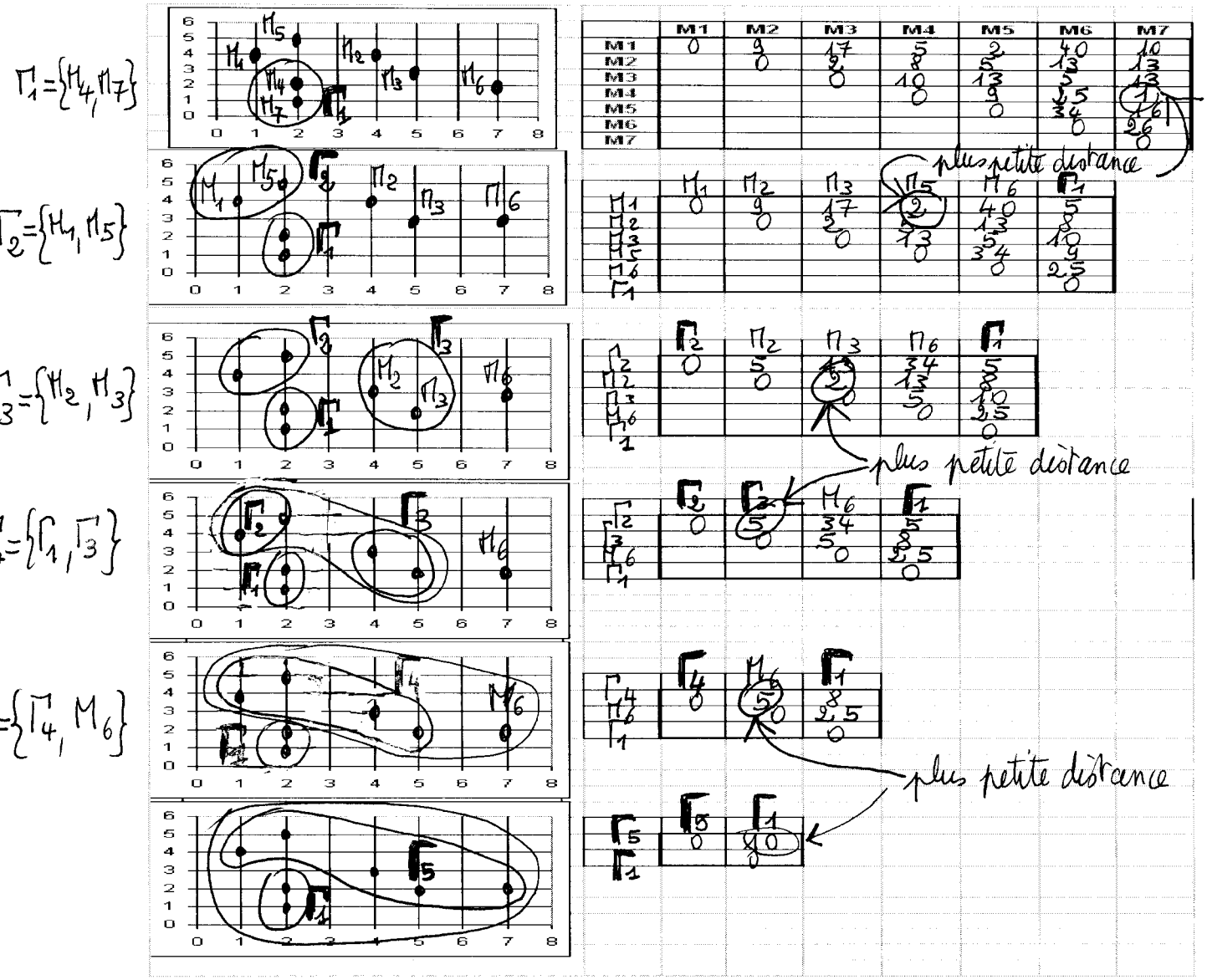
NOM :
PRENOM : Corrigé

Groupe :

Semaine du 26 mars 2012

Mathématiques pour la Biologie (semestre 2) : Feuille-réponses du TD 7
Classification hierarchique ascendante

Exercice 1. : On se propose de réaliser une classification des 7 points suivants en utilisant la méthode d'agglomération au plus proche voisin : $M_1 = (1\,;4)$, $M_2 = (4\,;4)$, $M_3 = (5\,;3)$, $M_4 = (2\,;2)$, $M_5 = (2\,;5)$, $M_6 = (7\,;2)$ et $M_7 = (2\,;1)$.



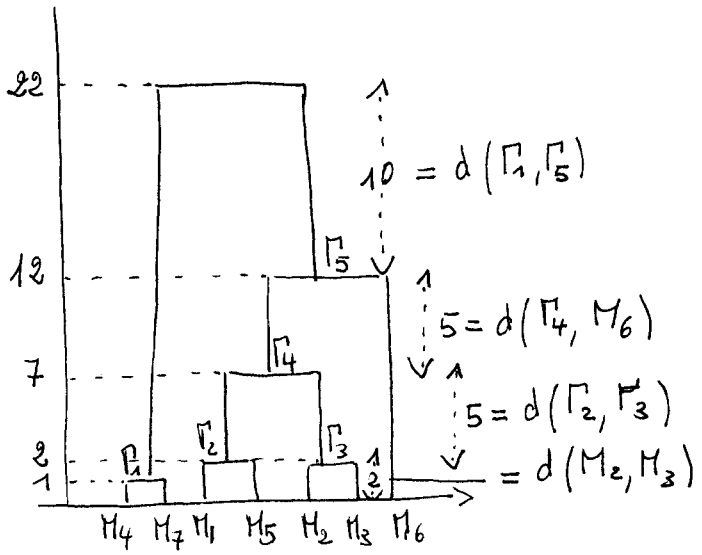
1. Calculer le carré de la distance euclidienne de M_1 à M_2 et de M_1 à M_3.

$d^2(M_1, M_2) = (4-1)^2 + (4-4)^2 = 9$

$d^2(M_1, M_3) = (5-1)^2 + (3-4)^2 = 16 + 1 = 17$

2. Compléter le premier tableau à droite représentant la matrice des distances des points tracés à gauche, en utilisant le carré de la distance euclidienne.

3. Sur le second dessin, agglomérer, en les entourant d'une courbe, les deux points les plus proches pour former une classe, Γ_1, puis compléter la deuxième matrice de distance en calculant notamment les distances (au plus proche voisin) de la nouvelle classe avec les 5 autres points.
4. Poursuivre la classification en complétant les tableaux suivants et en cerclant les classes, $\Gamma_2, \ldots$ créées au fur et à mesure.
5. Tracer un dendrogramme résumant cette classification.



Exercice 2. : *(Sujet inspiré d'un article de John Hartshorne, paru dans le journal de la "British Ecological Society")*

Un laboratoire d'écologie étudie les espèces micro-animales (larves, ..) présentes dans les rivières et les étangs. Il réalise, dans 6 sites de rivière, notés $R1$, $R2$, $R3$, $R4$, $R5$ et $R6$, et 3 sites d'étangs, notés $E1$, $E2$ et $E3$, des prélèvements répétés qui lui permettent d'avancer une liste des espèces présentes dans chacun de ces sites et de repérer les espèces présentes dans plusieurs sites à la fois. La matrice suivante contient, pour chaque paire de sites A et B, le nombre d'espèces communes aux 2 sites. Ainsi on y lit par exemple que 11 espèces sont présentes au site $R1$ et qu'il y a 7 espèces présentes à la fois au site $R1$ et au site $R2$.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	11	7	4	6	6	7	4	4	3
$R2$	7	15	8	8	9	6	3	3	2
$R3$	4	8	13	7	7	4	2	3	2
$R4$	6	8	7	15	7	6	6	8	6
$R5$	6	9	7	7	12	4	3	5	4
$R6$	7	6	4	6	4	10	6	5	5
$E1$	4	3	2	6	3	6	13	10	9
$E2$	4	3	3	8	5	5	10	15	11
$E3$	3	2	2	6	4	5	9	11	12

On se propose de regrouper les 9 sites en trois ou quatre classes composées de sites où ce sont pratiquement les mêmes espèces qui sont présentes. Pour réaliser cette classification, on propose de mesurer la distance entre deux sites A et B par la formule

$$d(A,B) = \frac{n_A + n_B - 2n_{AB}}{n_A + n_B}$$

où n_A (resp. n_B) désigne le nombre d'espèces présentes au site A (resp. au site B) et n_{AB} le nombre d'espèces en commun entre les sites A et B.

1. Calculer $d(R_1, R_3)$ puis $d(R_2, E_1)$.

On utilise le tableau, n_A et n_B se lisent sur la diagonale.

$$d(R_1, R_3) = \frac{11+13-2\times 4}{11+13} = \frac{16}{24} = \frac{2}{3} \simeq 0{,}666$$

$$d(R_2, E_1) = \frac{15+13-2\times 3}{15+13} = \frac{22}{28} \simeq 0{,}786.$$

2. Quelle est la distance de deux sites n'ayant pas d'espèces en commun et celle de deux sites ayant les mêmes espèces exactement ?

- S'il n'y a pas d'espèces en commun entre le site A et le site B alors $n_{AB} = 0$ donc $d(A,B) = \frac{n_A + n_B}{n_A + n_B} = 1$
- Si les deux sites ont exactement les mêmes spèces, alors $n_A = n_B$ et $n_{AB} = n_A = n_B$. Donc $d(A,B) = 0$

0 et 1 sont les deux valeurs maximale et minimale pour $d(A,B)$.

3. Après calculs, on obtient la matrice des distances suivante. Pouvez-vous compléter la colonne manquante de la matrice des distances suivante sans calcul ? Expliquez.

on se souvient que la matrice des distances est symétrique.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	0	0,462	0,666	0,538	0,478	0,334	0,666	0,692	0,74
$R2$	0,462	0	0,428	0,466	0,334	0,52	0,786	0,8	0,852
$R3$	0,666	0,428	0	0,5	0,44	0,652	0,846	0,786	0,84
$R4$	0,538	0,466	0,5	0	0,481	0,52	0,571	0,466	0,556
$R5$	0,478	0,334	0,44	0,481	0	0,636	0,76	0,63	0,666
$R6$	0,334	0,52	0,652	0,52	0,636	0	0,478	0,6	0,546
$E1$	0,666	0,786	0,846	0,572	0,76	0,478	0	0,285	0,28
$E2$	0,692	0,8	0,786	0,466	0,63	0,6	0,285	0	0,185
$E3$	0,74	0,852	0,84	0,556	0,666	0,546	0,28	0,185	0

4. La classification conduit au dendrogramme représenté ci-dessous.

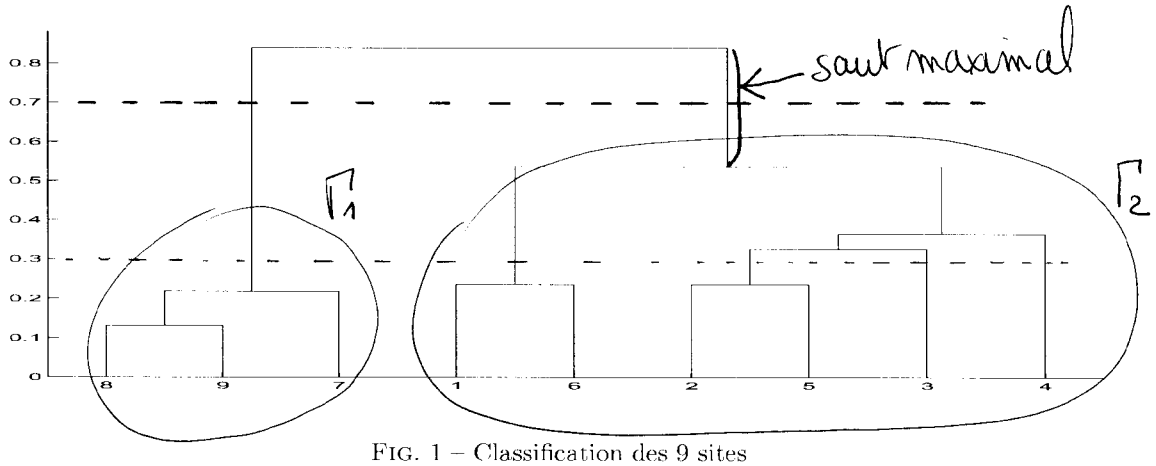


FIG. 1 – Classification des 9 sites

Décrire la composition des classes de la partition obtenue en coupant ce dendrogramme à la hauteur 0.7.

On obtient une partition en deux classes (qui distingue les sites d'étang et de rivière)

$\Gamma_1 = \{E_1, E_2, E_3\}$

$\Gamma_2 = \{R_1, R_2, R_3, R_4, R_5, R_6\}$

5. Même question si l'on coupe à 0.3.

On obtient une partition en 5 classes, une classe pour les sites d'étang mais 4 petites classes pour les sites de rivière.
$\Gamma_1 = \{E_1, E_2, E_3\}$ $\Gamma_2 = \{R_1, R_6\}$ $\Gamma_3 = \{R_2, R_5\}$
$\Gamma_4 = \{R_3\}$ $\Gamma_5 = \{R_4\}$

6. Quelle partition correspond au saut maximal ? Est-ce effectivement la partition qui vous semble la plus appropriée ?

Le "saut maximal" (on ne regarde que la plus "courte" des branches de la fourche ⌐¬3←) dans ce dendrogramme est le dernier. C'est donc la partition de la question 4 en $\Gamma_1 = \{E_1, E_2, E_3\}$ et $\Gamma_2 = \{R_1, R_2, R_3, R_4, R_5, R_6\}$ qui correspond au saut maximal.
C'est la plus appropriée car elle distingue effectivement étangs et rivières

7. Imaginer une situation concrête où une telle classification d'espèces micro-animales aurait pu être utilisée en pratique et la décrire.

Dans la mise en place d'une expérience pour étudier l'influence d'une facteur extérieur sur un milieu aquatique (température, modification du courant, pollution, ...) il convient de répéter l'expérience un grand nombre de fois et il faut le faire dans des sites aussi "semblables" que possible : on prendra des sites appartenant à une même classe. et pour cela il est indispensable de connaitre la liste des sites d'une même classe.