Mining RFID Behavior Data using Unsupervised Learning

Guénaël Cabanes, Younès Bennani, Dominique Fresneau

Mining RFID Behavior Data using Unsupervised Learning

Guénaël Cabanes*[1], Younès Bennani[1], and Dominique Fresneau[2]
[1]LIPN-CNRS UMR 7030, [2]LEEC, FRANCE

ABSTRACT

Radio Frequency IDentification (RFID) is an advanced tracking technology that can be used to study the spatial organization of individual's spatio-temporal activity. The aim of this work is firstly to build a new RFID-based autonomous system which can follow individuals' spatio-temporal activity, a tool not currently available. Secondly, we aim to develop new tools for automatic data mining. In this paper, we study how to transform these data to investigate the division of labor, the intra-colonial cooperation and conflict in an ant colony. We also develop a new unsupervised learning data mining method (DS2L-SOM: Density-based Simultaneous Two-Level - Self Organizing Map) to find homogeneous clusters (i.e., sets of individual which share a similar behavior). According to the experimental results, this method is very fast and efficient. It also allows a very useful visualization of the results.

Keywords: RFID; spatio-temporal data; automatic data mining; unsupervised learning; ants behavior.

INTRODUCTION

Radio Frequency IDentification (RFID) is an advanced tracking technology. The RFID tags, which consist of a microchip and an antenna, must be used with a reader that can detect simultaneously a lot of tags in a single scan. A computer has to be used to store the data about the position of each tag for each scan in a database. This allows different analyses.

RFID systems can be used to study animal societies. Animal societies are dynamic complex systems characterized by numerous interactions between individual members. Such dynamic structures stem from the synergy of these interactions, the individual capacities in information processing and the diversity of individual responses (Fresneau *et al.*, 1989). RFID, thanks to miniaturization, offers the advantage of automation and overcomes the constraints imposed by video analyzes. Indeed, video recording allows long-duration tracking, however the time for analyzes highly increases with the number of individuals monitored. It also imposes strong constraints (as the need of a minimum illumination and high contrast between the animals and the environment) and it does not work when the ant is moving in a reverse position which doesn't allow individual identification. The aim of this work is to develop a new RFID-based autonomous system to follow the spatio-temporal activity of groups, which is currently very difficult to study in its entirety and to develop new tools for automatic data processing. These objectives have necessarily led to an interdisciplinary project combining behavioral and complex systems sciences with computer and engineering sciences.

Since dynamic experimental data are extremely difficult to collect, behavioral sciences are dominated by static approaches (optimality, game theory). In order to understand the functioning of insect societies, the integration of both individual and collective types of analyzes is necessary. However, there are two essential challenges: the presence of autonomous units and the great influence of fluctuations. These challenges imply that adequate observation tools are available. Therefore, traceability technologies are greatly needed, e.g., RFID, which allows the automated monitoring of the localizations and of the movements of many individuals simultaneously. Those systems are now well developed and miniaturized enough to be used in insect societies (Streit *et al.*, 2003) and they are now almost operational.

However, experiments using RFID generate large datasets which need suitable analysis methods to allow a comprehensive understanding of the link between events and reveal behavioral patterns. In this paper, we investigate how to transform these data to study the division of labor, the intra-colonial cooperation and conflict in an ant colony. Ants, often caricatured and little known, have nevertheless a huge ecological impact and are considered as major energy catalysts. Their complex underground nests contribute to soil ventilation and ecosystem equilibrium because of their predatory and detritivore diets. Ants are very diverse and the *Formicidae* (11000 described species) exhibit a great variety of social structures (Passera & Aron, 2005). Division of labor is one of their fundamental characteristics (Hölldobler & Wilson, 1990). According to the needs of the colony, each individual in the colony can assume a basic behaviors, such as: nursing the queen and the brood (i.e: eggs, larva and cocoon), transporting food or building material, hunting, and so on. Ant colonies face rapid changes of environmental conditions and constraints through an important individual flexibility. The dynamic component of this phenomenon is the hallmark of our research on social organization and on colony performance during migrations (change of nest). A RFID device has been developed for these study organisms. Based on marketed products, it requires little development. It consists of a network of RFID readers in a constrained space with compulsory passageways in an artificial nest. These readers are connected to a detector which sends the information to a computer.

However, in this study, we don't have any prior knowledge about the structure of the social organisation of the colony, for example, the usual behavior of each ant and how many different behaviors can be expressed by the colony. In that case, most usual statistical approaches fail, as they need at least two different samples to compare. To analyze the internal structure of a unique data set, unsupervised classification methods (clustering methods), are very powerful. They allow an automatic detection of relevant sub-groups (or clusters) in a data set and are particularly suitable for data mining from experimental studies, for which we have generally little a priori information. Thus, in this paper, we analyze the collected data by using a new method of data mining, based on a clustering algorithm (DS2L-SOM: Density-based Simultaneous Two-Level - Self Organizing Map, Cabanes & Bennani (2007, 2008)). DS2L-SOM is an effective connectionist (i.e. neuro-inspired) clustering tool to find and simply represent a significant amount of information about the structure of data. We apply this method to data from experimental research because of its efficiency in the extraction of information in this field and the discovery of knowledges. This method allows the discovery of a topological space from a set of behavioral observations.

The remainder of this paper is organized as follows. We first presents the DS2L-SOM algorithm. Then we describes the experimental protocol of the behavioral study and we show some results and their interpretations. Conclusion and future works are finally given.

A TOPOGRAPHIC CONNECTIONIST UNSUPERVISED LEARNING

High dimension data may be sparse (the curse of dimensionality), making it difficult for a clustering algorithm to find any structure in the data. Indeed, when dimensionality increases, data become increasingly sparse in the space that it occupies. Definitions of density and distance between objects, which is critical for clustering and outliers detection, become less meaningful. To solve this problem, a large number of dimension reduction approaches have been developed and tested in different application domains and research communities. The main idea behind these techniques is to map each pattern into a lower dimensional space that preserves the topology of the data. The reduced data present in the lower dimensional representation can be used to perform clustering more efficiently. Various approaches have been proposed as a two-level clustering algorithms (Bohez, 1998; Hussin et al., 2004; Aupetit, 2005; Ultsch, 2005; Guérif & Bennani, 2006; Korkmaz, 2006).

The key idea of the two-level clustering approach based on a Self Organizing Map (SOM) (Kohonen, 1984, 2001) is first to combine the dimensionality reduction and the fast learning capabilities of SOM to construct a new reduced vector space, then to apply another clustering method in this new space to produce a final set of clusters in the second level (Hussin et al., 2004; Ultsch, 2005; Guérif & Bennani, 2006). Although the two-level methods are more interesting than the traditional approaches, the data segmentation obtained from the SOM is not optimal, since a part of the information is lost during the first stage (dimensionality reduction). Moreover, this separation in two stages is not suitable for a dynamic, i.e. incremental, segmentation of data despite the important needs for such analysis.

We propose a new unsupervised learning algorithm, DS2L-SOM (Cabanes & Bennani, 2008), which learns simultaneously the structure of the data and its segmentation using both distance and density information.

Principle of the SOM

The Kohonen SOM can be defined as a competitive unsupervised learning neural network. When an observation is recognized, activation of an output cell – competition layer – leads to inhibit the activation of other neurons and to reinforce itself. It is said that it follows the so called "Winner Takes All" rule. Actually, neurons are specialized in the recognition of one kind of observations.

SOM consists of a two dimensional map of neurons which are connected to n inputs according to n weights connections $w^{(i)} = (w_1^{(i)},..., w_n^{(i)})$ (prototype vectors) and to their neighbors with topological links. The training set is used to organize this map under topological constraints of the input space. Thus, a mapping between the input space and the network space is constructed; two nearby observations in the input space would activate two close units of the SOM. An optimal spatial organization is determined by the SOM from the input data. When the dimension of the input space is lower than three, both position of weights vectors and direct neighborhood relations between cells can be represented visually. Thus, a visual inspection of the map provides qualitative information about the map and the choice of its architecture. The winner neuron updates its prototype vector, making it more sensitive for latter presentation of that type of input. This allows different cells to be trained for different types of data. To achieve a topological mapping, the neighbors of the winner neuron can adjust their prototype vector towards the input vector as well, but to a less degree, depending on how far away they are from

the winner. Usually a radial symmetric Gaussian neighborhood function $K_{i,j}$ is used for this purpose.

The DS2L-SOM algorithm

Connectionist learning algorithms are often presented as a minimization of a cost function. In our case, it will correspond to the minimization of the distance between the input samples and the map prototypes, weighted by a neighborhood function $K_{i,j}$. To do that, we use a gradient algorithm. The cost function to be minimized is defined by:

$$R(w) = \frac{1}{N} \sum_{k=1}^{N} \sum_{j=1}^{M} K_{j,u^*(x^{(k)})} \parallel w^{(j)} - x^{(k)} \parallel^2$$

N represents the number of learning samples, M is the number of neurons (units) in the map, $u^*(x^{(k)})$ is the neuron having the weight vector closest to the input pattern $x^{(k)}$, known as the best match unit (BMU). $K_{i,j}$ is a positive symmetric kernel function: the neighborhood function. The relative importance of a neuron i compared to a neuron j is weighted by the value of the kernel function $K_{i,j}$ which can be defined as:

$$K_{i,j} = \frac{1}{\lambda(t)} \times e^{-\frac{d_1^2(i,j)}{\lambda^2(t)}}$$

$\lambda(t)$ is the temperature function modeling the topological neighborhood extent, defined as:

$$\lambda(t) = \lambda_i \left(\frac{\lambda_f}{\lambda_i}\right)^{\frac{t}{t_{max}}}$$

λ_i and λ_f are respectively initial and the final temperature (for example $\lambda_i = 2$, $\lambda_f = 0.5$). t_{max} is the maximum number allotted to the time (number of iterations for the x learning sample). $d_l(i,j)=r-k+s-m$ is the Manhattan distance defined between two neurons i and j on the map grid, with the coordinates (k, m) and (r, s) respectively.

In DS2L-SOM, each neighborhood connection is associated with a real value (v) which indicates the relevance of the connected neurons. The value of this connection is adapted during the learning process. Given the organization constraint of the SOM, both best closest prototypes of each data point must be connected by a topological connection. The value v of this connection will be adapted by the algorithm. It was proved by Martinetz (1993) that the so generated graph is topology-preserving optimally in a very general sense. In particular, each edge of this graph belongs to the Delaunay triangulation corresponding to the given set of reference vectors. Thus, at the end of the training, a set of inter-connected prototypes will be an artificial image of well separated sub-group of the whole data set. We propose also to associate each unit i to an estimation of the local data density $D^{(i)}$, so as to detect local fluctuation of density which defines the borders of touching clusters. For each data point, this density value will be increased for all units in function of the Euclidean distance between the related prototype $w^{(i)}$ and the data. This method of evaluation is similar to the one proposed by Pamudurthy et al. (2007).

At the end of the learning process, prototypes which are linked together by neighborhood connections such as $v > 0$ define the well separated clusters. Thus, we use a “Watersheds”

method (see Vincent & Soille (1991)) on the density map of each of these clusters to find locally low density area inside well separated clusters so as to characterize density defined sub-clusters. For each pair of adjacent subgroups, a density-dependent index is used (Yue et al., 2004) to determine if an area of low density is a reliable indicator of the data structure, or whether it should be regarded as a random fluctuation in the density. This process is very fast because of the reduced number of prototypes. The combined use of these two types of group definition can achieve good results despite the low number of prototypes of the map.

The DS2L-SOM algorithm can be used either during the learning of the prototypes of the SOM (stochastic version), in particular for not having to keep the data matrix in memory (Cabanes & Bennani, 2008), or after the learning of the prototypes (batch version) for greater flexibility and reduction of computing time. We present the batch version as follows.

The DS2L-SOM learning algorithm proceeds essentially in three phases:

Input : Distances matrix Dist(w, x) between the prototypes w and the N data x.
Output : Groups of similar units (the *clusters*).

1. Calculation of D and v :
   - Density estimate :

$$D^{(i)} = \sum_{k=1}^{N} \frac{e^{-\frac{Dist(w^{(i)},x^{(k)})^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

     with σ, a parameter to be chosen by user (σ = 0.1 in this study).

   - Calculation of the neighborhood connections values :
     - For each data x, find the two closest BMUs, $u^*(x)$ and $u^{**}(x)$:

$$u^*(x) = \text{argmin}_i\ (\text{Dist}(w^{(i)},x))$$
$$u^{**}(x) = \text{argmin}_{i \neq u^*(x)}\ (\text{Dist}(w^{(i)},x))$$

     - Let $v_{i,j}$ = be the number of data having i and j as BMUs.

2. Extract all groups of connected units : Let $P = \{C_i\}_{i=1..L}$ the set of the L groups of linked unit such as v > threshold (see fig.1(b)). Here threshold = 0.

3. For each $C_k \in P$ do :
   - Find the set M (C_k) of density maxima (i.e. density mode, see fig.1(c)).

$$M(C_k) = \{\text{unit } i \in C_k \mid D^{(i)} \geq D^{(j)}, \forall j \in \aleph(i)\}$$

     □(i) is the neighborhood of unit i.

   - Determine the threshold matrix :

$$S = [S(i,j)]_{i,j:1...|M(Ck)|}$$

with

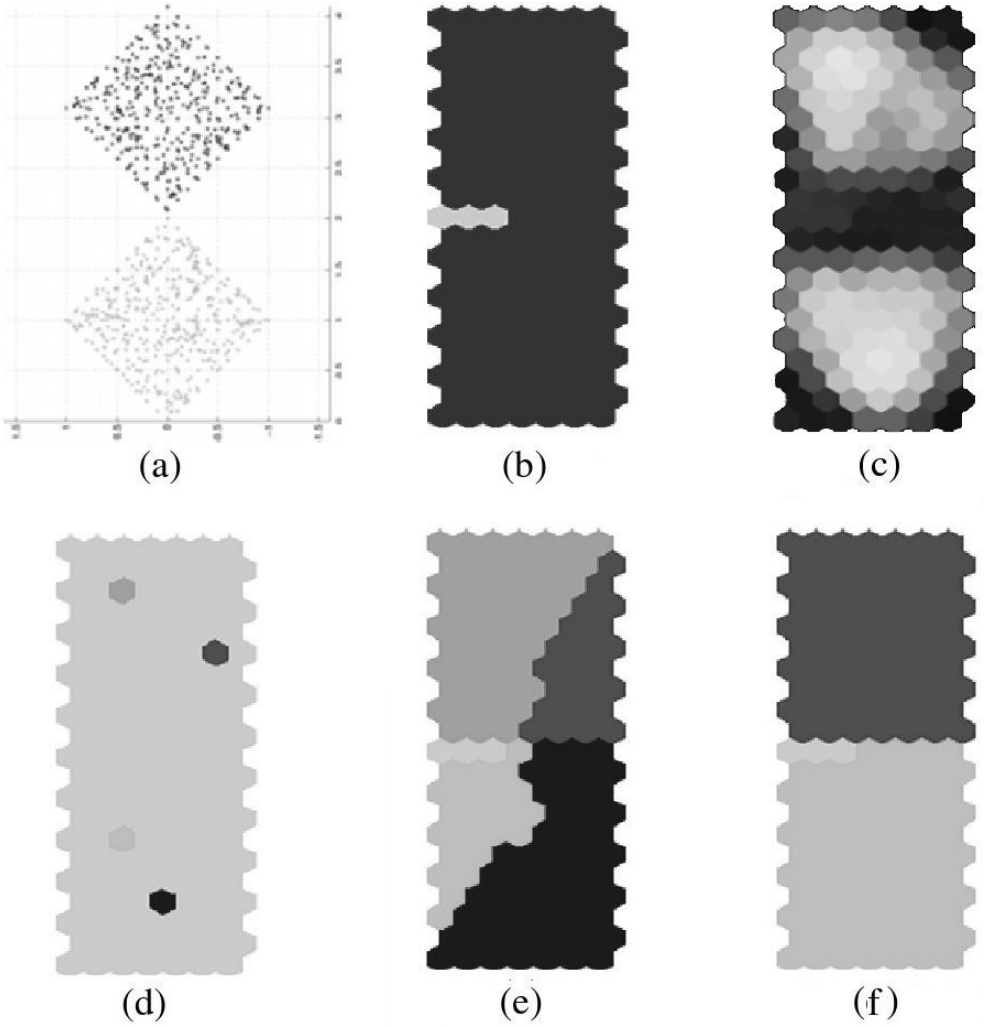
$$S(i, j) = (1/D^{(i)} + 1/D^{(j)})^{-1}$$

- For all unit $i \in C_k$, label the unit i with one element *label*(i) of $M(C_k)$, according to an ascending density gradient along the topological connections. Each label represents a micro-cluster.

- For each pair of neighbors unit (i, j) in C_k, if *label*(i) ≠ *label*(j) and if both $D^{(i)} > S(label(i), label(j))$ and $D(j) > S(label(i), label(j))$ then merge the two micro-clusters.

4. Return refined clusters.

Figure 1. Example of a sequence of the different stages of the refinement algorithm : (a) data base, (b) one set of connected units and some unconnected units (in gray), (c) density values for each unit, (d) density modes detection, (e) detection of subgroups associated to each mode, (f) Merging of irrelevant subgroups (final clustering)



Validity of the clustering algorithm

The quality of this clustering method is very good. Experimental results (see Cabanes & Bennani (2008)) demonstrated that the proposed clustering method achieves a better clustering quality than classical approaches. DS2L-SOM is able to discover irregular and intertwined clusters, while conventional partitional clustering algorithms can deal with convex clusters only. Finally, the greatest advantage of our approach is that the number of clusters is determined automatically during the learning process, i.e., no a priori hypothesis for the number of clusters is required. Furthermore, the DS2L-SOM clustering algorithm is a powerful tool for visualization of the obtained segmentation in two dimensions. Clusters are easily and clearly identifiable, as well as regions without data (unconnected neurons). As we can notice from figures 2 to 5, the results obtained by the DS2L-SOM algorithm are very close to reality. Figures 3 and 4 show that DS2L-SOM is able to detect density-defined clusters.

In these figures, each hexagon represents a prototype of the SOM together with its associated data. Hexagons showing the same color are in the same cluster. White hexagons are not part of any cluster.

Figure 2: Clustering of "Spirals" data

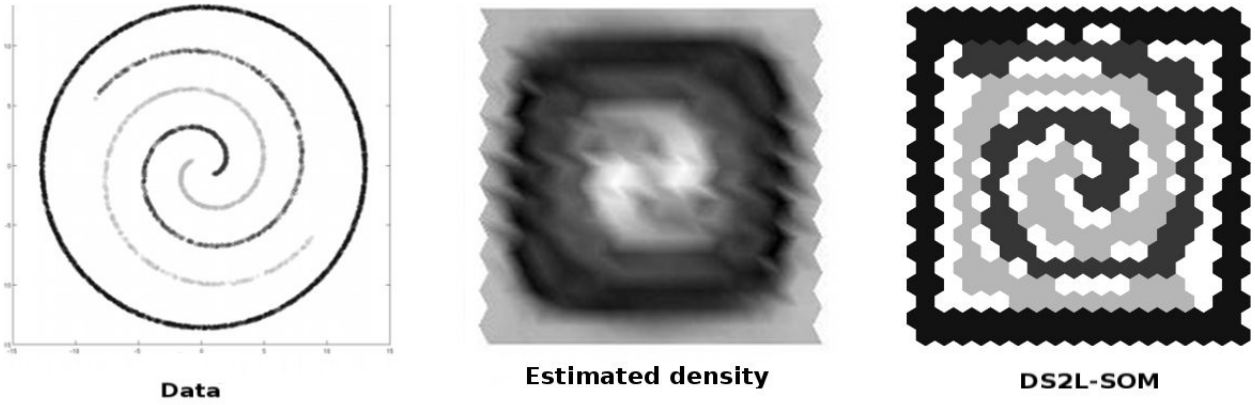


Figure 3: Clustering of "Engytime" data

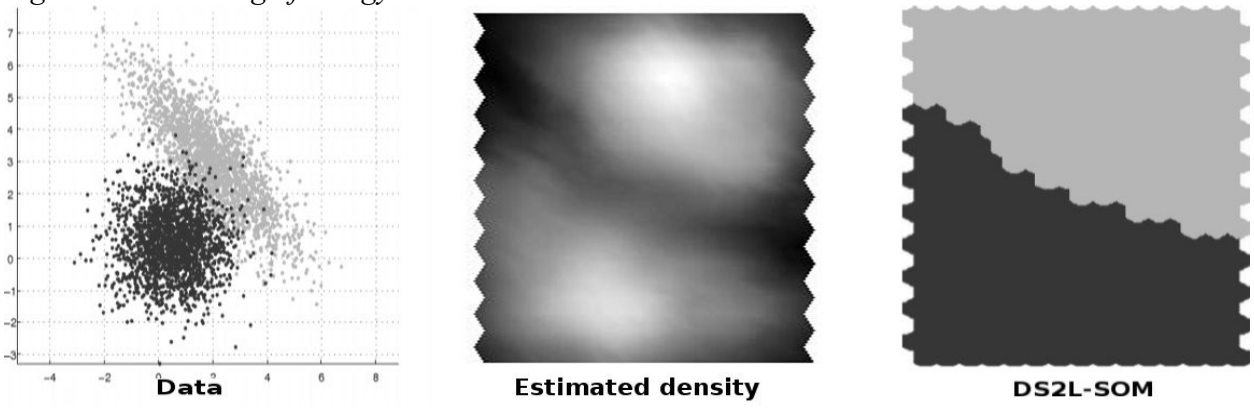


Figure 4: Clustering of "Wingnut" data

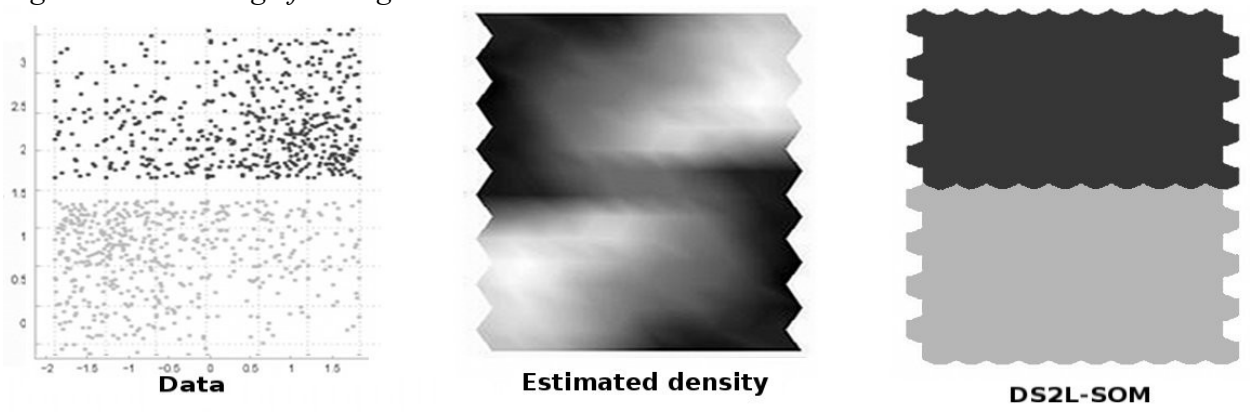
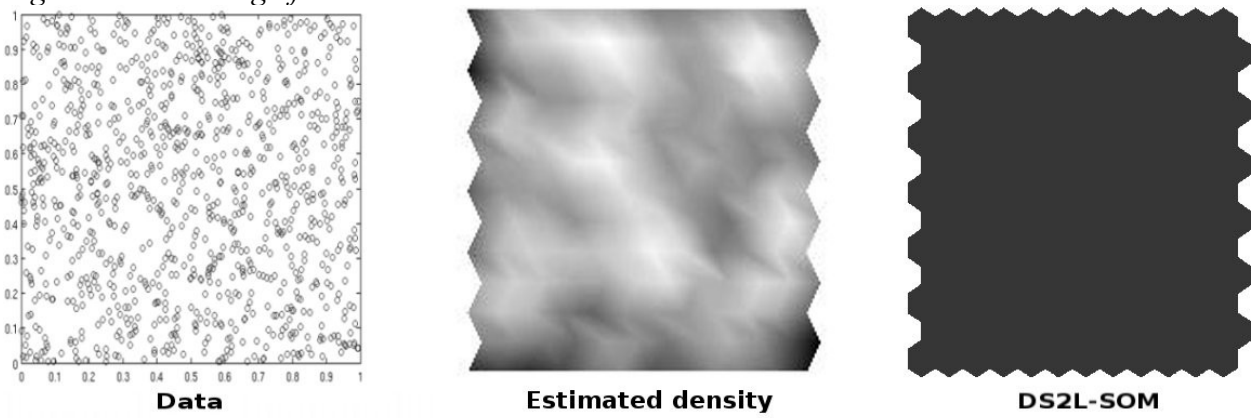


Figure 5: Clustering of "Random" data



RFID MONITORING OF AN ANTS COLONY

Among social animals, the *Formicidae* family certainly shows the greatest diversity of social structures and related behaviors. Its study is central in evolutionary biology: the "kin selection theory" (Hamilton, 1964) said that worker sterility is transmitted to the next generation through fertile kin. This criterion is fulfilled in some *Ponerines* society with simple familial structure (one queen mated with one male). This familial structure is the basis of the apparent harmony and cohesion of a colony.

The dynamic of task allocation (the process that results in specific workers being engaged in specific tasks) has widely been described in anterior theoretical studies, but the difficulty in acquiring the data and the lack of automatic tools discouraged the collection of associated experimental data. Yet, it is essential to find the rules that govern ant individual behavior and its integration at the colony scale. Understanding this phenomenon necessitates to be able to integrate different levels of analysis (from individual to colony). Thus, the individual monitoring of ant foragers (i.e. collectors and hunters) showed the elementary rules that each ant follows (Fresneau, 1985; Goss et al., 1989). Undoubtedly, the use of RFID technology will be very useful to obtain highly interesting results, such as a knowledge database about social behaviors and the analysis of its dynamic features.

However, RFID applied to ants poses some feasibility problems because weight limitations imply a good miniaturization of the tags and good performances of the readers. Streit et al. (Streit et al., 2003) recently used RFID technology to study bee longevity and the respective length of foraging and nurse activities. For this study, we chose a big-sized tropical ant *Pachycondyla tarsata*, which establish subterranean nests distributed in several interconnected chambers over 10m. Colonies of these species are typically composed of ten to a few thousand ants. In the colony, all individual have the same physical aptitudes and, in principle can assume any social role according to the social environment requirement. In normal condition, the colony lives in nest composed by several rooms, interconnected by tunnels, with a single entrance from the external environment. The eggs and the cocoons are kept in the most protected rooms, i.e. the most distant from the entrance. The nests are usually in the dark and the humidity level is very high.

RFID tag consists of a chip attached to an antenna weighting under 40 mg (i.e., 25% of an ant weight), glued on the animal thorax (Fig. 6). Preliminary tests showed that the tags don't disturb the ants' behavior and the colony dynamic significantly.

Measurement Device

A colony of *Pachycondyla tarsata* with a queen and 33 workers was monitored in the RFID device for 36 hours (about 270 000 scans). Each worker had a tag[1] attached to its thorax (Fig. 6).

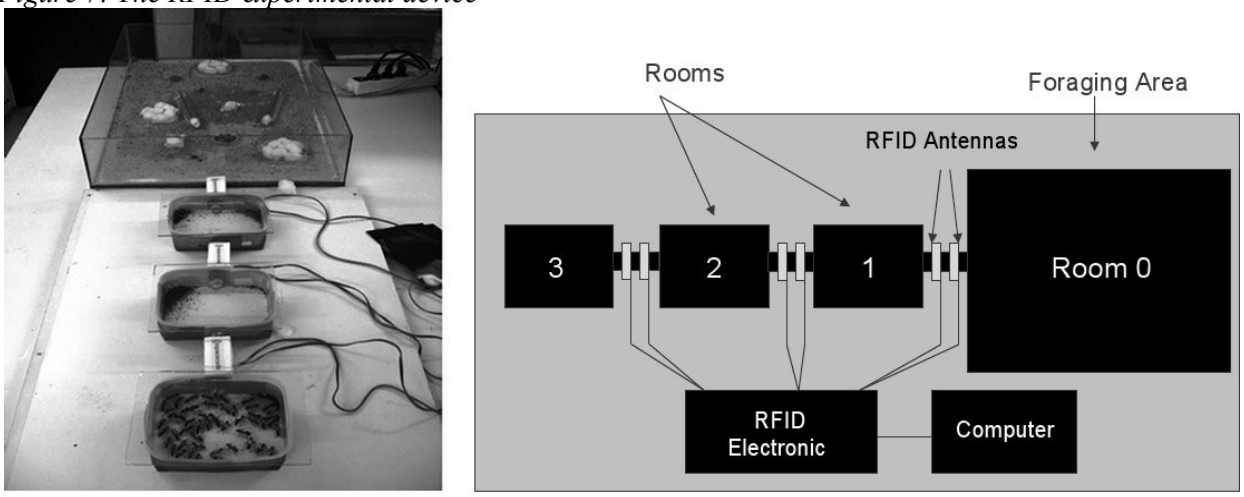
Figure 6: Ant with RFID tag

The experimental device is an artificial ant colony consisting of three rooms (Room 1, Room 2 and Room 3, dim 10 x 10cm) and a foraging area (Room 0, dim 0,5 X 1m), linearly connected by three tunnels (Fig. 7). The queen (not tagged) and its eggs stay permanently in Room 3, the farthest from the foraging area. Each tunnel is equipped with two RFID readers that detect the passage and the direction of tagged individuals moving between rooms. The position of an individual may be inferred unambiguously by the information provided by the six readers in the tunnels. The lack of detection implies that the individual is out of the tunnel and thus in one of the four rooms. The exact location of a tag (i.e., of an individual) can be deduced from the travel

1 Made by SpaceCode : http://www.spacecode-rfid.com/

direction. The information recorded by readers are handled by an RFID electronic, and then sent to a computer which creates and stores the data files.

Figure 7: The RFID experimental device

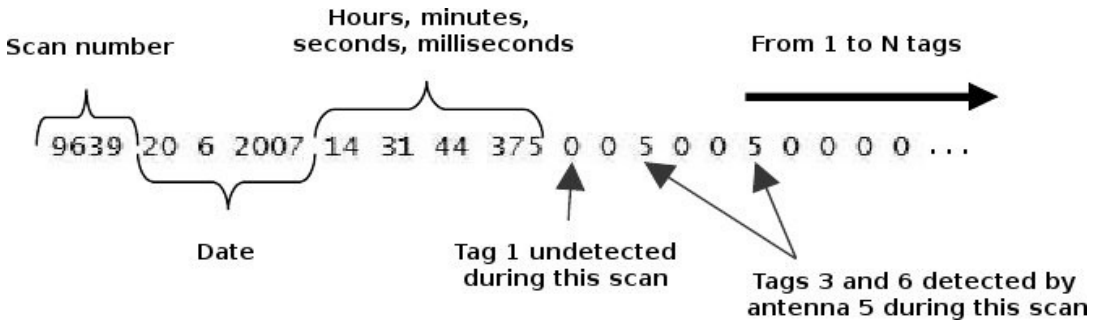


This experimental device allows us to monitor the amount of time spent by each colony member in each of the nest chambers. Indeed the different nest chambers segment social life in relation with the presence of the queen and brood. We have shown that the simple spatial monitoring of ants in the different chambers allows to predict with 99% probability the social status of individual workers (Fresneau et al., 1989). We aim at verifying and validating those results thanks to the RFID set-up. We also expect the spatial data collected to bring direct evidence of the existence of an organization between sub-castes with inactive individuals and more active ones, such as foragers.

Data

The data files are in text format. They indicate, for each antenna scan (about three scans per second), the scan number, the date, time, and, for each individual (i.e., for each tag), which antenna is activated (Fig. 8). If during a scan none is detected, nothing appears in the data file.

Figure 8: Example of a recorded scan in the data file



The recording system consists of four rooms (Room 0 to 3) connected to each other by three tunnels, each containing two RFID readers (antennas 1 to 6), which detect the passage of ants. If an ant moves from room 0 to room 1, it is detected successively by antennas 1 and 2. This allows us to infer the exact position of each ant at any moment (it is considered that an ant has changed room when it is detected by the second antenna). A simple treatment on these files makes it possible to obtain spatial information for each individual. For this study we take into consideration only the proportion of time spent in each room (time budget) for each individual.

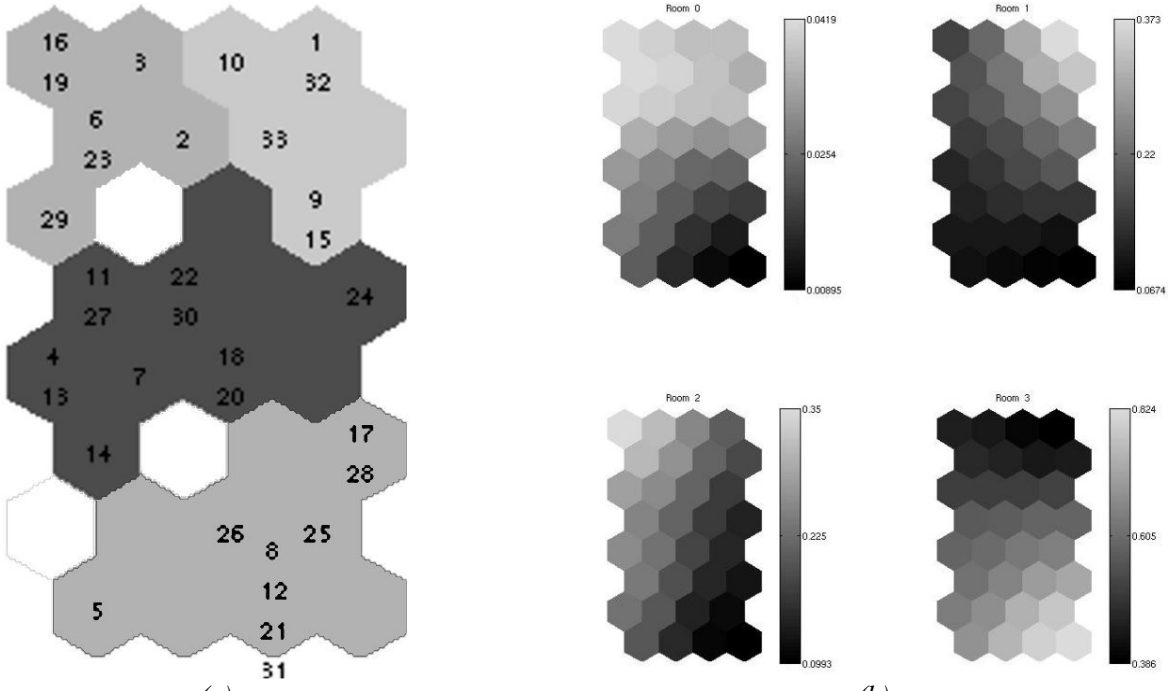
Therefore each individual was coded in a vectorial form, by the proportion of time spent in each area (i.e., 4 features in [0,1]), then we applied the algorithm DS2L-SOM on these data. The distance measure used by the algorithm for this study is the Chi^2 distance, more suitable than the Euclidean distance for proportion features.

Results

Figure 9(a) represents the map obtained with DS2L-SOM. Indeed, the DS2L-SOM clustering algorithm is a powerful tool for visualization of the obtained segmentation in two dimensions. Clusters are easily and clearly identifiable, as well as fields without data (unconnected neurons). In these figures, each hexagon represents a prototype of the SOM and its associated tags (i.e., ants). Two neighboring hexagons represent two similar prototypes and thus two similar behaviors.

The numbers inside the hexagons represent the tags associated to this prototype. Hexagons that share a color in the Fig. 9(a) belong to the same cluster, a cluster represents a set of ants which share a distinct behavior. White hexagons are not part of any cluster. The final segmentation of the map shows four types of behavior with regards to the occupation of rooms (Fig. 9(a)).

Figure 9: (a) map obtained from RFID data, (b) profiles of prototypes

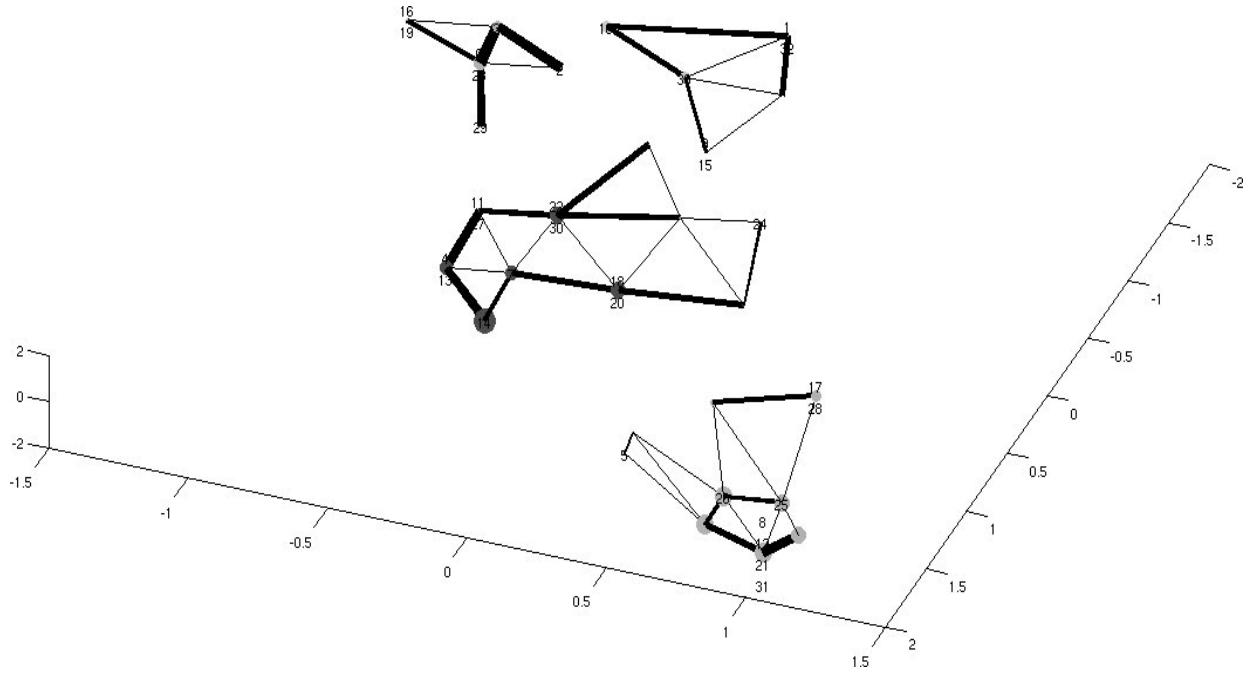


(a) *(b)*

We can characterize each type depending on the profile of their prototypes (Fig. 9(b)). Light hexagons (unlike dark ones) mean that individuals represented by this prototype spend much time in the related Room. Individuals that are represented in the "south" of the map are characterized by a significant occupation of Room 3 (the Queen's room), to the exclusion of others. On the contrary, the ants in the "north" spend more time in the foraging area, but while the "north-east" ants spend a lot of time in Room 1, "north-west" ants stay longer in the Room 2. Finally, individuals from the group represented in the "center" of the map present an intermediate profile, they are not characterized by a particular room occupancy when compared to other ants.

The representation according to a Sammon mapping of the prototypes (Fig.10) allows a more detailed analysis of the structure of each group and their relationships. Indeed, the clustering is accompanied by a set of information that may be used to expand the data analysis, such as the matrix of distances between prototypes, the density matrix and also the values of connections that can be used to determine the relative importance of each prototype for the representation of the data. Moreover, the map provides information on the relationships between groups, two groups close on the map being more similar than two distant groups. Finally, the presence of some unrepresentative prototypes (with null connections values between them and their neighbors) gives an idea about the shape of groups in the input space. We can represent all of this information in a single figure (figure10).

Figure 10: Sammon mapping of prototypes and their connections, from the RFID data



The spheres represent the prototypes shown in a three-dimensional space by a non-linear Sammon mapping (Sammon Jr., 1969) respecting the distances between prototypes. The spheres sizes are proportional to the density associated with each prototype. Colors depend on the associated cluster and connection thickness is proportional to the value associated with these connections.

With this additional information, we can see that the "south" group is composed of a dense nucleus of individuals (i.e., 8, 12, 21, 31, 26 and 25) which well representing their group (i.e., associated with well-connected prototypes) and a set of marginal individuals (5, 17 and 28) compared to that group (associated with prototypes little connected to the other). Prototypes of the nucleus are very close to each other and distant from other groups. This means that the individuals represented are similar to each other and highly specialized in their occupation of space (here Room 3, the queen's room). Marginal individuals show an intermediate behavior compared with the other groups: they are less specialized than the nucleus members.

Individuals in the "north-west" group are also well specialized in their space occupation, i.e., the Room 2 and the foraging area. Their prototypes are close and well-connected and, in particular, individuals 6 and 23 well representing their group. In contrast, individuals of the "north-east" group present a less specialized behavior. Their prototypes are more distant from one another, while properly connected.

Finally, the "center" group characterizes generalist ants. Most of the prototypes are well connected to the others and there is no marginal behavior (except for individual 24), but there is also an important distance between these prototypes, indicating a wide variety of behaviors between individuals in the group.

This segmentation undoubtedly results from the division of labor between individuals in the colony. The ants in the "north-west" group must be specialized in foraging and food processing. The "north-east" ants, which spend less time outdoors and have a more diverse spatial distribution probably handle maintenance tasks, while ants in the "south" group take care of the queen and brood. The "center" group could be composed of low-skilled individuals which have a more versatile activity (or maybe no particular activity) in the colony.

This study was replicated on an orphaned colony where the queen's absence is compensated with several workers eggs laying. The clustering generated strictly in the same results.

STRUCTURE OF DISPLACEMENTS DURING A NEST CHANGE

The content of his section follows from the study on social organization. It aims at studying the mechanisms leading to a colony changing nest. Migration is a widespread phenomenon in many species, but it remains a risky event because during the movement the queen and brood will be particularly vulnerable. The strategies used in nest choice and movement organization are therefore crucial for group survival. For example, the time cost of the queen's movement and the order of brood transport are two of the variables we aim at quantifying thanks to appropriate RFID devices.

We followed and analyzed the movement of a *Ponerine* ant colony composed of great sized ants easy to mark and follow. Ponerine ants are tropical species whose colonies consist in a few hundred to a few thousand individuals (Peeters, 1993). In these species, queens are often not very different from workers, but a wide variety of social structure exists, ranging from monogyny to polygyny, with some species even lacking a specific queen caste (Peeters, 1997; Monnin & Peeters, 1998). Our study species, *Pachycondyla tarsata*, establish monogyne colonies of up to 2500 monomorphic workers (Hölldobler, 1984). They build underground nests with interconnected chambers on surfaces as big as 1200 m^2, a 130m tunnel between several chambers has even been observed (Braun et al., 1994). Peripheral nests constitutes advanced positions which strengthen the network. The queen and brood are particularly mobile in order to optimize brood rearing. In the wild, solitary foragers look for termites through visual orientation (Hölldobler, 1980). They also exhibit mass recruitment through trail pheromones when an

important food source is discovered (Hölldobler, 1984) and prey transport is allocated between different foragers who do only part of the return trip. Workers also collect insect remains and contributes to resource turnover (Dejean et al., 1993). It is therefore an interesting model for laboratory studies of colony movements.

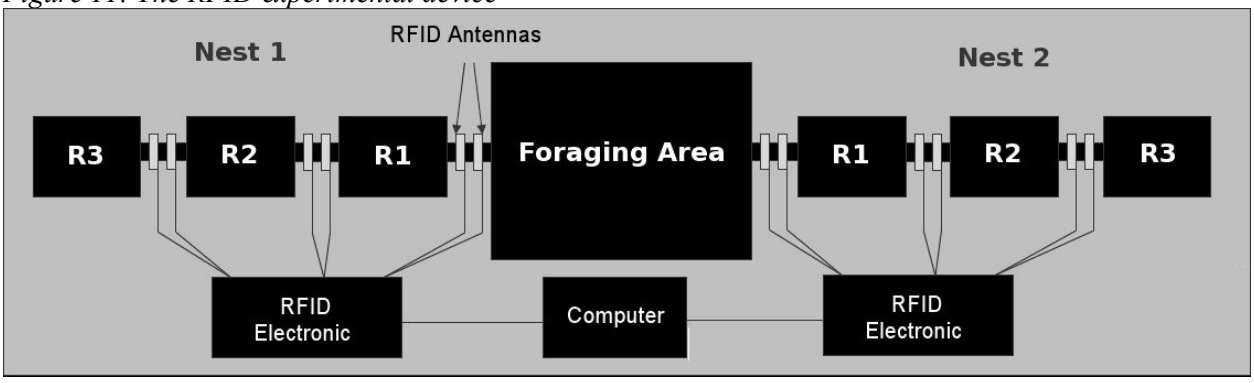
Measurement Device

The movement between nests of a colony of 55 *Pachycondyla tarsata* workers was monitored in the RFID device (about 4 hours recording). Each worker had a tag attached to its thorax.

The experimental device for this experiment consists of two artificial nests (N1 and N2) of three rooms each (Room 1, 2 and 3) and a foraging area, linearly connected by six tunnels (Fig. 11). At the beginning of the experiment, the brood is located in Room 3 of the first nest, the farthest from the foraging area. Each tunnel is equipped with two RFID readers (number 1 to 12 from Room 3 in Nest 2 to Room 3 in Nest 1) that detect the passage and the direction of tagged individuals between rooms. The information recorded by readers are handled by two RFID electronics and then sent to a computer which creates and store the data files.

At *time*=0 we switch on a strong neon light over the first nest and we open the entrance of the second nest, then we record the colony movement until the entire brood is moved into the second nest (~ 4 hours).

Figure 11: The RFID experimental device



The kinetics of departures from the original nest, passages through the foraging area and arrival sequences in the new protected nest will be studied. These movements necessitate a high coordination between workers and it is likely that foragers initiates the exploration of the new nest. It is also likely that foragers will initiate recruitments to lead more static ants (nurses and inactive ants) toward the new nest.

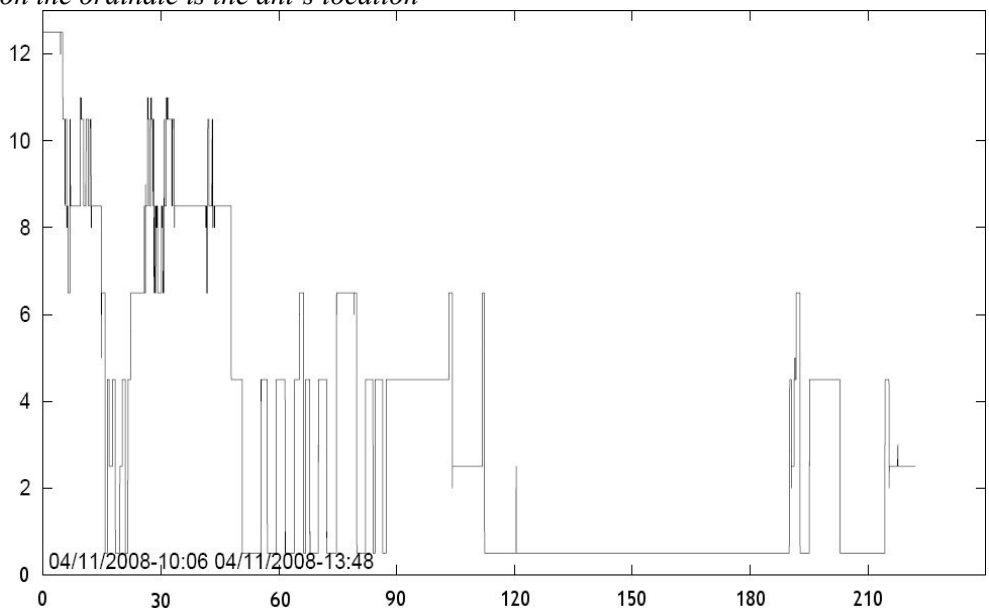
It is unknown that which ants will transport the brood and more importantly which ants will take care of the queen movement toward the new nest. Similarly, we do not know whether the established division of labor in a stable situation will influence particular individuals to play a key role in the organization of colony movement. Ultimately, what we focus on here is the logistics of dynamic sequences of movements between the original and the new nest. It is particularly difficult to follow this phenomenon in real-time and only a RFID system will allow the acquisition of such data and the extraction of significant patterns from them.

Data post-processing

The data files are in the same format as in part 3. They indicate, for each antenna scan (about three scans per second), the scan number, the date, time, and, for each individual (i.e., for each tag), which antenna is activated. If during a scan, none is detected, nothing appears in the data file. If an ant moves from one room to another, it is detected by two successive antennas, and this allows us to infer the exact position of each ant at any moment.

We used this information to produce the individual moving sequence of each ant. This sequence is a function that gives the ant's location at any time during the move (see Fig. 12 for an example). The location is coded by a reader number if the ant is under it and a float number to represent a room between two readers (for example if the ant is in Room 2 in Nest 2, between the readers 2 and 3, the location will be 2.5).

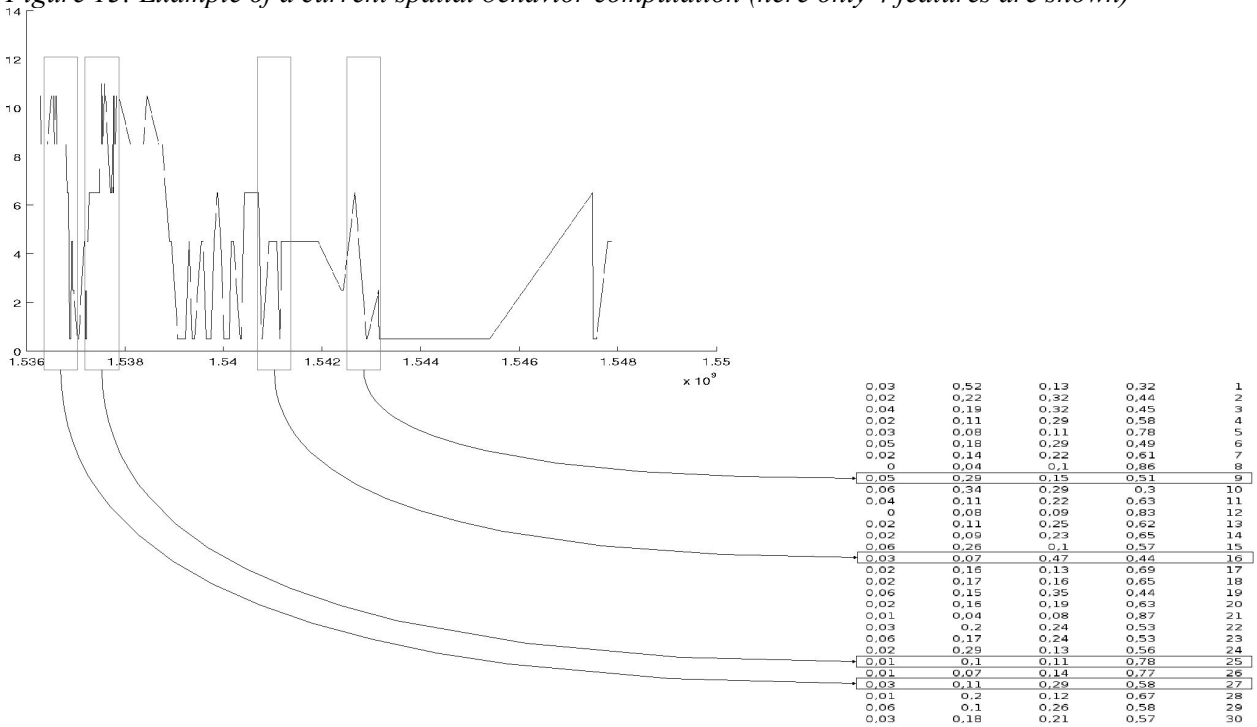
Figure 12: Example of a moving sequence : on the abscissa is the time from the start of the experiment (minutes), on the ordinate is the ant's location



However, what we would like to analyze is the variation of the ant's current spatial behavior over time. To do that, a current spatial behavior must be defined. Here, we can't just choose the current location, because in this way we would lose all dynamic information such as "the ant is moving quickly" or "the ant makes round trips between two rooms". Therefore, we define the current behavior as the time spent in each location (static information) and the number of exits from each location (dynamic information) during a 10 minutes time window centered on the current time (Fig. 13).

Obviously, this definition implies some correlations between the description of two current behaviors if they are separated by less than 10 minutes, as the two related windows overlap. This allow us to detect sudden changes in behavior. As there are 19 locations in the RFID device (7 Rooms and 12 readers), each temporal windows is coded in a vectorial form of 38 normalized features (one static and one dynamic feature for each location).

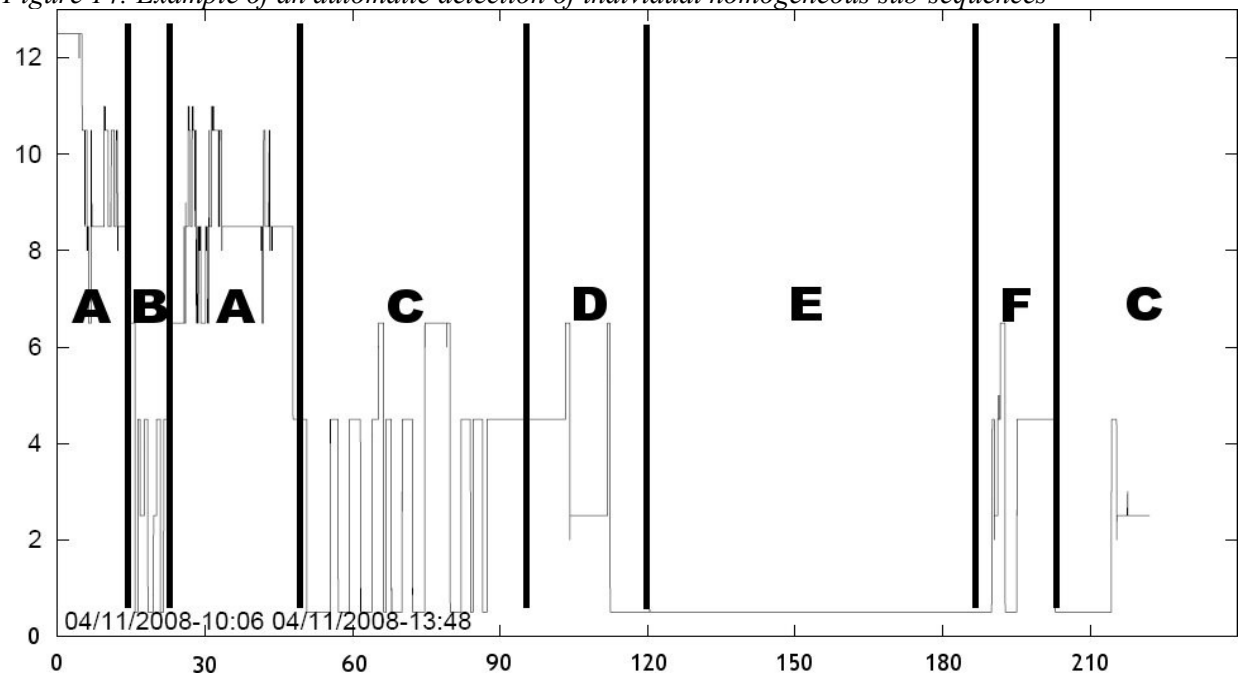
Figure 13: Example of a current spatial behavior computation (here only 4 features are shown)



Detection of individual homogeneous sub-sequences

In order to regroup similar current behaviors and to detect changes in current behaviors over time, we applied the algorithm DS2L-SOM on time windows from each individual sequences. The distance measure used by the algorithm for this experiment is the Euclidean distance. Figure 14 gives an example of the results obtained with DS2L-SOM for one ant.

Figure 14: Example of an automatic detection of individual homogeneous sub-sequences



When the light is switched on, the ant start to move quickly inside Nest 1 (location 7 and more) with some short outing in the foraging area (location 6.5). This homogeneous behavior was automatically found by DS2L-SOM, it's a set of current behaviors regrouped into one cluster (called A by the algorithm), we will call it "behavior A". After about 15 minutes, this ant moves from Nest 1 to Nest 2 and then moves quickly inside Nest 2. This has been detected as a homogeneous behavior (called B). After that the ant comes back to Nest 1 and shows a behavior very similar to the first one. DS2L-SOM automatically detects that the ant behaved in the same way again and gives it the same name. Then the ant comes back to Nest 2, but doesn't move as fast as in behavior B, constituting another behavior (called C). The ant stays in Nest 2 until the end of the experiment, expressing different behaviors (varying in static and dynamic features).

Thus, the proposed method is a powerful tool for detecting homogeneous behavioral sub-sequence inside a spatio-temporal RFID monitoring.

Detection of similar sub-sequences

The method used in part 4.3 allows us to analyze efficiently the behavior of each individual. We now need a method to compare all these individual sequences so as to perform an analysis at the collective level. The idea is to define a similarity measure between two set of prototypes (from DS2L-SOM) that represent two individual sub-sequences (clusters).

Let *dSim*(Ci, Cj) be the dissimilarity between the set of N_i prototype that represent the cluster Ci and the set of N_j prototypes that represent the cluster Cj :

$$dSim(C_i, C_j) = \frac{\sum_{k=1}^{N_i} Dn_i^{(k)} \parallel w_i^{(k)} - w_j^{(k^*)} \parallel^2 \left(Dn_i^{(k)} - Dn_j^{(k^*)} \right)^2}{N_i}$$

With Dn_i is the normalized density ($D^{(i)}$ / number of data in cluster i), k* is the first BMU in cluster j of the unit k.

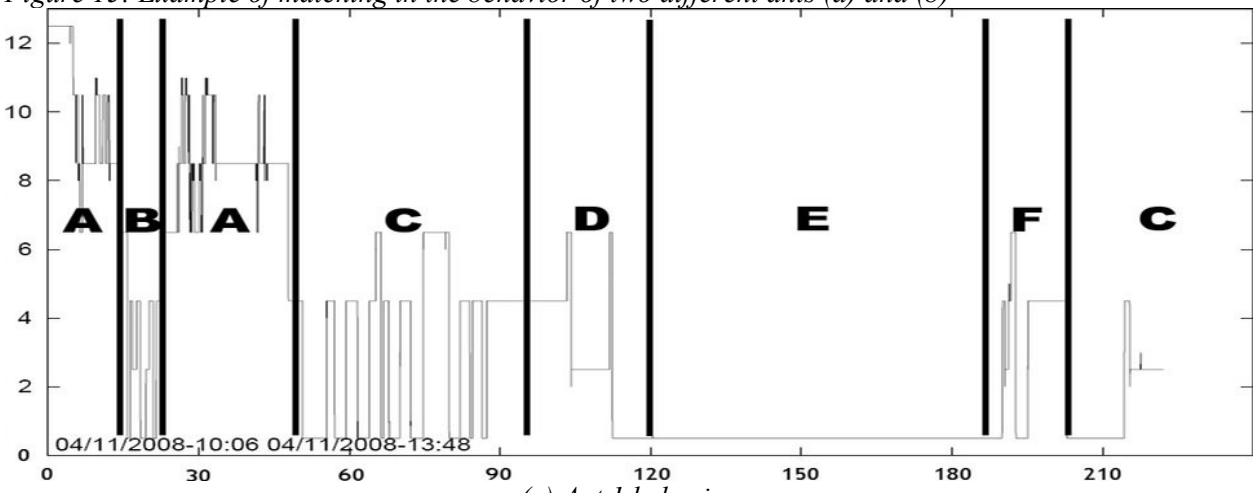
Let's note that the measure is not symmetric, as *dSim*(Ci, Cj) ≠ *dSim*(Cj, Ci). Therefore we define symmetric similarity *S* as :

$$dS(\text{Ci, Cj}) = (dSim(\text{Ci, Cj}) + dSim(\text{Cj, Ci}))/2$$

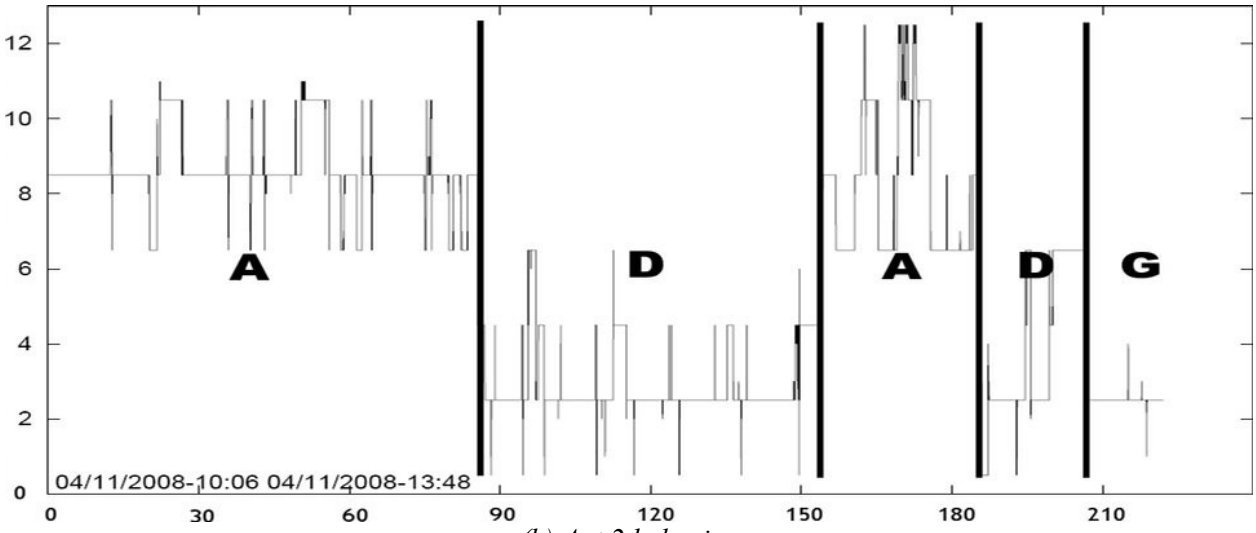
We used this measure to compute a similarity matrix with all the subsequences of all the ants. Then we computed a modified version of SOM that learn prototypes from a similarity matrix (Conan-Guez et al., 2006). The prototypes now represent a set of individual sub-sequences. We use DS2L-SOM on theses prototypes to find clusters of homogeneous subsequences. This allows us to compare the behaviors of different ants.

These clusters are then used to rename all the subsequences, so as to give the same name to subsequences that belong to the same cluster. From more than 600 individual subsequences, we found 45 clusters that represent different homogeneous behaviors. Fig. 15 show an example of matching obtained with this method: the first and the third behavior of ant 1 is automatically found to be the same as the first and the third behavior of ant 2. The method thus gives the same label to this four subsequences (here the label is A). It's the same for label D. On the opposite, some behaviors are only expressed by one of the two ants, like behaviors B, C, E, F and G, but they can be found in some other ant sequences.

Figure 15: Example of matching in the behavior of two different ants (a) and (b)



(a) Ant 1 behavior



(b) Ant 2 behavior

Validity of the proposed method

We propose a new algorithm for modeling trajectory structure (i.e. trajectory segmentation), which is based on prototypes, and a measure of dissimilarity between clusters. The advantages of this algorithm are not only the low computational cost and the low memory requirement, but also the high accuracy achieved in fitting the structure of the modeled trajectory, in comparison to the usual trajectory segmentation methods.

In order to check the validity of the obtained results, we compared it with some visual observations from a video record of the movement. A movie camera was placed over the foraging area and every ant moving across this area is filmed. This allowed us to detect only one apparent behavior: the transportation of larva and cocoon. Each ant can be identified visually thanks to some color painted on their tag. So we know which ant does transport and at what time this behavior occurred. We compared this with the results of the automatic analysis and we found that all the transportations subsequences were grouped into only 2 clusters (*S* an *T*). Thus all the

transportation sequences are combination of clusters *S* and *T* (Tab. 1). Moreover, only 2 ants have an *S* or a *T* subsequence without having been seen transporting (i.e. less than a 5% error).

Table 1: Corresponding between visual observation and automatic analysis (Number of ants)

	Transportation	Others
S + T	8	2
Others	0	42

This result shows the reliability in the clustering found by our automatic method. The difference between behavior *T* and *S* depends on where the ant gets the cocoon or larva: if it finds it in Room 3, the subsequence will belong to the cluster *T*, otherwise it will belong to the cluster *S*.

Thus, at the end of the automatic process, we have enough information to compare very efficiently the behavior of all ants in the colony during the move. This kind of information is very useful to biologists to understand complex behavioral systems like the ones observed in ant colonies. Our results are very interesting because they provide a description of each individual during the migration. It is important to know that our results rely on a blind analysis of the data. The comparison of those results with the observation reveals an excellent match between the automated data analysis and visual video recording analysis but with considerable time saving.

CONCLUSION

The new unsupervised clustering method (DS2L-SOM) used in this article is a very efficient data mining and visualization tool for behavioral studies based on RFID technology. It allows to discover groups defined either by distance or by density, whatever their form or the difference of densities between the groups or within a group. It is quite fast, suitable for continuous learning and allows a very simple and effective visualization with a non-linear projection of the data structure on a two-dimensional map. We were able to highlight the characteristics of spatial organization in ant colonies. Our approach also allows a detailed description of the characteristic behavior of every group of individuals. These descriptions allowed us to associate to each of these groups a social task. These deductions are perfectly compatible with the results of previous works using classic methods (Fresneau & Dupuy, 1988; Fresneau et al., 1989). Furthermore, the method allows to split different sequences during colony movement and to monitor individual performances which are impossible to evidence manually. Although we used few individuals in this study, the method is perfectly suitable for the study of thousands of individuals, with behaviors described by a large number of spatio-temporal parameters. The example we developed here can be applied to multiple situations where the control of individual abilities as well as their integration at the collective scale are needed, like in the study of flux regulation in social groups (crowd control, mass hysteria panic). Therefore, DS2L-SOM is a very powerful tool for processing and visualizing RFID data in experimental studies.

The next step in our study will be to model the sequences segmentation found in part 4.4 with a Markov Model method (Rabiner & Juang, 1986), in order to be able to describe very precisely the probabilistic behavior of each ant. This will also allow us to calculate probabilistic similarity between two sequences and we will try to find some correlation between the usual social organization and the actual behavior of the colony during movement. We also intend to carry on

our research using a new RFID system that allows to detect individual presence in different part of the foraging area, which constitutes a link between the colony and its environment. This will be used for a real-time monitoring of foraging trips which will help understanding ant behavior better. The research applications of this type of spatial data are numerous and promising. The success of the first experiments make it possible to gather databases of real-time movements in ants. These databases will allow us to unravel the complex social organisation of ant societies using the "numeric learning" techniques we developed.

In addition, by continuing the automation of these tracking systems it will be possible to control experimental devices which allow the modification of the environment according to the identity and the history of the individuals (controlled accesses to specific sectors or triggering of specific stimuli or reinforcement).

ACKNOWLEDGMENT

This work was supported in part by the *Sillages* project (N° ANR - 05 - BLAN - 017701) and the *CADI* project (N° ANR - 07 TLOG 003) financed by the ANR (Agence Nationale de la Recherche).

REFERENCES

Aupetit, M. (2005). Learning topology with the generative gaussian graph and the EM algorithm. In *Neural Information Processing Systems (NIPS).* Vancouver, B.C., Canada.

Bohez, E. L. J. (1998). Two level cluster analysis based on fractal dimension and iterated function systems (ifs) for speech signal recognition. *IEEE Asia-Pacific Conference on Circuits and Systems*, (pp. 291–294).

Braun, U., Peeters, C., & Hölldobler, B. (1994). The Giant Nests of the African Stink Ant Paltothyreus tarsatus (Formicidae, Ponerinae). *Biotropica*, 26(3), 308–311.

Cabanes, G. & Bennani, Y. (2007). A simultaneous two-level clustering algorithm for automatic model selection. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA'07)* (pp. 1176–1182). Cincinnati, Ohio, USA.

Cabanes, G. & Bennani, Y. (2008). A local density-based simultaneous two-level algorithm for topographic clustering. In *International Joint Conference on Neural Network (IJCNN),* Hong-Kong, China.

Conan-Guez, B., Rossi, F., & El Golli, A. (2006). Fast algorithm and implementation of dissimilarity self-organizing maps. *Neural Networks*, 19(6–7), 855–863.

Dejean, A., Beugnon, G., & Lachaud, J.-P. (1993). Spatial components of foraging behaviour in an African ponerine ant, Paltothyreus tarsatus. *Journal of Insect Behaviour*, 6, 271–285.

Fresneau, D. (1985). Individual foraging path fidelity: a novel strategy in a ponerine ant. *Ins. Soc.*, 32, 109–116.

Fresneau, D., Corbara, B., & Lachaud, J. (1989). Organisation Sociale et Structuration Spatiale Autour du Couvain chez Pachycondyla apicalis. *Actes coll. Insectes Sociaux*, 5, 83–92.

Fresneau, D. & Dupuy, P. (1988). Behavioural study of the primitive ant Neoponera apicalis. *Anim. Behav*., 36, 1389–1399.

Goss, S., Fresneau, D., Deneubourg, J.-L., Lachaud, J.-P., & Valenzuela-Gonzalez, J. (1989). Individual foraging in the ant Pachycondyla apicalis. *Oecologia*, 80, 65–69.

Guérif, S. & Bennani, Y. (2006). Selection of clusters number and features subset during a two-levels clustering task. In *Proceeding of the 10th International Conference Artificial intelligence and Soft Computing 2006* (pp. 28–33). Palma de Mallorca, Spain.

Hamilton, W. (1964). The genetical evolution of social behaviour. J*ournal of Theoretical Biology*, 7, 1–52.

Hölldobler, B. (1980). Canopy orientation: a new kind of orientation in ants. *Science*, 210, 86–88.

Hölldobler, B. (1984). Communication during foraging and nest-relocation in the African stink ant, Paltothyreus tarsatus. *Z.Tierpsychol*, 65, 40–52.

Hölldobler, B. & Wilson, E. (1990). *The ants*. Cambridge, MA: Harvard University Press.

Hussin, M. F., Kamel, M. S., & Nagi, M. H. (2004). An efficient two-level SOMART document clustering through dimensionality reduction. In *ICONIP* (pp. 158–165).

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Berlin: Springer-Verlag.

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Korkmaz, E. E. (2006). A two-level clustering method using linear linkage encoding. *International Conference on Parallel Problem Solving From Nature, Lecture Notes in Computer Science*, 4193, 681–690.

Martinetz, T. (1993). Competitive Hebbian Learning rule forms perfectly topology preserving maps. In S. Gielen & B. Kappen (Eds.), *Proceedings of the International Conference on Artificial Neural Networks (ICANN-93)*, Amsterdam (pp. 427–434). Heidelberg: Springer.

Monnin, T. & Peeters, C. (1998). Monogyny and regulation of worker mating in the queenless ant Dinoponera quadriceps. *Animal. Behavior*, 55, 299–306.

Pamudurthy, S. R., Chandrakala, S., & Sakhar, C. C. (2007). Local density estimation based clustering. *Prodeeding of International Joint Conference on Neural Networks*, (pp. 1338–1343).

Passera, L. & Aron, S. (2005). Les fourmis: comportement, organisation sociale et évolution. Ottawa: *Les Presses scientifiques du CRNC*.

Peeters, C. (1993). Monogyny and polygyny in ponerine ants with or without queens. In L. Keller (Ed.), *Queen Number and Sociality in Insects* (pp. 235–261). Oxford: Oxford University Press.

Peeters, C. (1997). Morphologically "primitive" ants: comparative review of social characters, and the importance of queen-worker dimorphism. In J. Choe & B. Crespi (Eds.), *The Evolution of Social Behaviour in Insects and Arachnids*. Cambridge: Cambridge.

Rabiner, L. R. & Juang, B. H. (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, (pp. 4–16).

Sammon Jr., J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computer*, 18(5), 401–409.

Streit, S., Bock, F., Pirk, C., & Tautz, J. (2003). Automatic life-long monitoring of individual insect behaviour now possible. *Zoology*, 106, 169–171.

Ultsch, A. (2005). Clustering with SOM: U*C. In *Proc. Workshop on Self-Organizing Maps* (pp. 75–82), Paris, France.

Vincent, L. & Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13, 583–598.

Yue, S.-H., Li, P., Guo, J.-D., & Zhou, S.-G. (2004). Using greedy algorithm: DBSCAN revisited II. *Journal of Zhejiang University SCIENCE*, 5(11), 1405–1412.