

## METHODS AND TECHNIQUES

## Discrimination and identification of morphotypes of *Banksia integrifolia* (Proteaceae) by an Artificial Neural Network (ANN), based on morphological and fractal parameters of leaves and flowers

Camilla Pandolfi<sup>1</sup>, Giuseppe Messina<sup>1,2</sup>, Sergio Mugnai<sup>1</sup>, Elisa Azzarello<sup>1</sup>, Elisa Masi<sup>1</sup>, Kingsley Dixon<sup>2</sup> & Stefano Mancuso<sup>1</sup>

<sup>1</sup> Department of Horticulture, University of Florence, viale delle Idee 30, 50019 Sesto Fiorentino (FI), Italy. [sergio.mugnai@unifi.it](mailto:sergio.mugnai@unifi.it) (author for correspondence)

<sup>2</sup> School of Plant Biology, University of Western Australia, 35 Stirling Highway, Crawley, WA 6907, Australia

An artificial neural network (back propagation neural network) based on morphological and fractal leaf and flower parameters was developed for the characterization of three *Banksia integrifolia* subspecies and the identification of nine unnamed morphotypes. Results indicated that the network can be effectively and successfully used to discriminate among morphotypes using simple dedicated instruments, such as a PC and an optical scanner. The new method also as a complementary approach to botanical identification, being capable of separating all the tested *Banksia integrifolia* accessions and of creating associations between the known subspecies and the unnamed accessions.

**KEYWORDS:** BPNN, image analysis, plant identification, taxonomy

### INTRODUCTION

Taxonomic identification by the use of “taxonomic keys” is traditionally based on a paper-based type of expert system, often involving a series of choices from successive groups of contrasting statements, culminating in a name (Gaston & O’Neill, 2004). Each statement describes the state of at least one morphological character or attribute. The effectiveness and accuracy of this type of identification is usually linked to the experience of the expert who compiled it but is also time consuming and affected by the environment that can alter morphological characters. Moreover, this approach may not fully discriminate where character states involve only subtle morphological differences. Characterisation of subtle character differences such those found in cultivar identification of plants has been performed by the use of bio-molecular techniques (chemotaxonomy, Eder & al., 1994; nucleic acids, Żebrowska & Tyrka, 2003; analysis of isozymes, Arzate-Fernández & al., 2005) but, though effective, are often resource and labour-intensive, and require skilled and experienced operators to be effectively exploited. A new approach for plant identification has been recently developed by the introduction of artificial neural networks (ANNs) (Clark & Warwick, 1998; Mancuso & al., 1998; Bari & al., 2003; Mugnai & al., 2007), which are information processing paradigms modelled as biological nervous systems, composed of a large number of highly interconnected processing elements (akin neurons) working in unison to solve specific problems (Veelenturf, 1995). An

ANN, like the human mind, learns by examples and is configured for a specific application, such as pattern recognition or data classification.

Fractal scaling is evident in natural objects from the micro-scale to the macro-scale; investigations show that non-fractal objects are the exception rather than the rule in many natural systems (West & al., 1999). Mancuso (1999) highlighted the value of fractal geometry of leaves for analysis of patterns in leaf morphology. For this reason, leaf fractal parameters were used in this study and considered as useful parameters for plant discrimination and identification.

The case study presented here involves the common and widespread Australian *Banksia integrifolia* L. f., a taxon that demonstrates a remarkably high level of morphological intraspecific variation (Thiele & Ladiges, 1994). *Banksia* is an iconic Australian genus that is widespread both in eastern and western parts of Australia and comprises 80 species, including trees, shrubs and groundcovers and growing in swamps, coastal habitats, forest and desert habitats. *Banksia integrifolia* is a highly variable species consisting of four subspecies: (1) subsp. *aquilonia* (A.S. George) K.R. Thiele, with long, narrow, acute leaves that are spirally arranged and with follicles that are typically slightly larger than those of the other varieties (George, 1981); (2) subsp. *integrifolia*, with short, usually obtuse, dull-green leaves; (3) subsp. *compar* (R. Br.) K.R. Thiele, with large, glossy leaves with undulate margins and a relatively obtuse tip; and (4) subsp. *monticola* K.R. Thiele, which was considered a montane

form of *B. integrifolia* for many years but recently promoted to subspecies rank (Thiele & Ladiges, 1994) due to differences in leaf shape and fruiting structure. Classical printed taxonomic keys based on leaf parameters and flower characteristics have already been used for the identification of *B. integrifolia* (Thiele & Ladiges, 1994) but, to date, there are no known computer-based identification systems relating to this species.

The aims of the work presented here are: (1) to perform univocal discrimination of twelve *B. integrifolia* morphotypes grown under uniform and controlled conditions by morphological and fractal characterization of leaves and flowers through image analysis; (2) to associate nine unknown morphotypes to three known *B. integrifolia* subspecies (subsp. *integrifolia*, subsp. *monticola*, subsp. *compar*) by the development of a dedicated ANN; (3) to attempt the development of a phenetic clustering of the tested *Banksia* accessions by the construction of a phenetic tree based on leaf and flower morphological and fractal characters.

## MATERIALS AND METHODS

**Plant material.** — Plant material was collected from a living collection located on The Banksia Farm (Mount Barker, WA, Australia, 34°38'15" S, 117°38'58" E). The selected morphotypes belonged to three *Banksia integrifolia* subspecies (subsp. *integrifolia*, subsp. *monticola*, subsp. *compar*) and nine unknown accessions that exhibited differences in habit, leaf and flower shapes. Voucher specimens of the subspecies are lodged in the herbarium of Kings Park Botanical Garden (KPBG) of Perth, Australia. Eighty leaves per accession were collected at random from healthy one-year-old branches on five mature plants, and split into two groups: forty for the training phase of ANN construction and the other forty for the validation phase (see below). Four inflorescences per accession were randomly collected, two at the fully open flower stage and the other two at 50% open flower stage. Eighty flowers per accession were selected to derive styles and styler hooks (i.e., style long and wiry protruding from the slit in the perianth and curving out in hook shape, typical of some *Banksia* species) for both the training and the validation phases of the ANN.

**Image acquisition and determination of morphometric parameters.** — A desktop optical scanner, set at 300 × 300 dpi, 16 million colours, was used to acquire leaf (Fig. 1) and flower (styles and styler hooks) images (Fig. 2). Fourteen morphometric parameters (Table 1) were determined in both leaves and flowers for each image through an image analysis software (UTHSCSA Image Tool 3.0). The identification of a single leaf/flower by a certain number of objective parameters, instead of

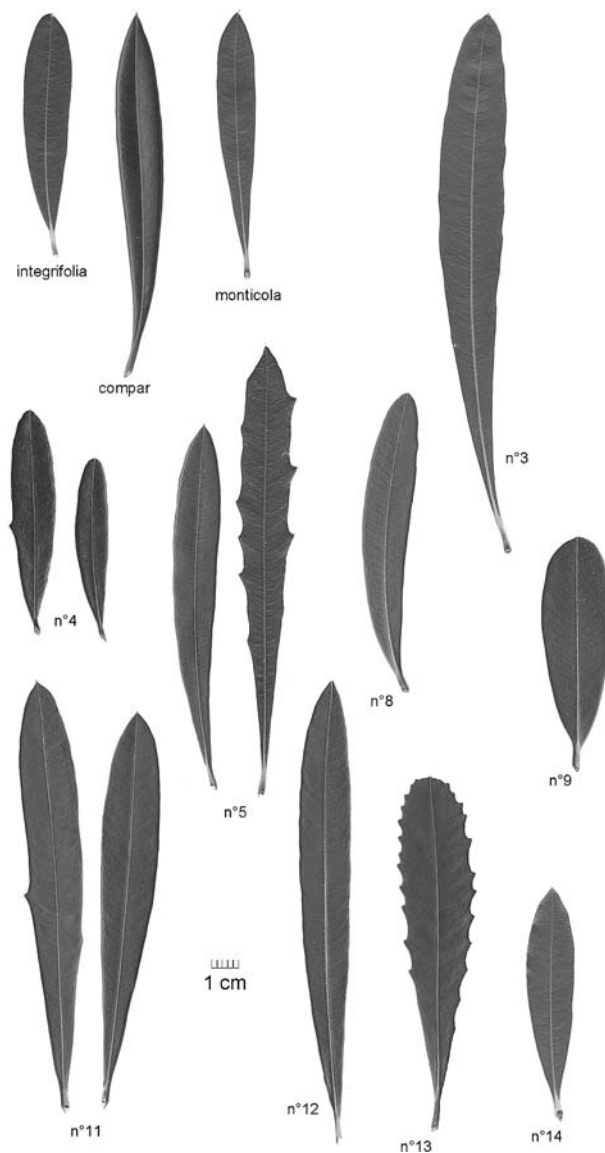


Fig. 1. Leaf images of the selected morphotypes, belonging to three *Banksia integrifolia* subspecies (subsp. *integrifolia*, subsp. *monticola*, subsp. *compar*) and nine unknown accessions differing in habit, leaf and flower shape.

using the entire and complete digital image *per se*, permitted to avoid the problem of the ‘curse of dimensionality’ (Bellman, 1961).

**Fractal geometry and fractal parameters.** — Leaf fractal parameters were calculated by a fractal image analysis software (HarFA, Harmonic and Fractal Image Analyzer 4.9.1). The leaf fractal spectrum was obtained using the method previously described by Mancuso (2002). In brief (Fig. 3, an output of HarFA), each leaf image was split into the three constituent colour channels (red, green, blue); each channel was set for a threshold colour value between 0 and 255 and the fractal dimension ( $D$ ) for each

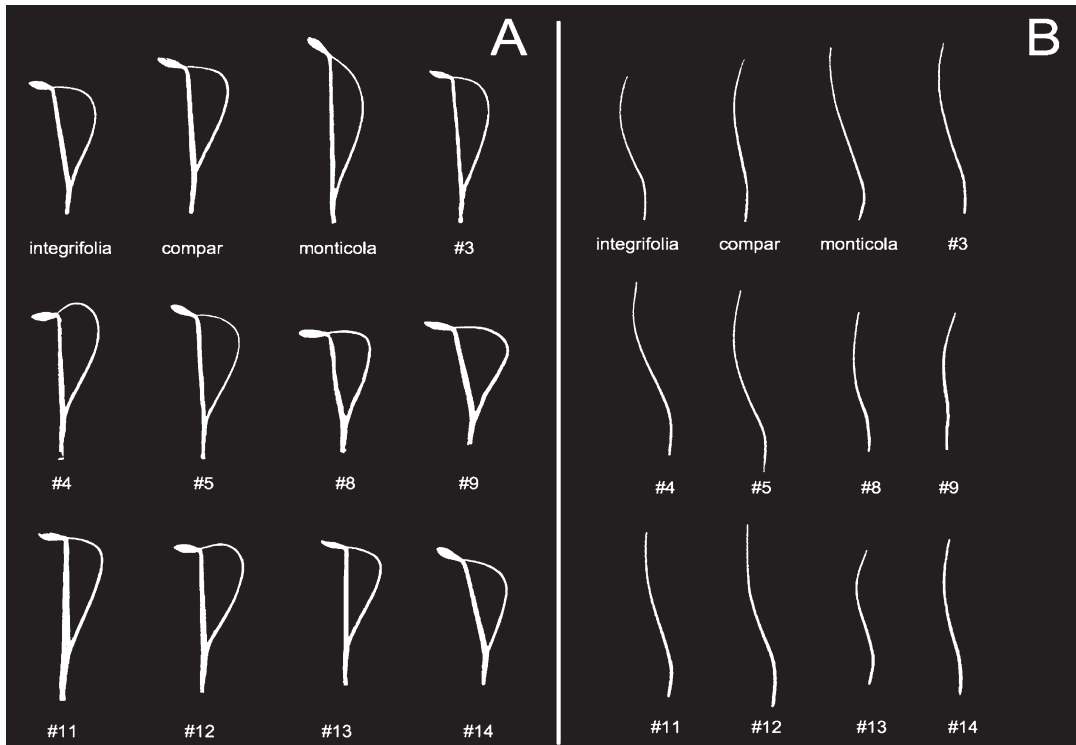


Fig. 2. Hooks (A) and style (B) images of the selected morphotypes.

Table 1. Morphological parameters of *Banksia integrifolia* organs (leaves, styles, flower hooks) calculated by the image analysis software.

| Parameter            | Definition   |
|----------------------|--|
| 1 Area               | Area of the organ  |
| 2 Perimeter          | Perimeter of the organ   |
| 3 Major axis length  | Length of the longest line that can be drawn through the organ                                 |
| 4 Minor axis length  | Length of the longest line that can be drawn through the organ perpendicular to the major axis |
| 5 Roundness          | Computed as: $(4 \times \pi \times \text{area}) / \text{perimeter}^2$                          |
| 6 Elongation         | Ratio of the length of the major axis to the length of the minor axis                          |
| 7 Feret diameter     | Diameter of a circle having the same area as the organ   |
| 8 Compactness        | Computed as: $\text{sqrt}(4 \times \text{area} / \pi) / \text{major axis length}$              |
| 9 Integrated density | Product of the mean grey level and the number of pixels in the image                           |
| 10 Min grey level    | Minimum grey level of the organ  |
| 11 Mean grey level   | Mean grey level of the organ   |
| 12 Median grey level | Median grey level of the organ   |
| 13 Mode grey level   | Mode grey level of the organ   |
| 14 Max grey level    | Maximum grey level of the organ  |

colour value was calculated. In fractal geometry,  $D$  is a statistic that provides an indication of how completely a fractal appears to fill space, as one zooms down to finer and finer scales. In our case,  $D$  was assessed using the box-counting method (Mancuso & al., 2003) and plotted against the colour intensity to obtain the fractal spectra

of the three channels. After drawing the baseline ( $D = 1$ ) that separates the fractal ( $>1$ ) from the non-fractal ( $<1$ ) zone of the spectrum, five fractal parameters (First X, Last X, Peak coordinates X and Y, and Total peak area) were calculated for each colour channel (Fig. 4). As previously reported by Mancuso & al. (2003), the green and

red channels were directly influenced by the phenotype, shifting their respective spectra to the left or to the right according to the different accessions (Fig. 5). As the blue channel seemed relatively unaffected, only green and the red channel results were selected as informative for the constitution of the neural network.

**Construction of the back propagation neural networks (BPNNs).** — A multi-layer neural network (MLNN) with the generalized delta rule for learning by

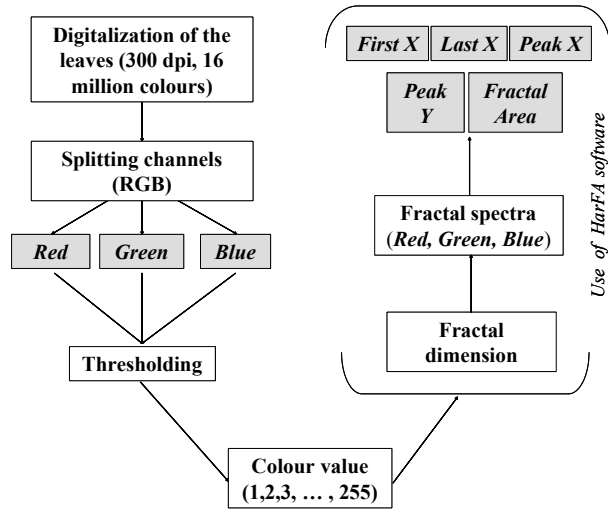


Fig. 3. Schematic diagram of the experimental protocol used to assess the application of fractals to the image analysis of leaves.

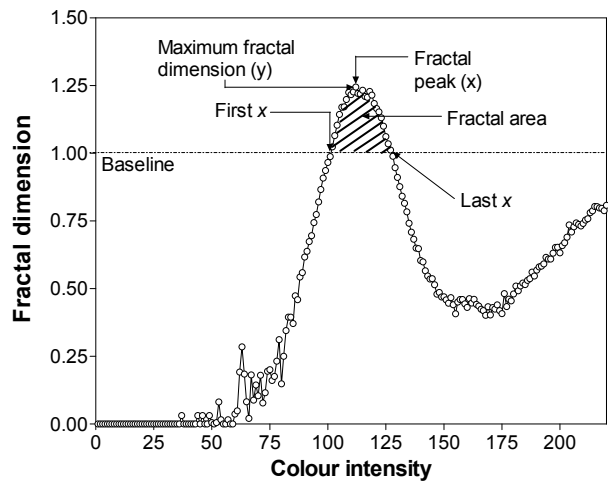


Fig. 4. Graphical representation of the five fractal parameters calculated from each colour channel: First X, Last X, Peak coordinates X and Y, and Total peak area. The baseline with the fractal dimension  $D = 1$  separates the fractal ( $>1$ ) from the non-fractal ( $<1$ ) zone of the spectrum. The figure represents an example of the output given by the software used for the fractal image analysis (HarFA, Harmonic and Fractal Image Analyzer 4.9.1).

a back-propagation learning algorithm (Rumelhart & al., 1986) is an effective system for learning discriminants for classes from a set of examples (Sejnowski & Rosenberg, 1987). It has been recently introduced as a useful method for plant identification (Mancuso & al., 1998; Clark & Warwick, 1998; Mancuso & Nicese, 1999): BPNN learns the class knowledge directly from the training dataset and, therefore, it is unnecessary to make any assumptions regarding the underlying probability density functions. Information about *a priori* probability can be adjusted after training (Hush & Horne, 1993), or by increasing the number of training patterns. Moreover, there is no need for retaining the training data and no extensive computation is involved in the classification of unknown patterns. BPNN

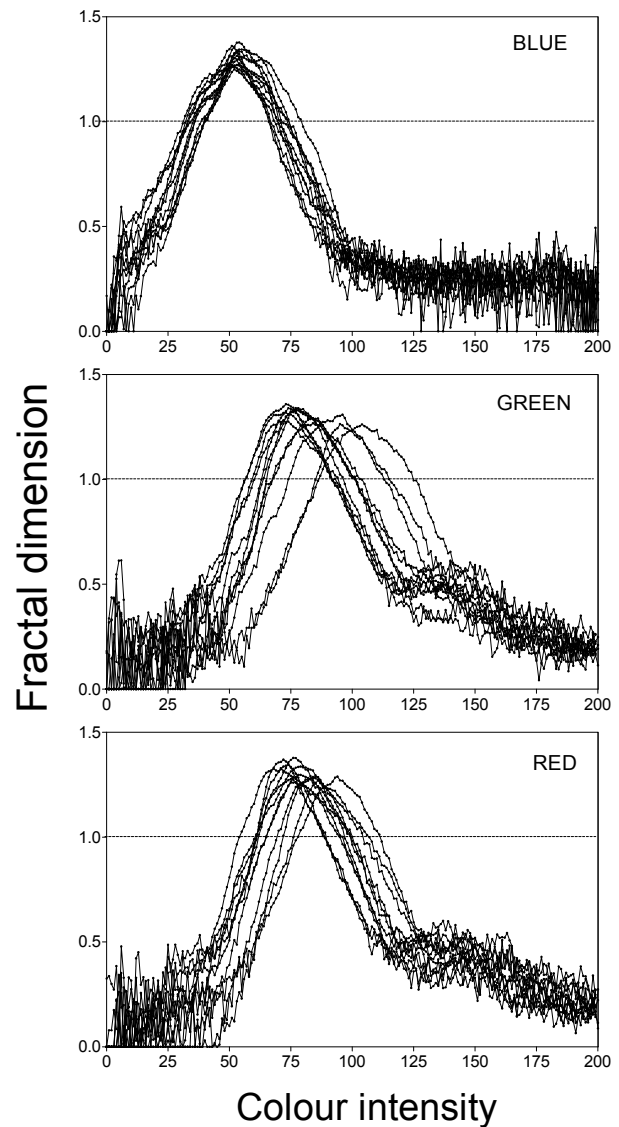


Fig. 5. Fractal spectra of the blue, green and red channels derived from the collected leaves belonging to the *Banksia integrifolia* accessions.

algorithm has also low memory requirements compared to other algorithms (Zhang & al., 2007) and usually reaches an acceptable level of error in a short time.

In details, a BPNN is an iterative gradient algorithm derived from a multilayer feed-forward network, or multilayer perceptron (MLP) and composed of some layers of neurons: an input layer, one or more hidden layers, and an output layer. The input layer collects the information about the system and passes them to the hidden layer(s) which processes the information initiated at the input. At the end, the output layer gets the observable response. The hidden layer(s) connects the input pattern  $x$  with the output pattern  $y$  through a series of interconnected weights (Eq. 1):

$$u = \sum_{i=0}^n x_i w_i \quad [1]$$

where  $u$  is the aggregated input signal;  $x_i$  is the input pattern;  $w_i$  is the weight of the  $i^{\text{th}}$  input vector which is connected to  $i^{\text{th}}$  processing element. Particular nodes were also used to shift the neuron transfer function and to improve the network performance, thanks to the backpropagation of errors (Rumelhart & al., 1986).

The most frequently used adjustment of weights is a simple gradient descent: in this case, each weight adjusts by a small amount proportional to the derivative of the error function ( $\delta E / \delta w_{ij}$ ) with respect to that weight, and in the opposite direction (Eq. 2):

$$\Delta w_{ij} = -\varepsilon(\delta E / \delta w_{ij}) \quad [2]$$

where  $\Delta w_{ij}$  is the adjustment of the weight  $w_{ij}$  connecting the unit  $i$  of one layer with the unit  $j$  of the following layer, while  $\varepsilon$  is the learning rate (value of 0.1).

In the present study a custom-developed algorithm for the neural network construction was developed, and a strict gradient descent (Rumelhart & McClelland, 1988) was used with the introduction of an averaging term (Eq. 3):

$$\Delta w_{ij} (\text{iteration } n) = -\varepsilon(\delta E / \delta w_{ij}) + \mu \Delta w_{ij} (\text{iteration } n-1) \quad [3]$$

where the coefficient  $\mu$  is referred as the momentum (value of 0.3).

In order to obtain a model with validation capabilities, the collected 80 leaves per accession were split into two datasets. A training set of 40 leaves was needed to build the neural model, which implies that the processing elements were able to change the output in response to the input change by adjusting the connecting weights  $w_i$  (Eq. 1). A validation set, composed by the other 40 leaves, was used to verify the correctness of the model.

The training phase was considered complete when the ANN achieved the desired statistical accuracy as it produced the required outputs for a given sequence of inputs. The correct network structure was created after

stopping the learning process when the root mean square error (RMS error, Eq. 4) was minimized:

$$\text{RMS error} = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \text{out}_{ij})^2}{N \times M}} \quad [4]$$

where  $y_{ij}$  is the element of the matrix ( $N \times M$ ) for the training set, and  $\text{out}_{ij}$  is the element of the output matrix ( $N \times M$ ) of the neural network ( $N$  is the number of variables in the matrix and  $M$  is the number of samples). The training phase was iterated until the RMS error became  $<0.06$ , and the difference between the RMS in two consecutive periods was  $<0.0001$ .

The validation test was also critical in order to verify that the network did not simply memorize the training set but learned the general patterns involved within an application. At this stage, input data derived from the validation set were put into the constructed ANN and the predicted outputs were then evaluated. The predicted and the experimental values were then compared to measure the network performance. When the validation test is well performed, unknown data can be evaluated and the relative outputs are predicted.

In this study, leaf and flower parameters obtained from the image analysis were used as inputs during both the training and the validation phases as independent variables, while the twelve accessions of *Banksia integrifolia* represented the outputs. From preliminary tests, 40 was considered the minimum significant number of leaves/flowers per phase. In fact, factors in the hidden layer, such as learning scheme, numbers of nodes of the output and input and connections between them play an important role (Zurada, 1992). The determination of the best configuration for the ANN was performed by a preliminary test among several hierarchical ANNs. The minimum error was achieved using one hidden layer composed by 50 nodes (or neurons). In Eq. 1,  $u$  must be further processed by an activation function, which determines the process inside the neurons. For our purposes, the hidden layer was activated by a logistic sigmoid function (Eq. 5), which allows the representation of non-linear relationship (Montague & Morris, 1994):

$$y = \frac{1}{(1 + e^{-x})} \quad [5]$$

The node's activation function (except for the input layer, which uses the input itself) controls the output signal strength for the unit. These sigmoid functions set the output signal strength between 0 and 1, acting like an output gate that can be opened (1) or closed (0). As the function is continuous, it is also possible for the gate to be partially

opened (i.e., output signal strength between 0 and 1). In an ideal case only the class of the output representing a given accession would show a value of 1 (correct identification) while all the other classes would show the value 0 (incorrect identification). However, due to the natural variation among leaves, the output of the expected class usually tends to report a value <1.

Two artificial neural networks were built in this experimental work. The first BPNN was based only on the leaf morphometric and fractal parameters inserted in the training phase as input signals. Flower (styles and stylar hooks) morphometric parameters were added later to the network as additional input signals together with the previous leaf parameters, in order to build a more informative, complete and discriminating network. Neural network outputs from the second BPNN were then used to measure the dissimilarities or distances between accessions when forming the clusters inside a dedicated phenetic tree.

**Data analysis.** — Morphometric parameters were subjected to one-way ANOVA and their means separated by Tukey’s Multiple Comparison Test ( $n = 30$ ,  $P < 0.05$ ; Fowler & al., 1998). To obtain a similarity matrix among accessions, the COSINE pattern similarity measure was performed (Bartish & al., 1999). A cluster analysis based on the unweighted pair group method with arithmetic averages (UPGMA, Sokal & Michener, 1958) was performed using this matrix, and a phenetic tree with Euclidean distances showing how the accessions clustered was produced by NTSYSpc 2.2 (Exeter Software).

**Table 2. The most informative leaf morphometric parameters obtained by leaf analysis. Data are reported as means ( $n = 40$ ) and sorted from the highest to the lowest leaf area values. Means were separated by Tukey’s test: different letters show statistically significant differences for  $P < 0.05$ .**

| Morpho-type  | Area [mm <sup>2</sup> ] | Perimeter [mm] | Major axis [mm] | Minor axis [mm] | Elongation |
|--------------|-------------------------|----------------|-----------------|-----------------|------------|
| 3            | 2,967 a                 | 420 a          | 192 a           | 24 bc           | 7.93 b     |
| 5            | 2,645 a                 | 400 a          | 181 a           | 22 cd           | 8.31 ab    |
| 13           | 2,486 a                 | 340 b          | 144 c           | 28 a            | 5.08 e     |
| 12           | 2,416 b                 | 381 b          | 174 b           | 20 e            | 8.57 a     |
| 11           | 2,234 b                 | 343 b          | 159 bc          | 20 de           | 7.67 bc    |
| 8            | 2,091 bc                | 296 c          | 135 d           | 22 de           | 6.20 d     |
| compar       | 2,013 bc                | 325 b          | 149 bc          | 20 de           | 7.28 c     |
| 9            | 1,901 c                 | 244 d          | 106 ef          | 26 b            | 4.06 f     |
| integrifolia | 1,430 d                 | 244 d          | 103 f           | 20 e            | 5.16 e     |
| monticola    | 1,266 de                | 264 cd         | 116 e           | 16 f            | 6.99 c     |
| 14           | 944 ef                  | 193 d          | 83 g            | 17 f            | 4.88       |
| 4            | 777 f                   | 180 e          | 81 g            | 13 g            | 6.06 d     |

## RESULTS

The most informative leaf (Table 2), hooks (Table 3) and styles (Table 4) morphometric parameters enabled a preliminary discrimination of the tested morphotypes, so permitting the creation of rough groups of similarity. The discrimination based on the leaf area indicated that accessions 3, 5 and 13, which showed the greatest values, should be grouped together, whereas the smallest values found in 4 and 14 should compose another group of similarity. Leaf perimeter almost followed the leaf area trend, except for 13 that, despite the highest leaf area, showed a reduced perimeter compared to 3 and 5. The other two parameters (leaf major and minor axes) provided other levels of discrimination based on leaf shape. The same conclusions could be obtained by using flower parameters. Similarities among the morphotypes were analysed and distinguished by the construction of the dedicated ANNs. The BPNN outputs can be graphically represented by a XY-graph for each accession, with the accession names on the x-axis, and the y-axis reporting the output signal strength. Each graph aims to show how the BPNN was able to discriminate the selected accession in comparison with the others.

The first BPNN was built using only leaf morphometric and fractal parameters: in this case, the network almost completely discriminated the relative morphotypes which presented an output value higher than 0.5, which is considered the threshold value for a successful discrimination in plant species identification processes (Pandolfi & al., 2006), with the exception of morphotypes

**Table 3. The most informative hook morphometric parameters obtained by flower image analysis. Data are reported as means ( $n = 40$ ) and sorted from the highest to the lowest leaf area values. Means were separated by Tukey’s test: different letters show statistically significant differences for  $P < 0.05$ .**

| Morpho-type  | Area [mm <sup>2</sup> ] | Perimeter [mm] | Major axis [mm] | Minor axis [mm] | Elongation |
|--------------|-------------------------|----------------|-----------------|-----------------|------------|
| 12           | 104.49 a                | 69.35 bc       | 25.48 d         | 8.84 a          | 2.99 e     |
| 11           | 101.72 a                | 72.82 a        | 28.50 b         | 6.35 d          | 4.56 bc    |
| 5            | 100.58 a                | 69.25 bc       | 27.01 c         | 6.91 c          | 3.92 d     |
| 9            | 92.50 b                 | 61.95 d        | 22.59 e         | 7.63 b          | 2.99 e     |
| 4            | 85.11 c                 | 67.49 c        | 25.66 d         | 9.49 a          | 2.72 e     |
| compar       | 83.91 cd                | 66.08 c        | 25.74 cd        | 6.06 de         | 4.27 cd    |
| 3            | 82.76 cd                | 66.96 c        | 26.36 cd        | 5.55 e          | 4.82 b     |
| monticola    | 81.91 cd                | 70.24 b        | 30.35 a         | 4.54 f          | 6.85 a     |
| integrifolia | 81.13 cd                | 58.57 e        | 22.07 e         | 6.14 de         | 3.73 d     |
| 8            | 79.18 cd                | 59.51 de       | 21.40 e         | 6.67 cd         | 3.27 de    |
| 14           | 76.29 de                | 60.82 de       | 24.67 d         | 5.48 e          | 4.60 bc    |
| 13           | 70.34 e                 | 62.32 d        | 23.94 d         | 5.44 e          | 4.47 bc    |

5, 11 and 12 (Fig. 6 in the online version of this article; Table 5). Among the discriminated morphotypes, the highest output value was shown by 4 (0.99), which means that this accession was completely and unequivocally discriminated by the dedicated BPNN. However, even though morphotype 5 reported an average output of 0.45, this result can lead to a successful discrimination because the output graph reported no concurrent and significant peaks in correspondence with other accessions. On the contrary, morphotypes 11 and 12 showed a very different behaviour, because both the accessions reported a concurrent and almost equally significant peak in its own graph corresponding to the other accession. This means that the first BPNN, exclusively based on leaf parameters, was not completely able to unequivocally discriminate between morphotypes 11 and 12, due to a strong morphological and fractal similarity in leaf shape.

In order to improve the effectiveness of the first BPNN, flower morphometric parameters were added as inputs to build a second, more discriminating and informative BPNN (Fig. 7 in the online version of this article). This improvement led to a more powerful network compared to the previous BPNN, with all the output values referring to the tested accessions higher than 0.9 (Table 5). This means a total, complete and unequivocal discrimination of any single morphotype.

To illustrate the ability of the second BPNN to discriminate among the accessions, a phenetic tree based on the morphological similarities among the different OTUs (*Banksia* morphotypes) and derived from the ANN results

**Table 5. Output level of similarity in the two dedicated BPNNs.**

|                     | Leaves | Leaves and flowers |
|---------------------|--------|--------------------|
| <i>integrifolia</i> | 0,84   | 0,97               |
| <i>compar</i>       | 0,85   | 0,98               |
| <i>monticola</i>    | 0,91   | 0,98               |
| 3                   | 0,77   | 0,88               |
| 4                   | 0,99   | 1,00               |
| 5                   | 0,45*  | 0,94               |
| 8                   | 0,80   | 0,98               |
| 9                   | 0,90   | 0,97               |
| 11                  | 0,43** | 0,95               |
| 12                  | 0,44** | 0,95               |
| 13                  | 0,72   | 0,97               |
| 14                  | 0,89   | 0,99               |

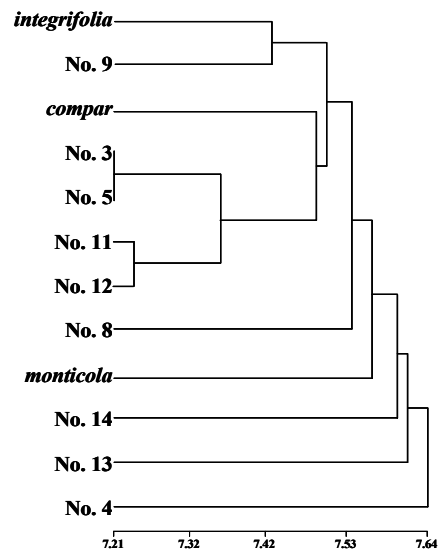
\* indicates a successful discrimination despite the low output value as no concurrent peaks were monitored.

\*\* indicates a unsuccessful discrimination as concurrent peaks were monitored.

was constructed (Fig. 8). As expected, a close cluster formed by morphotypes 11 and 12 was found, suggesting a strong morphological relationship between these two accessions. More morphotypes 3 and 5 were closely grouped together. All these accessions were associated with subsp. *compar*. Moreover, morphotype 9 was recognized as subsp. *integrifolia*. On the contrary, the ANN

**Table 4. The most informative styles morphometric parameters obtained by flower image analysis. Data are reported as means (n = 40) and sorted from the highest to the lowest leaf area values. Means were separated by Tukey's test: different letters show statistically significant differences for P < 0.05.**

| Morpho-type         | Area [mm <sup>2</sup> ] | Perimeter [mm] | Major axis [mm] | Minor axis [mm] | Elongation |
|---------------------|-------------------------|----------------|-----------------|-----------------|------------|
| <i>compar</i>       | 18.62 a                 | 69.51 c        | 32.03 c         | 0.68 a          | 47.32 c    |
| 12                  | 18.13 ab                | 71.97 ab       | 33.35 b         | 0.64 b          | 52.41 b    |
| 11                  | 18.06 ab                | 69.71 c        | 32.59 c         | 0.68 a          | 48.24 c    |
| 4                   | 17.88 b                 | 71.78 b        | 32.94 c         | 0.64 b          | 52.04 b    |
| <i>monticola</i>    | 17.06 c                 | 73.70 a        | 34.24 a         | 0.62 c          | 55.53 b    |
| 14                  | 16.84 cd                | 64.17 d        | 29.95 e         | 0.65 b          | 46.27 cd   |
| 3                   | 16.64 d                 | 68.15 c        | 31.51 c         | 0.62 c          | 51.12 b    |
| 9                   | 16.53 d                 | 60.08 e        | 27.76 f         | 0.70 a          | 39.89 f    |
| <i>integrifolia</i> | 15.46 e                 | 62.24 de       | 28.52 e         | 0.66 ab         | 43.63 e    |
| 5                   | 15.17 e                 | 72.68 a        | 33.22 bc        | 0.53 d          | 63.24 a    |
| 13                  | 14.18 f                 | 60.36 e        | 27.45 f         | 0.61 c          | 45.18 d    |
| 8                   | 14.16 f                 | 58.02 f        | 27.07 f         | 0.63 bc         | 43.17 e    |



**Fig. 8. Phenetic tree based on UPGMA analysis of morphological similarity estimates (COSINE similarity matrix derived from the second BPNN outputs, obtained using both leaf and flower parameters) among the tested morphotypes of *Banksia integrifolia*. The numbers at the x-axis report the Euclidean distances among the accessions.**

was not able neither to associate subsp. *monticola* with the unnamed accessions, nor to cluster morphotypes 4, 8, 13 and 14, which remained completely distinguished. In this case, a clear association between morphotypes 4, 8, 13, 14 and one of the subspecies was not unequivocally possible.

## DISCUSSION

A certain number of techniques for the identification of biological samples should be used to assign new specimens to taxonomic groups (Weeks & Gaston, 1997), such as linear discriminant analysis, but mostly of them are based on the restrictive assumptions of multivariate normality and homogeneity of variance-covariance matrices (Boddy & Morris, 1993). On the contrary, the use of ANN for biological species identification should be promoted and supported as more appropriate, because ANNs are powerful pattern-recognition and data-analysis tools which make no preliminary assumptions about data. Here we report that our BPNN can discriminate among different *Banksia* morphotypes, with the second BPNN being capable of successfully identify all the tested *Banksia integrifolia* morphotypes through the image analysis of both leaves and flowers. On the contrary, the use of leaf parameters alone failed to properly discriminate among morphotypes 11 and 12.

Similar positive results in the use of BPNN for plant identification were previously obtained in olive (Bari & al., 2003; Mancuso & Nicese, 1999), chestnut (Mancuso & al., 1999), *Rollinia* (Mariño & Tressens, 2001), grapevine (Mancuso & al., 1998), *Tilia* spp. (Clark, 2004) and *Camellia japonica* (Mugnai & al., 2007). In this study, the introduction of fractal parameters as quantitative input layers in our ANN revealed an improvement in the reliability of the system. The fractal spectrum has been recently used as a useful tool to assess the hardiness and cold tolerance of some Australian species such as *Callistemon* and *Grevillea* spp. (Mancuso & al., 2003, 2004), and was introduced as a botanical identification key by Bari & al. (2003) on *Olea europaea* L. and by Mugnai & al. (2007) on *Camellia japonica* L.

The BPNN was also a powerful tool to detect similarities among the morphotypes, as a phenetic tree based on cluster analysis of its result can be built. Cluster analysis from the second BPNN created groups composed by similar unknown morphotypes and a related *Banksia* subspecies (Fig. 8). For example, morphotype 9 was associated with subsp. *integrifolia*, whereas morphotypes 3, 5, 11 and 12 were identified as *B. integrifolia* subsp. *compar.* On the contrary, morphotypes 4, 8, 13 and 14 could not be unequivocally associated with any subspecies as their Euclidean distances did not permit a clear and unequivocal clustering. This level of dissimilarity did not permit

unequivocal association with any subspecies, probably due to an intrinsic high morphological variability inside these accessions. In fact, the success of a BPNN largely depends on the quantity, validity, and accuracy of training data, as neural networks are better able to train and learn to generalize when the presented data are rich in variation. In our case, the creation of an ANN based on both flower and leaves morphometric and fractal parameters can effectively and unequivocally recognize the tested morphotypes. It is well known that it is often easier to correctly identify a specimen to a higher taxon than it is to identify to the rank of subspecies. For this reason our results should be considered a promising approach to the correct identification of specimens, even if the association of the unknown accessions with the known subspecies appeared to be rather controversial. One of the main disadvantages of an ANN is the need of a suitable set of example data for the achievement of an excellent quality in the training set, as this operation can be usually defined as the key point of all the ANN building process. ANN should be considered applicable only to problems where there are ample data for network training, and an equal number of samples from each group in the training set in order to avoid imbalance in network training (Weeks & Gaston, 1997). Particular care should be directed to the choice of the plant material, which must be healthy, well-developed and representative of the specimens to be associated. Most studies of automated identification systems have employed training sets with a relatively small number of samples (5–10) per species (Gaston & O'Neill, 2005), but they must be ideally larger to improve the accuracy of identifications. The use of 40 leaves and flowers from five plants in the training set of the present work should have reduced the risk of a wrong estimation of the specimen variability.

The BPNN built on all the calculated parameters had the highest output values and the highest level of apparent success in discrimination of morphotypes. A major limitation of this approach is that plant material needs to include flowering material which necessarily restricts the times at which the analysis can be performed. In ecological and botanical studies it is generally important to be able to recognize the species *in situ*, and specimens with flowers are not always available. In our case, a network exclusively based on morphometric and fractal parameters of leaves can be effectively and successfully used, though not always unequivocally, to discriminate among *Banksia integrifolia* accessions, but the inclusion of flower morphometric parameters can lead to a more powerful, detailed and informative network, with a high discrimination capacity.

In conclusion, the application of a BPNN is proposed as a complementary method of botanical identification, being capable of separation of all the tested *Banksia integrifolia* accessions and of creating associations between known taxa (here: subspecies) and unnamed accessions.



## ACKNOWLEDGEMENTS

The authors would like to thank Kings Park Botanic Garden staff (Dr. Matthew Barrett, Russell Barrett, Mr. Bob Dixon) and the institution for hosting this project (providing plant material, scanner, travelling expenses) and for assistance with *Banksia* identification, and Mr. and Mrs. Collins, owners of 'The Banksia Farm', for access to their collection of *Banksia* species from eastern Australia. The authors are also indebted to the two anonymous reviewers who improved the quality of the manuscript with useful suggestions.

## LITERATURE CITED

- Arzate-Fernández, A.M., Mejía-González, C.O., Nakazaki, T., Okumoto, Y. & Tanisaka, T. 2005. Isozyme electrophoretic characterization of 29 related cultivars of lily (*Lilium* spp.). *Pl. Breed.* 124: 71–78.
- Bari, A., Martin, A., Boulouha, B., Gonzalez-Andujar, J. L., Barranco, D., Ayad, G. & Padulosi, S. 2003. Use of fractals and moments to describe olive cultivars. *J. Agric. Sci.* 141: 63–71.
- Bartish, I.V., Rumpunen, K. & Nybom, H. 1999. Genetic diversity in *Chaenomeles* (Rosaceae) revealed by RAPD analysis. *Pl. Syst. Evol.* 214: 131–145.
- Bellman, R. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton.
- Boddy, L. & Morris, C.W. 1993. Analysis of flow cytometry data—a neural network approach. *Binary* 5: 17–22.
- Clark, J.Y. 2004. Identification of botanical specimens using artificial neural networks. Pp. 87–94 in: Fogel, D. (ed.), *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE press, San Diego.
- Clark, J.Y. & Warwick, K. 1998. Artificial keys for botanical identification using a multilayer perceptron neural network (MLP). *Artificial Intelligence Rev.* 12: 105–115.
- Eder, R., Wendelin, S. & Barna, J. 1994. Classification of red wine cultivars by means of anthocyanin analysis. *Mitt. Klosterneuburg* 44: 201–212.
- Fowler, J., Cohen, L. & Jarvis, P. 1998. *Practical Statistics for Field Biology*, 2nd ed. Wiley, Chichester.
- Gaston, K.J. & O'Neill, M.A. 2005. Automated species identification: why not? *Philos. Trans., Ser. B* 359: 655–677.
- George, A. 1981. The genus *Banksia* L. f. (Proteaceae). *Nuytsia* 3: 239–473.
- Hush, D.R. & Horne, B.G. 1993. Progress in supervised neural network. *IEEE Signal Processing Mag.* 10: 8–39.
- Mancuso, S. 1999. Fractal geometry-based image analysis of grapevine leaves using the box counting algorithm. *Vitis* 38: 97–100.
- Mancuso, S. 2002. Discrimination of grapevine (*Vitis vinifera* L.) leaf shape by fractal spectrum. *Vitis* 41: 137–142.
- Mancuso, S., Pisani, P.L., Bandinelli, R. & Rinaldelli, E. 1998. Application of an artificial neural network (ANN) for the identification of grapevine genotypes. *Vitis* 37: 27–32.
- Mancuso, S. & Nicese, F.P. 1999. Identifying olive (*Olea europaea* L.) cultivars using artificial neural networks. *J. Amer. Soc. Hort. Sci.* 124: 527–531.
- Mancuso, S., Ferrini, F. & Nicese, F.P. 1999. Chestnut (*Castanea sativa* L.) genotype identification: an artificial neural network approach. *J. Hort. Sci. Biotechnol.* 74: 777–784.
- Mancuso, S., Nicese, F.P. & Azzarello, E. 2003. The fractal spectrum of the leaves as a tool for measuring frost hardness in plants. *J. Hort. Sci. Biotechnol.* 78: 610–616.
- Mancuso, S., Nicese, F.P. & Azzarello, E. 2004. Comparing fractal analysis, electrical impedance and electrolyte leakage for the assessment of cold tolerance in *Callistemon* and *Grevillea* spp. *J. Hort. Sci. Biotechnol.* 79: 627–632.
- Mariño, S.I. & Tressens, S.G. 2001. Artificial neural networks application in the identification of three species of *Rollinia* (Annonaceae). *Ann. Bot. Fenn.* 38: 215–224.
- Montague, G. & Morris, J. 1994. Neural-network contributions in biotechnology. *Trends Biotechnol.* 12: 312–324.
- Mugnai, S., Pandolfi, C., Azzarello, E., Masi, E. & Mancuso, S. 2007. *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. *Pl. Syst. Evol.* 270: 95–108.
- Pandolfi, C., Mugnai, S., Azzarello, E., Masi, E. & Mancuso, S. 2006. Fractal geometry and neural networks for the identification and characterization of ornamental plants. Pp. 213–225 in: Teixeira da Silva, J. (ed.), *Floriculture, Ornamental and Plant Biotechnology: Advances and Topical Issues*, vol. 4. Global Science Books, Kyoto.
- Rumelhart, D.E. & McClelland, J.L. 1988. *Exploration in Parallel Distributed Processing*. MIT Press, Cambridge.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986. Learning internal representations by error propagation. Pp. 213–222 in: Rumelhart, D.E. & McClelland, J.L. (eds.), *Parallel Distributed Processing*. MIT Press, Cambridge.
- Sejnowski, T.J. & Rosenberg, C. 1987. Parallel networks that learn to pronounce English text. *Complex Systems* 1: 145–168.
- Sokal, R.R. & Michener, C.D. 1959. A statistical method for evaluating systematic relationship. *Univ. Kansas Sci. Bull.* 28: 1409–1438.
- Thiele, K. & Ladiges, P.Y. 1994. The *Banksia integrifolia* L. f. species complex (Proteaceae). *Austral. Syst. Bot.* 7: 393–408.
- Veelenturf, L.P.J. 1995. *Analysis and Application of Artificial Neural Networks*. Prentice Hall, London.
- Weeks, P.J.D. & Gaston, K.J. 1997. Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. *Biodiv. Cons.* 6: 263–274.
- West, G.B., Brown, J.H. & Enquist, B.J. 1999. The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* 284: 1677.
- Żebrowska, J.I. & Tyrka, M. 2003. The use of RAPD markers for strawberry identification and genetic diversity studies. *Food Agric. Environm.* 1: 115–117.
- Zhang, Z., Wang, Y., Fan, G. & Harrington P.D.B. 2007. A comparative study of multilayer perceptron neural networks for the identification of rhubarb samples. *Phytochem. Anal.* 18: 109–114.
- Zurada, J.M. 1992. *Introduction to Artificial Neural Systems*. Ed. Saujanya Books, Mumbai.

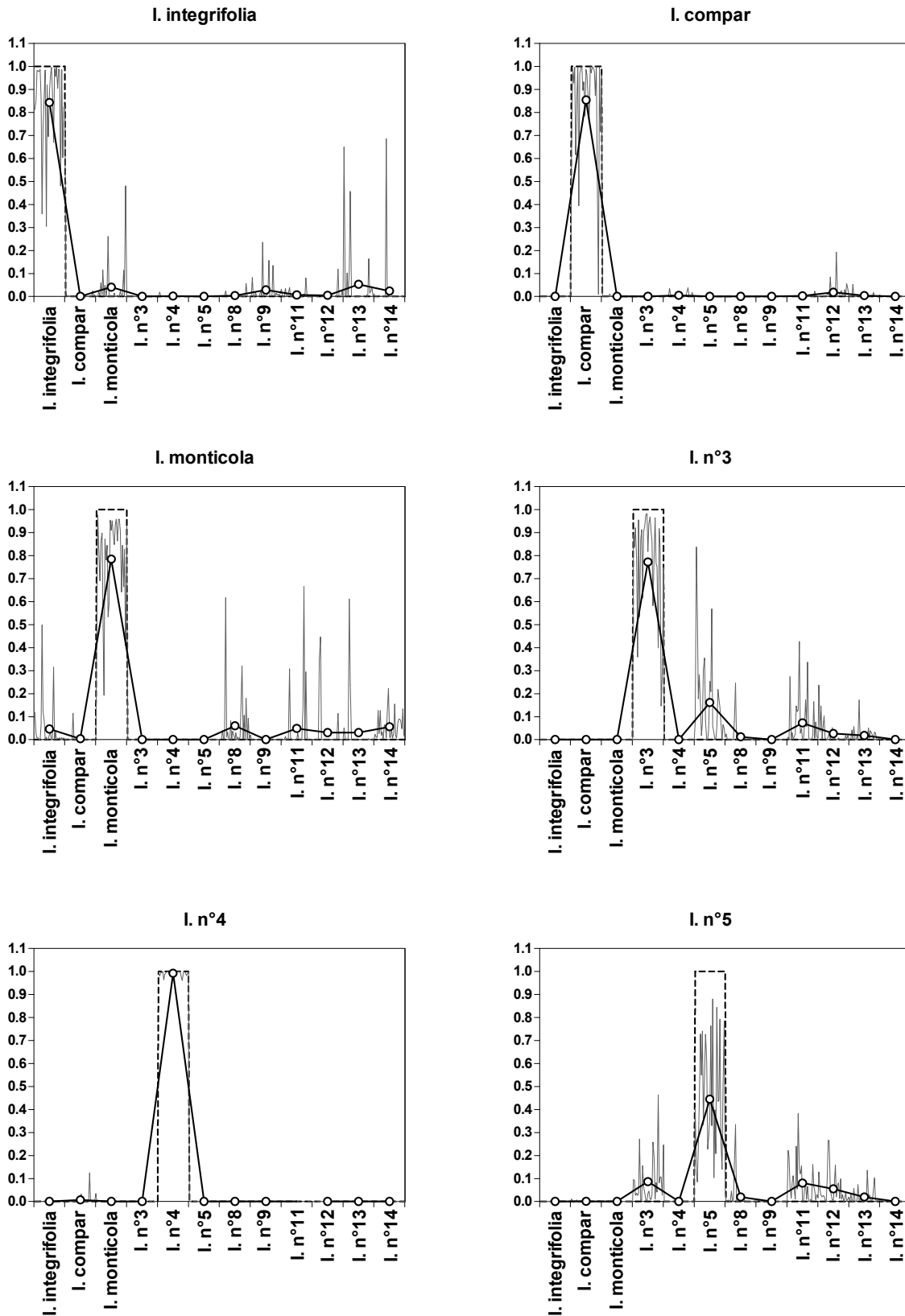
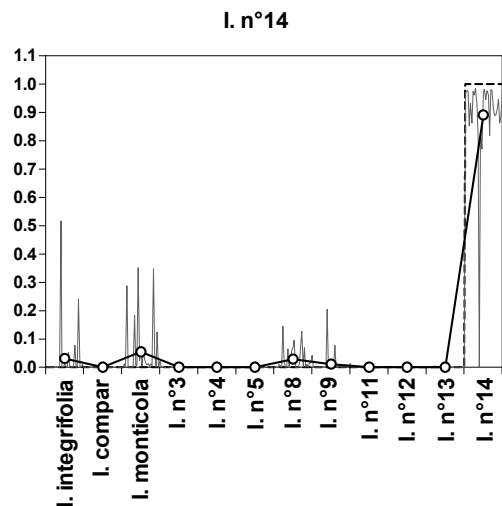
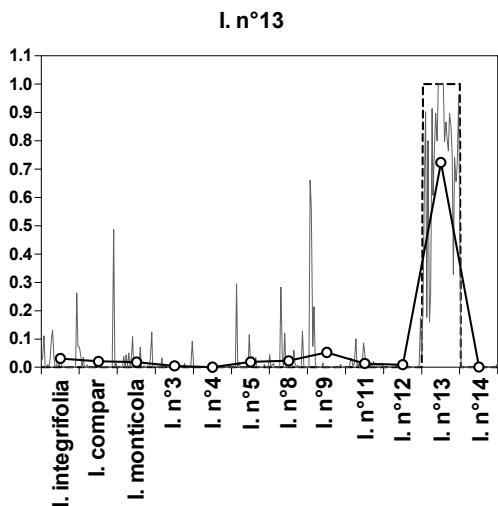
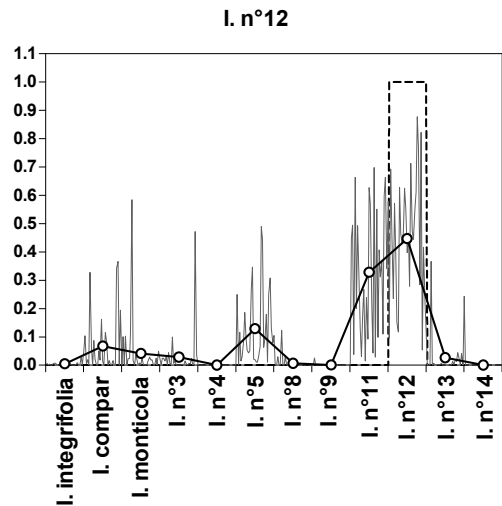
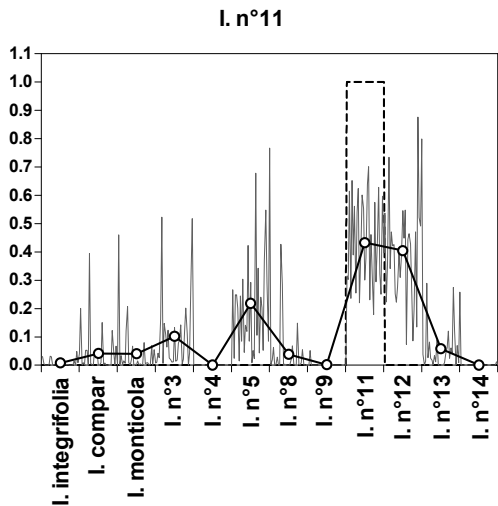
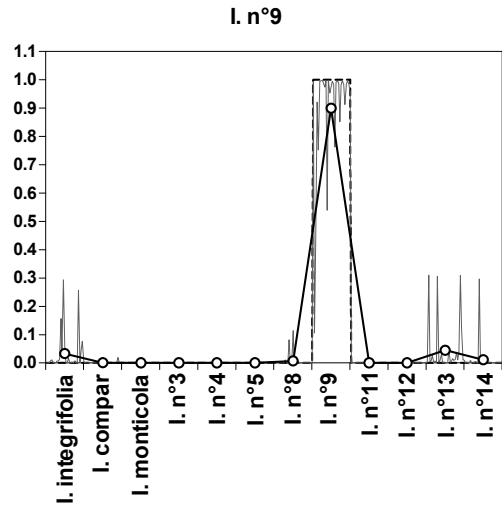
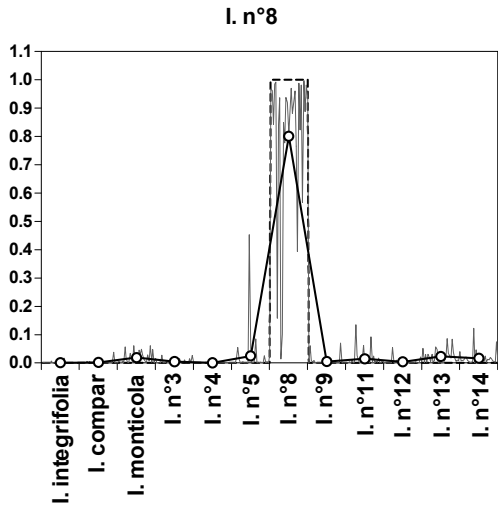


Fig. 6. Outputs of the first BPNN based on leaf morphometric and fractal parameters. Each frame shows the BPNN outputs for a given *Banksia* acces- ▶



► sion. Reported symbols and lines show the output data. x-axis reports the name of the accessions, whereas y-axis reports the output signal strength.

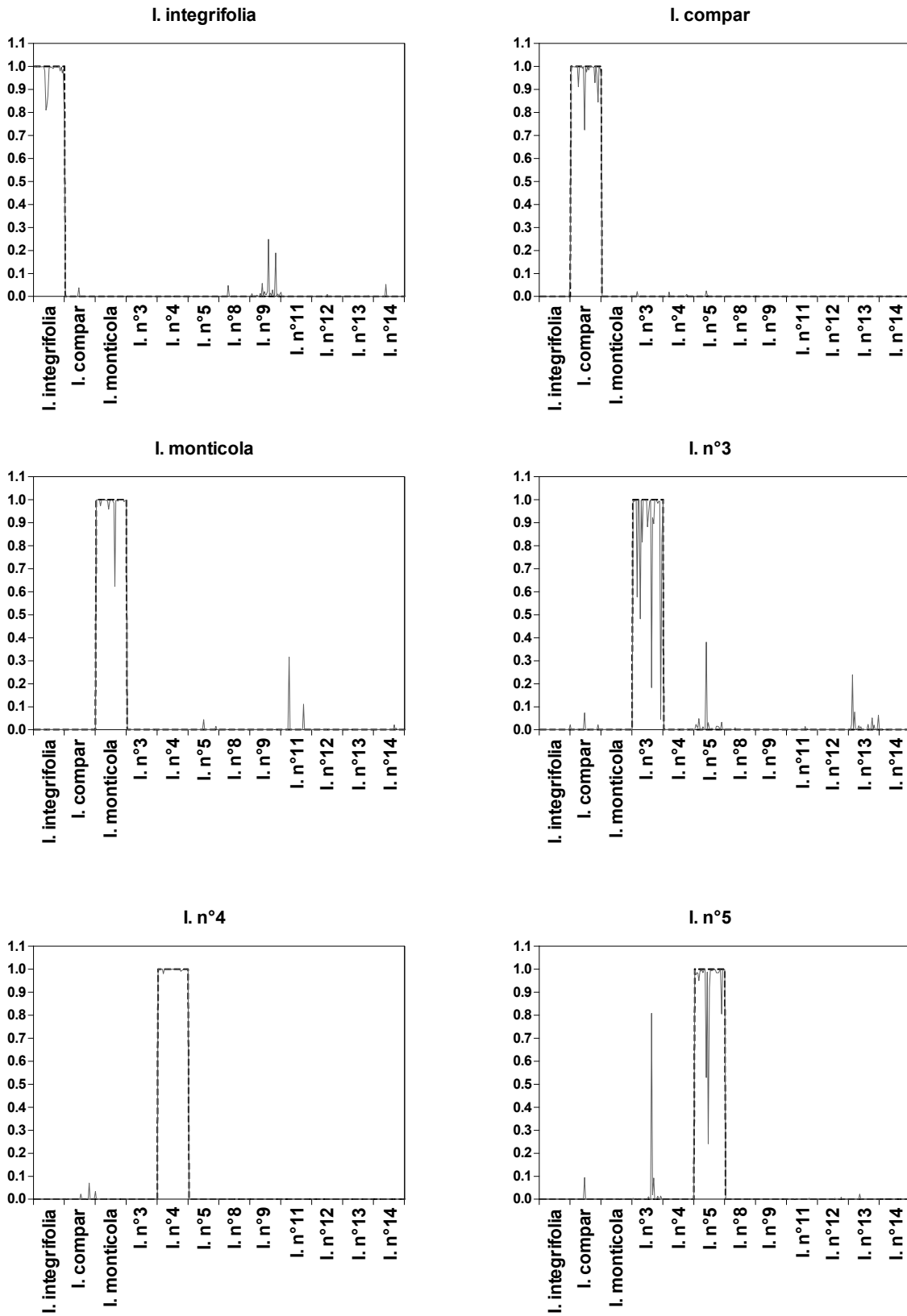
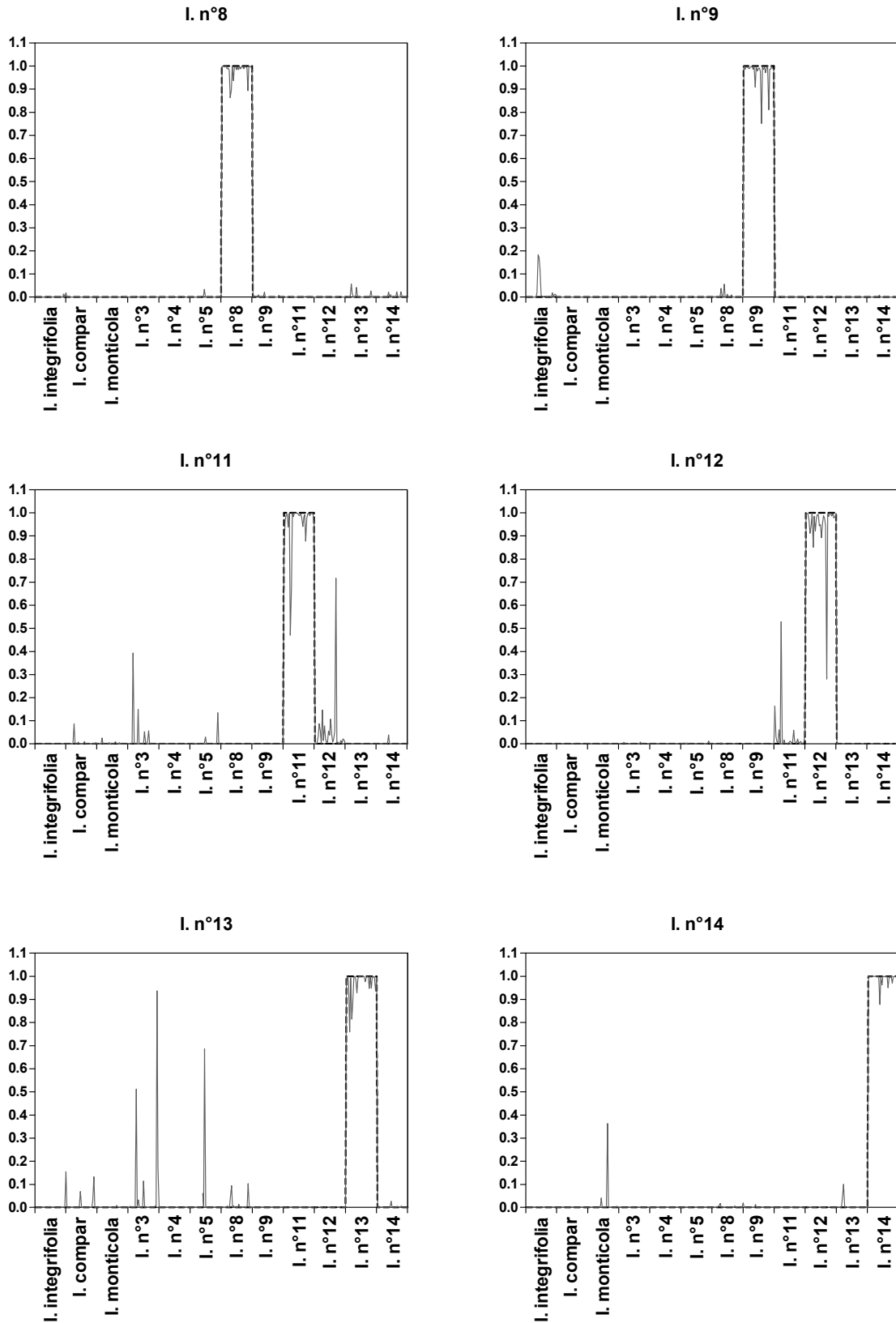


Fig. 7. Outputs of the second BPNN based on leaf and flower morphometric and fractal parameters. Each frame shows the BPNN outputs for a given ►



► *Banksia* accession. Reported lines show the output data. x-axis reports the name of the accessions, whereas y-axis reports the output signal strength.