Chemometrics and Intelligent Laboratory Systems, 9 (1990) 107–114
Elsevier Science Publishers B.V., Amsterdam

Pattern Recognition Studies in Chemical Communication: Nestmate Recognition in *Camponotus floridanus*

BARRY K. LAVINE [1,*], LAURENCE MOREL [2], ROBERT K. VANDER MEER [2], ROBERT W. GUNDERSON [3], JIAN H. HAN [1], ANTHONY BONANNO [1] and ARTHUR STINE [1]

[1] *Department of Chemistry, Clarkson University, Potsdam, NY 13676 (U.S.A.)*
[2] *USDA-ARS, Insects Affecting Man and Animals Laboratory, Gainesville, FL 32604 (U.S.A.)*
[3] *Department of Electrical Engineering, Utah State University, Logan, UT 84322 (U.S.A.)*

(Received 27 November 1989; accepted 27 February 1990)

ABSTRACT

Lavine, B.K., Morel, L., Vander Meer, R.K., Gunderson, R.W., Han, J.H., Bonanno, A. and Stine, A., 1990. Pattern recognition studies in chemical communication: nestmate recognition in *Camponotus floridanus*. *Chemometrics and Intelligent Laboratory Systems*, 9: 107–114.

The combination of gas chromatography and pattern recognition (GC/PR) analysis is a powerful tool for investigating complicated biological problems. Clustering, mapping, and principal component modelling are necessary to analyze large chromatographic data sets and to seek meaningful relationships between chemical constitution and biological variables.

We have applied GC/PR to the problem of deciphering the complex chemical messages of *Camponotus floridanus* (a carpenter ant) and have learned that GC traces of soaks obtained from the carpenter ants are characteristic of their colony of origin, social caste, and social experience. In this study gas chromatographic data obtained from 119 red carpenter ants was analyzed using principal component analysis and the FCV clustering algorithm.

INTRODUCTION

Nestmate recognition is the ability of a worker ant to discriminate workers belonging to the same colony from alien workers. It has been documented in many species of social insects [1]. In ants and bees chemical signals are the only nestmate recognition cues known [2]. These chemical signals are present on the (insect's) cuticle as a result of genetically controlled production, e.g. the ant *Pseudomyrmex ferruginea*, [3] and/or adsorption of chemicals from the environment, e.g. the ant *Solenopsis invicta* [4]. Indeed, social insects have demonstrated their great virtuosity by exhibiting a reliance for nestmate recognition on either genotypic or environmental factors or some combination of the two [1].

We have investigated [5] the nature of nestmate recognition cues and the effect of social experience on these cues in *Camponotus floridanus*, a

0169-7439/90/$03.50

highly evolved social insect, and now wish to report that the colony of origin, social caste, and social experience of these ants can be directly correlated to specific concentration patterns of cuticular compounds as represented by their gas chromatographic (GC) profiles. In this study, capillary column gas chromatography was used to analyze the soaks obtained from 119 red carpenter ants; principal component analysis and fuzzy pattern recognition techniques were then used to analyze the GC traces of the soaks. The focus of this report will be on the analytical methodology used to solve this rather interesting classification problem, with particular emphasis on the clustering techniques used to identify the various fingerprint patterns in the GC data.

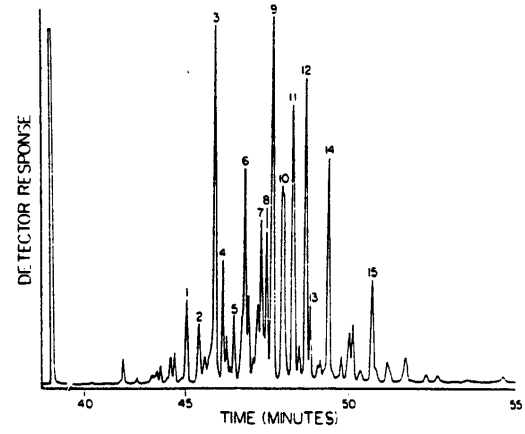

Fig. 1. A gas chromatographic trace of the hydrocarbon extract obtained from a forager showing the 15 peaks used for pattern recognition analysis.

EXPERIMENTAL

For this study 119 ants (see Table 1) were obtained from two different laboratory colonies (A and B) which were maintained in the USDA-ARS Fire Ant Project Laboratory in Gainesville, Florida, U.S.A. Ants from both colonies were fed regularly with honey–water (1 : 1), and immature insects. Five different work categories were represented in the data set: (1) foragers, (2) normal callow workers less than 12 hours old, (3) normal five-day-old callow workers, (4) naive callow workers less than 12 hours old, and (5) naive five-day-old callow workers. Foragers were collected at the honey source in the foraging area. Callow workers were removed from each colony as they emerged from their cocoon and were maintained in small Petri dishes with nurses (normal callows) or without nurses (naive callows).

TABLE 1

Camponotus floridanus data set

Worker category	Colony	Number of specimens
Forager	A	21
5-day-old callow	A	10
0-day-old callow	A	11
5-day-old naive callow	A	12
0-day-old naive callow	A	11
Forager	B	13
5-day-old callow	B	10
0-day-old callow	B	10
5-day-old naive callow	B	11
0-day-old naive callow	B	10
Total		119

The chemical composition of the cuticle was determined by first soaking individual ants in 150 μl of hexane. After three hours, the ant was removed from the hexane wash, and the solution was transferred via a Pasteur pipet to a clean vial. GC analysis was performed on the extract using a Varian 3700 gas chromatograph equipped with a flame ionization detector. A representative GC trace of a soak obtained from a forager is shown in Fig. 1. A 30-m DB-1 fused silica capillary column (J&W Scientific) was used in the analysis, and the column was temperature programmed from 50 °C to 285 °C at 5 °C/minute. Further details regarding the collection of the GC data can be found elsewhere [5].

No qualitative differences were apparent in the chemical composition of the soaks obtained from ants that were from different colonies or from different worker categories. If the compounds comprising the cuticle do in fact play an important role in nestmate recognition, it would seem likely that quantitative differences in the soaks would result in concentration profiles characteristic of the colony of origin and the worker's

category. In order to investigate this hypothesis, data base of information concerning the relative concentration of the compounds comprising the cuticle (i.e. the GC traces) which we compiled was analysed using pattern recognition techniques.

PATTERN RECOGNITION METHODOLOGY

For pattern recognition (PR) analysis, each gas chromatogram was represented by a data vector $X = (x_1, x_2, x_3, x_4, \ldots, x_j, \ldots x_n)$ where component x_j is the area of the jth peak. In this study the chromatographic data were normalized to constant sum using the total area of the forty GC peaks. Each peak was then expressed as percent of total area to indicate the relative concentration. Of the 40 peaks comprising each chromatogram, only 15 were considered for pattern recognition analysis (see Fig. 1). Each of the 15 had an area representing more than 1% of the total in every chromatogram. Computer integration of these peaks always yielded reliable results. Furthermore, each of these peaks was well resolved and readily identifiable in all of the chromatograms. Because this feature selection process was carried out on the basis of a priori considerations, the probability of exploiting random variation in the data was minimized.

Of the 119 ant samples, 103 comprised the training set (see Table 2). There were 53 ants from colony A and 50 ants from colony B in the training set. Of the 53 ants from colony A, fifteen were foragers, ten were 5-day-old callows, eight were 0-day-old callows, ten were 5 day-old-naive callows, and ten were 0-day-old naive callows. Of the 50 ants from colony B, each of the five different worker categories were represented by 10 ant samples. The prediction set (see Table 3) consisted of 12 samples from colony A (six foragers, three 0-day-old callows, one 0-day-old naive callow and two 5-day-old naive callows), and 4 samples from colony B (3 foragers and one 5-day-old naive callow). Members of the prediction set were chosen by random lot. The raw data vectors for the prediction set ants were sent to our laboratory under separate cover and served as blind samples in the study. The class assignment of each prediction set sample was made known after the colony of origin and worker category of the ants were postulated.

TABLE 2

Training set

Worker category	Colony	Number of specimens
Forager	A	15
5-day-old callow	A	10
0-day-old callow	A	8
5-day-old naive callow	A	10
0-day-old naive callow	A	10
Forager	B	10
5-day-old callow	B	10
0-day-old callow	B	10
5-day-old naive callow	B	10
0-day-old naive calow	B	10
Total		103

TABLE 3

Prediction set

Worker category	Colony	Number of specimens
Forager	A	6
0-day-old callow	A	3
5-day-old naive callow	A	2
0-day-old naive callow	A	1
Forager	B	3
5-day-old naive callow	B	1
Total		16

In this study we used the FCV clustering algorithm [6–8] to seek relationships between the GC profiles of the ants and the biological variables; age, social experience, and colony of origin. The FCV clustering algorithm attempts to fit each of the c classes in the data set to a linear model (i.e. principal component model) of the form

$$x = v + \sum t_j d_j \qquad (1)$$

where x denotes a prototypical class membership vector, v is the center of the class in the n-dimensional space R_n, the vectors $\{d_j\}$ are an orthonormal set spanning a subspace of R_n, and t_j are the coordinates of the prototypical vector in the subspace. An interesting feature of the FCV cluster-

ing algorithm is that each data vector in the training set is assumed to contribute to the modelling of each of the classes within the data. The actual algorithm consists fo solving simultaneously the following set of equations.

$$v_i = \sum_{k=1}^{n} (u_{ik})^m x_k / \sum_{k=1}^{n} (u_{ik})^m \quad (2)$$

$$u_{ik} = 1 / \sum_{j=1}^{c} (D_{ik}/D_{jk})^{1/m-1} \quad (3)$$

$$S_i = \sum_{k=1}^{n} (u_{ij})^m (x_k - v_i)(x_k - v_i)^T \quad (4)$$

$$D_{ik} = \left(|x_i - v_i|^2 - \sum_{j=1}^{r} \langle x_k - v_{ij}, d_{ij} \rangle^2 \right)^{1/2} \quad (5)$$

The membership value of sample k with respect to class $i (i = 1, 2, 3, \ldots)$ is u_{ik}, and these values are subjected to the condition $0 < u_{ik} < 1$ and $\Sigma u_{ik} = 1$. D_{ik} is the distance of sample k from cluster center i, v_i is the center of cluster (i.e. class) i, d_{ij} is a unit eigenvector corresponding to the jth largest eigenvalue of the fuzzy within-cluster scatter matrix, $\mathbf{S}_i$, m is a fixed weighting exponent which usually is assigned a value of 2, and r defines the shape of the cluster ($r = 0$ for round clusters, $r = 1$ for linear varieties, and so forth).

To obtain an approximate solution to this set of four equations, the user must supply the starting cluster centers. The class membership values, the within-cluster scatter matrix for each cluster, and the distance of each sample from each cluster are computed in rapid succession. New cluster centers are then computed for the samples in the final step of the first iteration. The algorithm continues by using these new cluster centers as the starting point for a second iteration through the same set of four equations. This process continues until convergence is achieved. The number of iterations required to achieve convergence depends upon the minimum pre-specified change criterion for the class membership values (which has been set at 0.005 in our studies).

Usually one chooses $m = 2$, but by increasing m, less weight is attached to the importance of samples with small membership values. This means that the higher the value of m, the fuzzier the algorithm becomes, in the sense that points whose membership values are uniformly low through the iterative procedure tend to become increasingly ignored in determining the membership functions (i.e. clusters) and the defining linear varieties. It is this feature of the FCV clustering algorithm that is particularly appealing when one suspects the data may not exist in compact well separated clusters. The ability to 'tune out' noise in the data by adjusting m can be of great value in obtaining favorable and meaningful clustering results.

One of the advantages of the FCV clustering algorithm is the possibility of using the fuzziness of a given cluster configuration as an indicator of its quality. This can be achieved by computing the cluster validity coefficient (CVC). It is a measure of the separation between clusters and is determined by computing the ratio of the distance between the two cluster centers to the weighted scatter of the two clusters [9]. The larger the value of the coefficient, the better the separation between clusters. By successively increasing the value of m, the effect of samples with poor class membership values can be filtered out. An indication of the cluster quality can therefore be obtained by comparing the values of the cluster discriminant from computations where m is increased stepwise. If there is little change in the value of the CVC, the conclusion is that most of the data have shared membership values close to either zero unity, which suggests the presence of two distinct clusters (i.e. classes) in the data. On the other hand, a marked increase in the value of the CVC as m increases would be taken as an indication of substantial overlap between the two data clusters.

When investigating the data with the FCV clustering algorithm, one may choose to search for round clusters in the data by specifying $r = 0$, or find the best fit of the data to linear clusters by specifying $r = 1$, or in general attempt to fit the data to other geometric shapes by setting $r \geq 2$. This feature of the FCV algorithm allows the investigator to compare the fit of the data to a number of geometrically distinct cluster shapes and to select the fit which appears to best represent the actual structure of the data. Clearly, this feature results in a very definite advantage.

RESULTS AND DISCUSSION

In the first step of our study principal component analysis [10] was employed to examine the structure of the data. In Fig. 2 the results of a principal component mapping experiment are shown for the 103 ants samples in the training set. The 53 ant samples from colony A are well separated from the 50 ant samples from colony B in the two-dimensional map. It is important to note this projection is made without the use of information about the class assignments of the samples, i.e. colony A or colony B. The resulting separation is, therefore, a strong indication of real differences in the cuticular patterns of these ants as reflected in their GC profiles.

The FCV clustering algorithm was then used to partition the data into two classes and to assess the degree of separation between the two classes (i.e. colony A and colony B). The starting centers for this clustering experiment were a forager from colony A and a 0-day-old naive callow from colony B. The values of m and c were set at 2, and the value of r was set at 0. (For GC profile data, we have learned from previous experience that r should usually be set to zero.) The FCV clustering experiment was performed on the 15 raw variables. The algorithm converged after 25 iterations. The 53 samples from colony A had a larger class membership value for cluster one than cluster two, and the 50 samples from colony B had a larger class membership value for cluster two than cluster one. The CVC value was computed to be 6.20. CVC values were also computed for other values of m. Only minor changes in the CVC were observed for increasing values of m. The FCV clustering experiment was repeated using different starting centers (a 5-day-old callow from colony A and a forager from colony B), and the same results were achieved. Evidently, it is reasonable to divide the training set into two distinct classes: colony A and colony B.

The principal component models developed in the FCV clustering experiment were validated using the 16 samples from the prediction set. The chromatograms in the prediction set were fitted to the class models and were assigned a membership value for each class. An ant sample would be assigned to colony A if its class membership value was greater than 0.50 for cluster 1. If the ant sample had a class membership value greater than 0.50 for cluster 2, it would be assigned to colony B. The results of this experiment are summarized in Table 4. Every chromatogram in the prediction set was correctly classified. This result demonstrates that information derived solely from the hydrocarbons can categorize ant specimens as to colony of origin. Because differences in odor for the laboratory colonies are only due to genotypic factors, we would expect differences between field colonies which would also include environmental factors to be even more pronounced.

The GC profiles were also found to be characteristic of the worker's category. The training set of 103 chromatograms was further subdivided into two smaller sets — one of 53 chromatograms (colony A) and the other of 50 chromatograms (colony B). Each set was analyzed separately using the FCV cluster algorithm. For colony A, r was set equal to 0, m was set equal to 2, and c was varied from 2 to 8. As c increased in value, clustering of samples on the basis of social caste was observed. When $c = 6$, an interesting result was obtained. Five of the fifteen foragers had a large membership value (greater than 0.60) for cluster six. None of the other samples had a large membership value for that cluster. The other ten foragers had a large membership value for cluster one. Only two other samples (0-day-old naive callows) had a large membership value for cluster one. All of the 5-day-old callows had a large membership value for cluster 2. Only two other samples (5-day-old naive callows) had a large class membership value for cluster two. Six 0-day-old callows had a large membership value for cluster three. None of the other samples had a large class membership value for this cluster. The remaining eight 5-day-old naive callows had a large membership value for cluster four, and eight 0-day-old naive callows has a large membership value for cluster five. When the ant samples were assigned to the cluster where they had the highest class membership value, it was evident the data could be partitioned into different classes on the basis of the worker's category. The results of this experiment are summarized in Table 5.

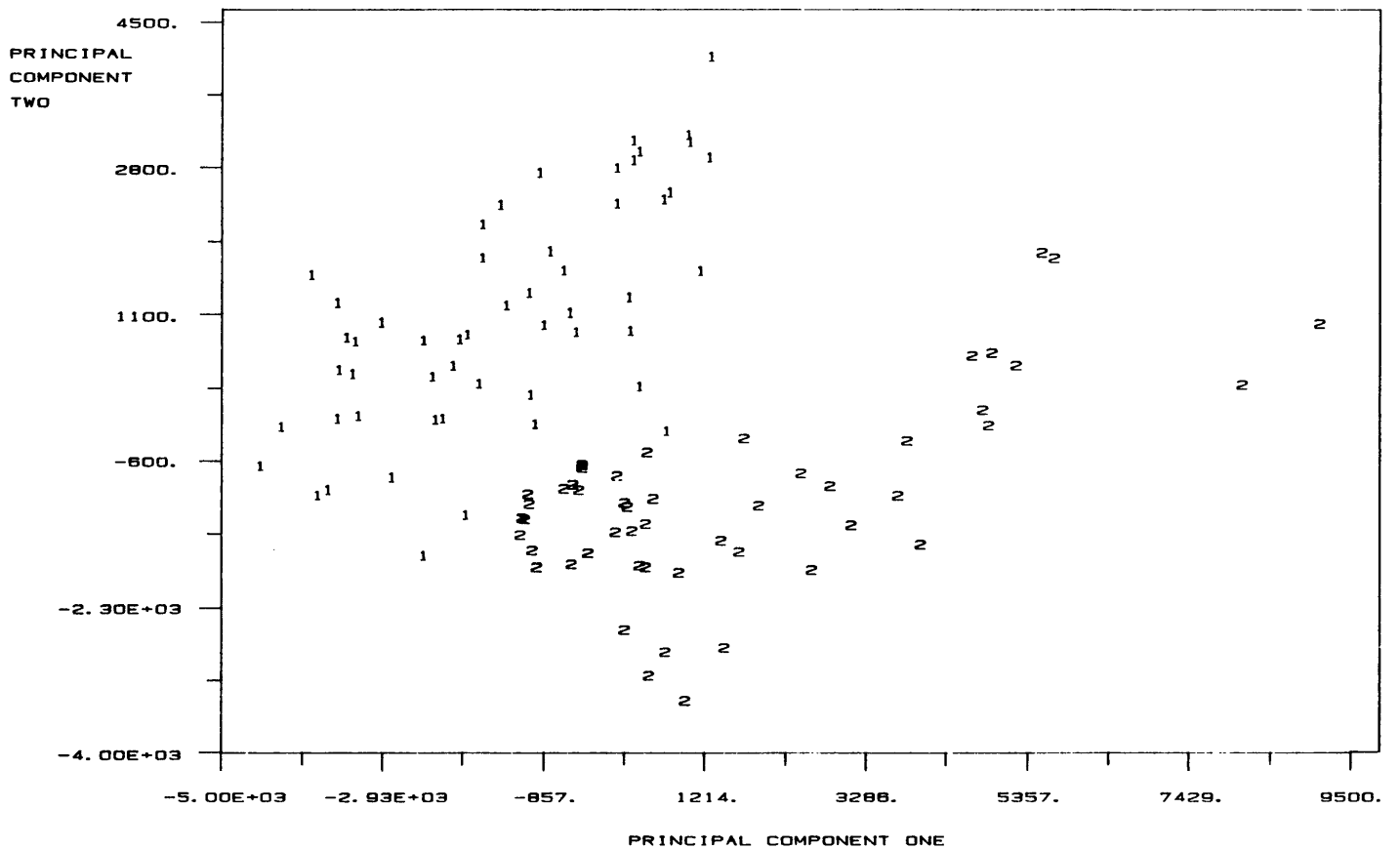

Fig. 2. Plot of the first two principal component of the 15 GC peaks obtained from individual workers. The 1s represent colony A and the 2s represent colony B. The first two principal components account for 60.8% of the total cumulative variance.

TABLE 4

Prediction set results for colony of origin

Worker category	Colony	Membership value	
		Cluster 1	Cluster 2
Forager (#1)	A	0.75	0.25
Forager (#2)	A	0.80	0.20
Forager (#3)	A	0.70	0.30
Forager (#4)	A	0.65	0.35
Forager (#5)	A	0.85	0.15
Forager (#6)	A	0.74	0.26
5-day-old naive callow (# 1)	A	0.88	0.12
5-day-old naive callow (#2)	A	0.90	0.10
0-day-old callow (#1)	A	0.87	0.13
0-day-old callow (#2)	A	0.82	0.18
0-day-old callow (#3)	A	0.82	0.18
0-day-old naive callow	A	0.79	0.21
Forager (#1)	B	0.28	0.72
Forager (#2)	B	0.20	0.80
Forager (#3)	B	0.30	0.61
5-day-old naive callow	B	0.22	0.78

The six principal component models developed in this experiment were validated using 12 of the 16 samples from the prediction set. These 12 samples were from colony A. The samples were fitted to each principal component model, and the class membership values for each sample were computed. Each sample was assigned to the cluster where it had the highest membership value. Ten of the twelve samples were correctly classified. A 0-day-old callow was classified as a 5-day-old callow (i.e it was assigned to cluster 2 instead of cluster 3), and a forager was classified as a 5-day-old callow (i.e. it was also assigned to cluster 2 instead of cluster 1 or 6). Clustering experiments were performed for the 50 ant samples from colony B, and favorable classification results were also obtained (see Table 6). The prediction set results were equally encouraging. All four samples from colony B in the prediction set were correctly classified. (The three foragers had the largest class membership value for cluster 1, and the 5-day-old naive callow had the largest class membership value for cluster 2.)

TABLE 5

Training set results for colony A

Number of samples	Caste	Maximum membership cluster
10	Foragers	1
2	0-day-old naive callows	
10	5-day-old callows	2
2	5-day-old naive callows	
6	0-day-old callows	3
8	5-day-old naive callows	4
8	0-day-old naive callows	5
2	0-day-old callows	
5	Foragers	6

TABLE 6

Training set results for colony B

Number of samples	Caste	Maximum membership cluster
7	Forager	1
2	0-day-old naive callows	
9	5-day-old naive callows	2
2	0-day-old callows	
5	0-day-old naive callows	3
3	Foragers	4
10	5-day-old callows	5
8	0-day-old callows	6
1	5-day-old naive callow	
3	0-day-old naive callows	7

The results of this study demonstrate that foragers and callows (normals or naives) can be differentiated from one another on the basis of their GC profile. This implies a direct relationship between the concentration pattern of the cuticular components and the social caste of these ants. The results also demonstrate that the GC traces of the ants convey information about their social experience. For example, we were able to readily differentiate 5-day-old callows from 5-day-old naive callows by examining the GC traces of their hexane soaks. The only difference between 5-day-old callows and 5-day-old naive callows is that 5-day-old callows have interacted with the nurses whereas

the 5-day-old naive callows do not participate in such an interaction. It seems probable this interaction leads to an exchange of cuticular chemicals which contain the nestmate recognition label. Behavioral studies carried out in conjunction with the GC/PR analysis [5] demonstrate the importance of this interaction. If the callows do not interact with the nurses, they will not acquire the full colony label.

The pattern recognition methods used in the study also deserve comment. Massart and Kaufman in their classic text, *The Interpretation of Analytical Data by the use of Cluster Analysis* ([11] p. 39), note that a clustering method combined with a mapping and display technique, preferably principal component analysis, is the best approach to take for tackling a classification problem. These techniques do not utilize information about the class assignment of the sample. Therefore, if the results from the principal component mapping and clustering experiments confirm our a priori assumption that we have about the structure of the data, we therefore have a very strong indication that real differences exist between chromatograms belonging to samples from different colonies or worker categories. Although supervised pattern recognition techniques could be used for classification problems, we believe that we have used the appropriate pattern recognition methodology given the scope and nature of the study.

CONCLUSION

GC traces representing hydrocarbon extracts could be related to the social experience, colony or origin and social caste of the ants using pattern recognition techniques. These results support a potential role for cuticular hydrocarbons in nestmate recognition. They also demonstrate the GC/PR is an important technique for evaluating the informational content of highly complex chemical communicatory systems.

ACKNOWLEDGEMENTS

This study was partially supported by Contract FO8635-90-C-0105 between Clarkson University and the United States Air Force. The authors wish to thank Janet Lavine for many valuable discussions.

REFERENCES

1 M.D. Breed and B. Bennet, Kin recognition in highly eusocial insects, in D.J.C. Fletcher and C.D. Michener (Editors), *Kin Recognition in Animals*, Wiley, New York, 1986, pp. 243–286.
2 B. Holldobler and C.D. Michener, Mechanisms of identification and discrimination in social hymenoptera, in H. Markl (Editor), *Evolution of Social Behavior: Hypothesis and Empirical Tests*, Verlag Chemie, Weinheim, 1980, pp. 433–439.
3 A. Mintzer, Nestmate recognition and incompatability between colonies of the acacia ant *Pseudomyrmex ferruginea*, *Behavioral Ecology and Sociobiology*, 10 (1982) 165–168.
4 M.S. Oblin, Nestmate recognition cues in laboratory and field colonies of *Solenopsis invicta Buren* (Hymenoptera: Formicidae). Effect of the environment and role of the hydrocarbons, *Journal of Chemical Ecology*, 12 (1986) 1965–1975.
5 L. Morel, R.K. Vander Meer and B.K. Lavine, Ontogeny of nestmate recognition cues in the red carpenter ant (*Camponotus floridanus*), *Behavioral Ecology and Sociobiology*, 22 (1988) 175–183.
6 J.C. Bezdek, C. Coray, R. Gunderson and J. Watson, Detection and characterization of cluster substructure II. Fuzzy *c*-varieties and convex combinations thereof, *SIAM Journal of Applied Mathematics*, 40 (1981) 358–372.
7 R.W. Gunderson and T. Jacobson, Cluster analysis of beer flavor components. I. Some new methods in cluster analysis, *ASBC Journal*, 41 (1983) 73–77.
8 T. Jacobsen and R.W. Gunderson, Trace element distribution in yeast and wort samples: An application of the FCV clustering algorithms, *International Journal of Man-Machine Studies*, 19 (1983) 105–116.
9 R.W. Gunderson and K.E. Thrane, Monitoring polycyclic aromatic hydrocarbons: an environmental application of fuzzy *c*-varieties pattern recognition, in J.J. Breen and P.E. Robinson (Editors), *Environmental Applications of Chemometrics*, ACS Symposium Series 292, American Chemical Society, Washington, DC, 1985, pp. 130–148.
10 I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
11 D.L. Massart and L. Kaufman, *The Interpretation of Analytical Data by the Use of Cluster Analysis*, Wiley, New York, 1983, p. 39.