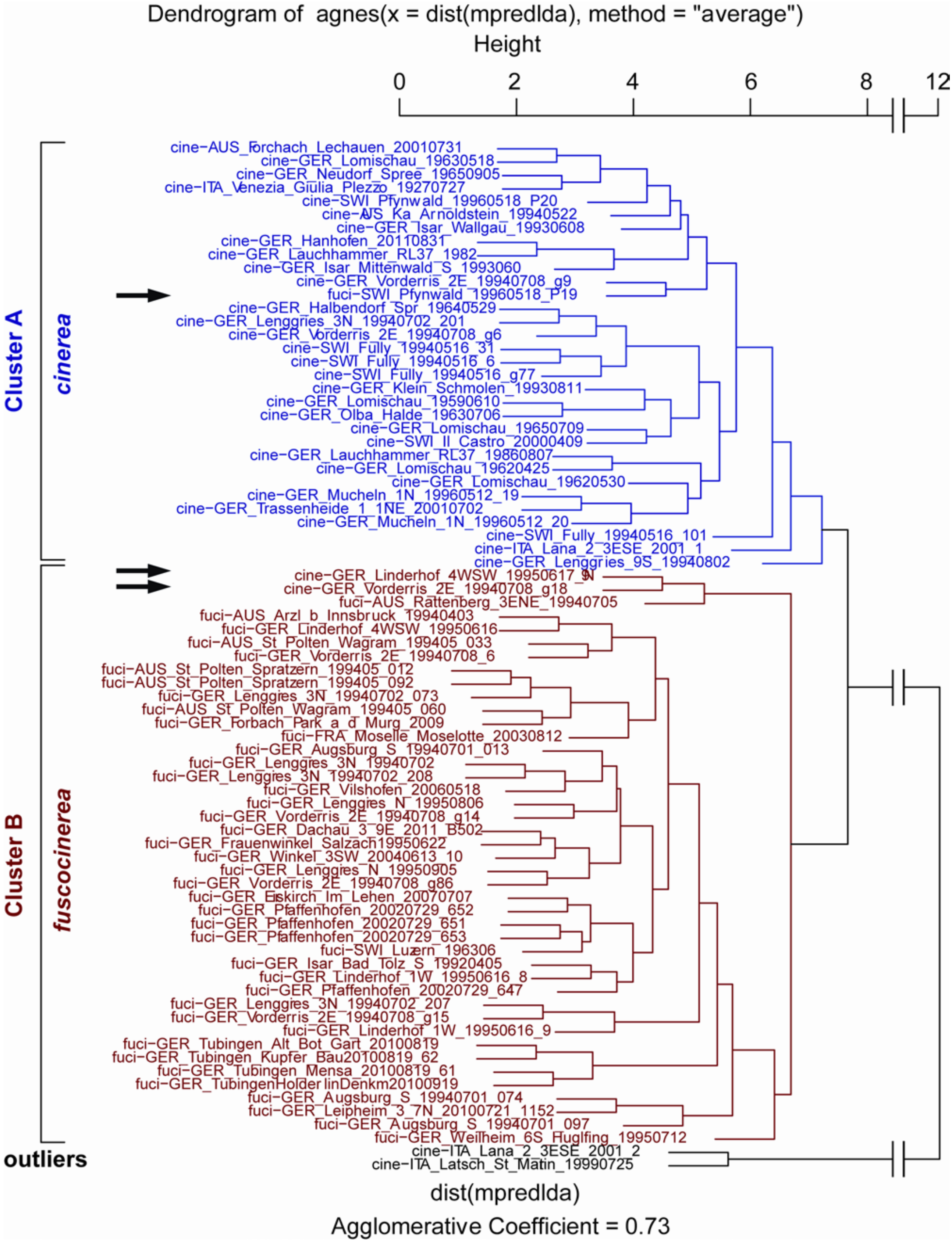

Fig. 6: UPGMA dendrogram of the cryptic species *Formica cinerea* and *F. fuscocinerea*. Sequence of information in the string designating the samples: Advanced Species Hypothesis – country – locality – date – sample number. Deviations from advanced species hypotheses: 1.3% error in Ward's method, 3.8% error and 2.5% outliers in UPGMA. Arrows mark the erroneously placed nest samples.

theless have the disadvantage of requiring an a priori determination of the number of classes. Furthermore they are unable to expose substructures. The dendrograms of both UPGMA and Ward's method, in contrast, provide an immediate idea of sub-clusters of possibly genuine biological significance. A further disadvantage of NMDS-*K*-Means, but not of *K*-Means alone, is an extremely long computing time: A PC having 2MB RAM and a 32-bit CPU needed three minutes to process 156 samples and, because of an exponential growth of computing time, it failed to process 230 samples within two hours.

As one factor probably influencing the extraordinarily positive results shown in Table 1, we must consider the dominance of the morphological methods in the determination of Advanced Species Hypotheses and the comparably low contribution of independent sources of information such as morphology of sexuals, chorology, behavior, ecology, or genetics. The latter data sets are required to test hypotheses formulated by exploratory analyses of morphological data. Expectably, if these Advanced Species Hypotheses would have been determined in each sample by a fully integrative taxonomy including morphology, nuclear DNA and cuticular hydrocarbons (GALIMBERTI & al. 2012), the deviation would probably not appear to be so minute – the more disciplines are involved the higher the probability of conflicts. However, SCHLICK-STEINER & al. (2010), evaluating 184 systematic studies on arthropods that reported diversity at the species level using more than one discipline (morphology, DNA analysis, ecology, enzymes, behavior, life history, cytogenetics, chemistry), found morphology to show a higher congruence with the final species hypotheses formed by integrative taxonomy than analysis of nuDNA, mtDNA or ecology. Given this general arthropod finding and our long-term experience with the biology of many of the ant species considered here, we are confident that the Advanced Species Hypotheses in this paper are close to biologically meaningful entities. The method is applicable to any group of eusocial organisms in which nest members are conspecific such as ants, bees, wasps, termites, gall-making aphids, thrips, weevils, pistol shrimps, or mole rats.

The four clustering methods are not effective tools for analysis of hybridization

We checked the performance of UPGMA, Ward's method, NMDS-*K*-Means and *K*-Means in five published cases of interspecific hybridization (Tab. 2). The delimitation of hybrids and parental species was based in these earlier studies on data on morphology (in all five cases), nuDNA (in three cases), mtDNA (in two cases) and allozymes (in two cases). We found an acceptable clustering only in *Messor* (realistically shown by all four methods), the Ward clustering of *Formica aquilonia / polyctena*, and the *K*-Means clustering of *Formica polyctena / rufa*. In all other cases, the four clustering methods presented confused pictures that did not allow any reasonable hypothesis formation. Accordingly, we do not recommend applying any of these methods in order to identify hybrids. This failure is explained by the fluctuating, to-and-fro nature of phenotypic characters in hybrids. In the same hybrid, one character may approach the situation of parental species A, the second character may be close to parental species B, and a third character may be intermediate (SEIFERT 1984, 1999, 2006b, KULMUNI & al. 2010, SEIFERT & al. 2010, STEINER & al. 2011, BAGHERIAN & al. 2012). This induces a great deal of instability in node formation or cluster allocation. Backcrosses of F1 hybrids with one of the parental species, which are likely to occur in the two *Formica* cases and in *Messor*, would also make the whole picture more diffuse. A robust method for hybrid identification, successfully tested in each of these five cases and another one (BAGHERIAN & al. 2012), is a linear discriminant analysis in which a priori hypotheses are only given to samples reliably identified as pure parental species while problematic or putative hybrid samples are run as wild-cards. The placement along this discriminant vector highly correlates with nuDNA data (KULMUNI & al. 2010, SEIFERT & al. 2010). Principal component analysis was also found to be useful in analyses of hybridization if the character set under consideration was reduced to the most indicative selection.

Conclusion and recommendation

The applications of Ward's method, UPGMA and *K*-Means clustering described here allow the formation of sound hypotheses on formerly unexpected cryptic morphological entities in ants. The outstanding performance is probably caused by running a hypothesis-free linear discriminant analysis that determines the position of nest samples in the Euclidean space by calculating nest sample (group) centroids and creates a distance matrix. The Euclidean distance matrix is a precondition of cluster algorithms, but using the centroid position of nests instead of the raw data of primary characters as input apparently reduces accidental errors during agglomerative node formation and in *K*-Means clustering. We propose to name the whole procedure NC-Ward, NC-UPGMA and NC-*K*-Means (with "NC" meaning "nest centroid"). These clustering methods will help accelerate the often complicated and tedious taxonomical decision making process dramatically. The method is applicable to any group of eusocial organisms such as ants, bees, wasps, termites, gall-making aphids, thrips, weevils, pistol shrimps, or mole rats. In general, NC-Clustering can be applied for all cohesive systems providing repeats of definitely conspecific elements – e.g., leaves and flowers of the same plant, a coral "head" of genetically identical polyps, an aphid colony produced by a single fundatrix. On the intraspecific level, AGNES dendrograms can also depict regional population differentiation if the input data have a high quality. Hence, it is considered a suitable tool to address zoogeographical questions as well.

Summing up, we recommend the following procedure as the best operational routine for morphological clustering of ant samples:

(1) running a hypothesis-free LDA to determine nest-sample centroids in the Euclidean space;

(2) running Ward's method in order to showing structures the most clearly and with a low average error;

(3) running UPGMA in order to indicate outliers. Such samples can be checked afterwards for possible reasons of their outstanding position;

(4) running *K*-Means in order to have a low error rate and testing a possible over-structuring by Ward's method by modifying the *K* value.

(5a) If there is already a previous and feasible species hypothesis, the starting hypothesis of the confirmative LDA is prepared by integrating the cluster information of Ward's

Tab. 2: Deviation in per cent of five clustering methods from identifications of hybrids and parental species based on integrative taxonomy. "*K*-Means no LDA" is a *K*-Means clustering using nest sample means of primary morphological data – i.e., not using centroid data generated by a hypothesis-free LDA. N_S = number of samples; N_C = number of characters taken into consideration (an arrow in this column indicates a repetition of the previous case but with a reduced number of characters). The references for publications indicated [24 - 28] are as follows: [24] KULMUNI & al. (2010), [25] SEIFERT & al. (2010), [26] STEINER & al. (2011), [27] PUSCH & al. (2006), [28] SEIFERT (2006b).

Example	N_C	N_S	**UPGMA**	**Ward**	**NMDS**	***K*-Means**	***K*-Means no LDA**	**References**
Formica aquilonia / aquilonia x polyctena / polyctena	16	145	7.59	3.45	35.86	17.24	34.48	[24]
Formica polyctena / polyctena x rufa / rufa	13	140	7.14	7.14	13.54	1.43	14.28	[25]
Messor minor / minor x wasmanni / wasmanni	17	30	6.67	6.67	6.67	6.67	20.00	[26]
Temnothorax crassispinus / crassispinus x nylanderi / nylanderi	18	143	6.99	6.99	30.07	32.16	42.65	[27]
Lasius jensi / jensi x umbratus / umbratus	14	199	18.59	18.59	26.13	24.12	50.00	[28]
Mean percentage of five cases			9.39	8.57	22.45	16.32	32.28	

method and *K*-Means. The previous hypothesis is maintained for all samples in which both Ward's method and *K*-Means are congruent with the previous hypothesis. All samples in disagreement with the hypothesis according to at least one method (Ward's method or *K*-Means) are marked as wild-cards – i.e., no species hypotheses were imposed on them. The iterative run of the confirmative LDA is done as shown in Figure 2.

(5b) If there exists no feasible species hypothesis, the confirmative LDA is run by setting the samples for which the clustering of Ward's method and *K*-Means disagrees as wild-cards. The iterative LDA is then run in the same way as in (5a). Steps (1) - (4) and (5b) represent a completely explorative determination of morphospecies that is easily programmed as an automatic system. From the data in Table 1 it can be concluded that a completely explorative procedure is expected to have a low error rate because the confirmative LDA corrects most of the clustering errors – i.e., even if a clustering disagreed by 10% with the Advanced Species Hypotheses, most of these deviations were rectified by the LDA.

Acknowledgements

We wish to thank three unknown referees as well as Karsten Wesche (Senckenberg Museum of Natural History Görlitz), Zsolt Pénzes (Szeged University) and David Sheets (Canisius College Buffalo) for giving valuable suggestions on how to improve our manuscript and Thomas Cooper for linguistic corrections.

References

BAGHERIAN, A., MÜNCH, W. & SEIFERT, B. 2012: A first demonstration of interspecific hybridization in *Myrmica* ants by geometric morphometrics (Hymenoptera: Formicidae). – Myrmecological News 17: 121-131.

BAUR, H. & LEUENBERGER, C. 2011: Analysis of ratios in multivariate morphometry. – Systematic Biology 60: 813-825.

BROCK, G., PIHUR, V., DATTA, S. & DATTA, S. 2008: clValid, an R package for cluster validation. – Journal of Statistical Software 25: 1-22.

CREMER, S., UGELVIG, L.V., DRIJFHOUT, F.P., SCHLICK-STEINER, B.C., STEINER, F.M., SEIFERT, B., HUGHES, D.P., SCHULZ, A., PETERSEN, K.S., KONRAD, H., STAUFFER, C., KIRAN, K., ESPADALER, X., D'ETTORRE, P., AKTAÇ, N., EILENBERG, J., JONES, G.R., NASH, D.R., PEDERSEN J.S. & BOOMSMA J.J. 2008: The evolution of invasiveness in garden ants. – Public Library of Science One 3: e3838.

CSŐSZ, S., RADCHENKO, A. & SCHULZ, A. 2007: Taxonomic revision of the Palaearctic *Tetramorium chefketi* species complex (Hymenoptera: Formicidae). – Zootaxa 1405: 1-38.

CSŐSZ, S. & SCHULZ, A. 2010: A taxonomic review of the Palaearctic *Tetramorium ferox* species-complex (Hymenoptera: Formicidae). – Zootaxa 2401: 1-29.

CSŐSZ, S. & SEIFERT, B. 2003: *Ponera testacea* EMERY, 1895 stat.nov. – a sister species of *P. coarctata* (LATREILLE, 1802) (Hymenoptera: Formicidae). – Acta Zoologica Academiae Scientarum Hungaricae 49: 201-214.

DAY, W.H.E. & EDELSBRUNNER, H. 1984: Efficient algorithms for agglomerative hierarchical clustering method. – Journal of Classification 1: 7-24.

EDWARDS, R.J. & PARKER, J.D. 2011: UPGMA worked example. – <http://www.southampton.ac.uk/~re1u06/teaching/upgma/>, retrieved in May 2012.

FELTENS, R., GÖRNER, R., KALKHOF, S., GRÖGER-ARNDT, H. & VON BERGEN, M. 2010: Discrimination of different species from the genus *Drosophila* by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry. – BioMed Central Evolutionary Biology 10: 95.

FERREIRA, R.S., POTEAUX, C., DELABIE, J.H.C., FRESNEAU, D. & RYBAK, F. 2010: Stridulations reveal cryptic speciation in neotropical sympatric ants. – Public Library of Science One 5: e15363.

GALIMBERTI, A., SPADA, M., RUSSO, D., MUCEDDA, M., AGNELLI, P., CROTTINI, A., FERRI, E., MARTINOLI, A. & CASIRAGHI, M. 2012: Integrated operational taxonomic units (IOTUs) in echolocating bats: a bridge between molecular and traditional taxonomy. – Public Library of Science One 7: e40122.

GOLOBOFF, P.A., MATTONI, C.I. & QUINTEROS, A.S. 2006: Continuous characters analyzed as such. – Cladistics 22: 589-601.

GOTZEK, D., BRADY, S.G., KALLAL, R.J. & LAPOLLA, J.S. 2012: The importance of using multiple approaches for identifying emerging invasive species: the case of the rasberry crazy ant in the United States. – Public Library of Science One 7: e45314.

HARTIGAN, J.A. & WONG, M.A. 1979: A *K*-Means clustering algorithm. – Applied Statistics 28: 100-108.

KAUFMAN, L. & ROUSSEEUW, P.J. 1990: Finding groups in data: an introduction to cluster analysis. – John Wiley and Sons, New York, pp. 342.

Klingenberg, C.P. 2011: MorphoJ: an integrated software package for geometric morphometrics. – Molecular Ecology Resources 11: 353-357.

Kruskal, J.B. 1964: Nonmetric multidimensional scaling: a numerical method. – Psychometrika 29: 115-129.

Kulmuni, J., Seifert, B. & Pamilo, P. 2010: Segregation distortion causes large-scale differences between male and female genomes in hybrid ants. – Proceedings of the National Academy of Sciences of the United States of America 107: 7371-7376.

Lachenbruch, P. & Mickey,M. 1968: Estimation of error rates in discriminant analysis. – Technometrics 10: 1-11.

Legendre, P. & Legendre, L. 1998: Numerical ecology. – Elsevier, Amsterdam, 853 pp.

Lesaffre, E., Willems, J.L. & Albert, A. 1989: Estimation of error rate in multiple group logistic discrimination. The approximate leaving-one-out method. – Communication in Statistics: Theory and Methods 18: 2989-3007.

Lloyd, S. 1982: Least squares quantization in PCM. – IEEE Transactions on Information Theory 28: 129-137.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. 2012: Cluster: cluster analysis basics and extensions. R package version 1.14.3. – <http://CRAN.R-project.org/package =cluster>, retrieved in October 2012.

Moder, K., Schlick-Steiner, B.C., Steiner, F.M., Cremer, S., Christian, E. & Seifert, B. 2007: Optimal species distinction by discriminant analysis: comparing established methods of character selection with a combination procedure using ant morphometrics as a case study. – Journal of Zoological Systematics and Evolutionary Research 45: 82-87.

Nascimento, D.G., Rates, B., Santos, D.M., Verano-Braga, T., Barbosa-Silva, A., Dutra, A.A.A., Biondi, I., Martin-Eauclaire, M.F., Lima, M.E.D. & Pimenta, A.M.C. 2006: Moving pieces in a taxonomic puzzle: venom 2D-LC / MS and data clustering analyses to infer phylogenetic relationships in some scorpions from the Buthidae family (Scorpiones). – Toxicon 47: 628-639.

Odong, T.L., van Heerwaarden, J., Jansen, J., van Hintum, T.J.L. & van Eeuwijk, F.A. 2011: Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for molecular marker data? – Theoretical and Applied Genetics 123: 195-205.

Pusch, K., Seifert, B., Foitzik, S. & Heinze, J. 2006: Distribution and genetic divergence of two parapatric sibling ant species in Central Europe. – Biological Journal of the Linnean Society 88: 223-234.

R Development Core Team 2012: R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria. – <http://www.R-project.org/>, retrieved in March 2013.

Ross, K.G., Gotzek, D., Ascunce, M.S. & Shoemaker, D.D. 2010: Species delimitation: a case study in a problematic ant taxon. – Systematic Biology 2010: 162-184.

Rousseeuw, P., Struyf, A., Hubert, M. & Maechler, M. 2012: Cluster: cluster analysis extended Rousseeuw et al. R package version 1.14.2. – <http://CRAN.R-project.org/package=cluster/>, retrieved in October 2012.

Rousseeuw, P.J. 1986: A visual display for hierarchical classification. In: Diday, E., Escoufier, Y., Lebart, L., Pages, J., Schektman, Y. & Tomassone, R. (Eds.): Data analysis and informatics 4. – Elsevier, Amsterdam, pp. 743-748.

Schlick-Steiner, B.C., Seifert, B., Stauffer, C., Christian E, Crozier, R.H. & Steiner, F.M. 2007: Without morphology, cryptic species stay in taxonomic crypsis following discovery. – Trends in Ecology & Evolution 22: 391-392.

Schlick-Steiner, B.C., Steiner, F.M., Seifert, B., Stauffer, C., Christian, E. & Crozier, R.H. 2010: Integrative taxonomy: a multi-source approach to exploring biodiversity. – Annual Review of Entomology 55: 421-438.

Seifert, B. 1984: A method for differentiation of the female castes of *Tapinoma ambiguum* Emery and *Tapinoma erraticum* (Latreille) and remarks on their distribution in Europe north of the Mediterranean Region. – Faunistische Abhandlungen Museum Tierkunde Dresden 11: 151-155.

Seifert, B. 1988: A revision of the European species of the ant subgenus *Chthonolasius*. – Entomologische Abhandlungen Museum Tierkunde Dresden 51: 143-180.

Seifert, B. 1991: *Lasius platythorax* n.sp., a widespread sibling species of *Lasius niger*. – Entomologia Generalis 16: 69-81.

Seifert, B. 1992: A taxonomic revision of the Palaearctic members of the ant subgenus *Lasius* s.str. (Hymenoptera: Formicidae). – Abhandlungen und Berichte des Naturkundemuseums Görlitz 66: 1-67.

Seifert, B. 1999: Interspecific hybridisations in natural populations of ants by example of a regional fauna (Hymenoptera: Formicidae). – Insectes Sociaux 46: 45-52.

Seifert, B. 2000: A taxonomic revision of the ant subgenus *Coptoformica* Mueller, 1923. – Zoosystema (Paris) 22: 517-568.

Seifert, B. 2003a: The ant genus *Cardiocondyla* (Insecta: Hymenoptera: Formicidae) – a taxonomic revision of the *C. elegans*, *C. bulgarica*, *C. batesii*, *C. nuda*, *C. shuckardi*, *C. stambuloffii*, *C. wroughtonii*, *C. emeryi*, and *C. minutior* species groups. – Annalen des Naturhistorischen Museums Wien, Serie B, 104: 203-338.

Seifert, B. 2003b: A taxonomic revision of the *Formica cinerea* group (Hymenoptera: Formicidae). – Abhandlungen und Berichte des Naturkundemuseums Görlitz 74: 245-272.

Seifert, B. 2004: *Hypoponera punctatissima* (Roger) and *H. schauinslandi* (Emery) – two morphologically and biologically distinct species (Hymenoptera: Formicidae). – Abhandlungen und Berichte des Naturkundemuseums Görlitz 75: 61-81.

Seifert, B. 2005: Rank elevation in two European ant species: *Myrmica lobulicornis* Nylander, 1857, stat.n. and *Myrmica spinosior* Santschi, 1931, stat.n. (Hymenoptera: Formicidae). – Myrmecologische Nachrichten 7: 1-7.

Seifert, B. 2006a: *Temnothorax saxonicus* (Seifert, 1995) stat. n., comb.n. – a parapatric, closely-related species of *T. sordidulus* (Müller, 1923) comb.n. and description of two new closely-related species, *T. schoedli* sp.n. and *T. artvinense* sp.n., from Turkey (Hymenoptera: Formicidae). – Myrmecologische Nachrichten 8: 1-12.

Seifert, B. 2006b: Social cleptogamy in the ant subgenus *Chthonolasius* – survival as a minority. – Abhandlungen und Berichte des Naturkundemuseums Görlitz 77: 251-276.

Seifert, B. 2007: Die Ameisen Mittel- und Nordeuropas. – lutra, Tauer, 368 pp.

Seifert, B. 2008: Removal of allometric variance improves species separation in multi-character discriminant functions when species are strongly allometric and exposes diagnostic characters. – Myrmecological News 11: 91-105.

Seifert, B. 2009: Cryptic species in ants (Hymenoptera: Formicidae) revisited: We need a change in the alpha-taxonomic approach. – Myrmecological News 12: 149-166.

Seifert, B. 2011: A taxonomic revision of the Eurasian *Myrmica salina* species complex (Hymenoptera: Formicidae). – Soil Organisms 83: 169-186.

Seifert, B. 2012a: Clarifying naming and identification of the outdoor species of the ant genus *Tapinoma* Förster, 1850 (Hymenoptera: Formicidae) in Europe north of the Mediterranean region with description of a new species. – Myrmecological News 16: 139-147.

SEIFERT, B. 2012b: A review of the West Palaearctic species of the ant genus *Bothriomyrmex* EMERY, 1869 (Hymenoptera: Formicidae). – Myrmecological News 17: 91-104.

SEIFERT, B. & GOROPASHNAYA, A. 2004: Ideal phenotypes and mismatching haplotypes – errors of mtDNA treeing in ants (Hymenoptera: Formicidae) detected by standardized morphometry. – Organisms, Diversity & Evolution 4: 295-305.

SEIFERT, B., KULMUNI, J. & PAMILO, P. 2010: Independent hybrid populations of *Formica polyctena* X *rufa* wood ants (Hymenoptera: Formicidae) abound under conditions of forest fragmentation. – Evolutionary Ecology 24: 1219-1237.

SEIFERT, B., SCHLICK-STEINER, B.C. & STEINER, F.M 2009: *Myrmica constricta* KARAVAJEV, 1934 – a cryptic sister species of *Myrmica hellenica* FINZI, 1926 (Hymenoptera: Formicidae). – Soil Organisms 81: 53-76.

SEIFERT, B. & SCHULTZ, R. 2009a: A taxonomic revision of the *Formica rufibarbis* FABRICIUS, 1793 group (Hymenoptera: Formicidae). – Myrmecological News 12: 255-272.

SEIFERT, B. & SCHULTZ, R. 2009b: A taxonomic revision of the *Formica subpilosa* RUZSKY, 1902 group (Hymenoptera: Formicidae). – Myrmecological News 12: 67-83.

SEPPÄ, P., HELANTERÄ, H., TRONTTI, K., PUNTTILA, P., CHERNENKO, A., MARTIN, S.J. & SUNDSTRÖM, L. 2011: The many ways to delimit species: hairs, genes and surface chemistry. – Myrmecological News 15: 31-41.

STEINER, F.M., SCHLICK-STEINER, B.C. & SEIFERT, B. 2009: Morphology-based taxonomy is essential to link molecular research to nomenclature. – Contributions to Natural History 12: 1295-1315.

STEINER, F.M., SEIFERT, B., GRASSO, D.A., LE MOLI, F., ARTHOFER, W., STAUFFER, C., CROZIER, R.H. & SCHLICK-STEINER, B.C. 2011: Mixed colonies and hybridisation of *Messor* harvester ant species (Hymenoptera: Formicidae). – Organisms Diversity & Evolution 11: 107-134.

STEINER, F.M., SEIFERT, B., MODER, K. & SCHLICK-STEINER, B.C. 2010: A multisource solution for a complex problem in biodiversity research: description of the cryptic ant species *Tetramorium alpestre* sp.n. (Hymenoptera: Formicidae). – Zoologischer Anzeiger 249: 223-254.

STRUYF, A., MIA, H. & ROUSSEEUW, P.J. 1997: Integrating robust clustering techniques in S-PLUS. – Computational Statistics & Data Analysis 26: 17-37.

TABACCHI, M., MANTEGAZZA, R., SPADA, A. & FERRERO, A. 2006: Morphological traits and molecular markers for classification of *Echinochloa* species from Italian rice fields. – Weed Science 54: 1086-1093.

TUKEY, J.W. 1977: Exploratory Data Analysis. – Addison Wesley Longman Inc., Reading, MA, 688 pp.

WARD, J.H. 1963: Hierarchical grouping to optimize an objective function. – Journal of the American Statistical Association 48: 236-244.

WARD, P.S. & SUMNICHT, T.P. 2012: Molecular and morphological evidence for three sympatric species of *Leptanilla* (Hymenoptera: Formicidae) on the Greek island of Rhodes. – Myrmecological News 17: 5-11.

WIENS, J.J. 2000: Reconstructing phylogenies from allozyme data: comparing method performance with congruence. – Biological Journal of the Linnean Society 70: 613-632.