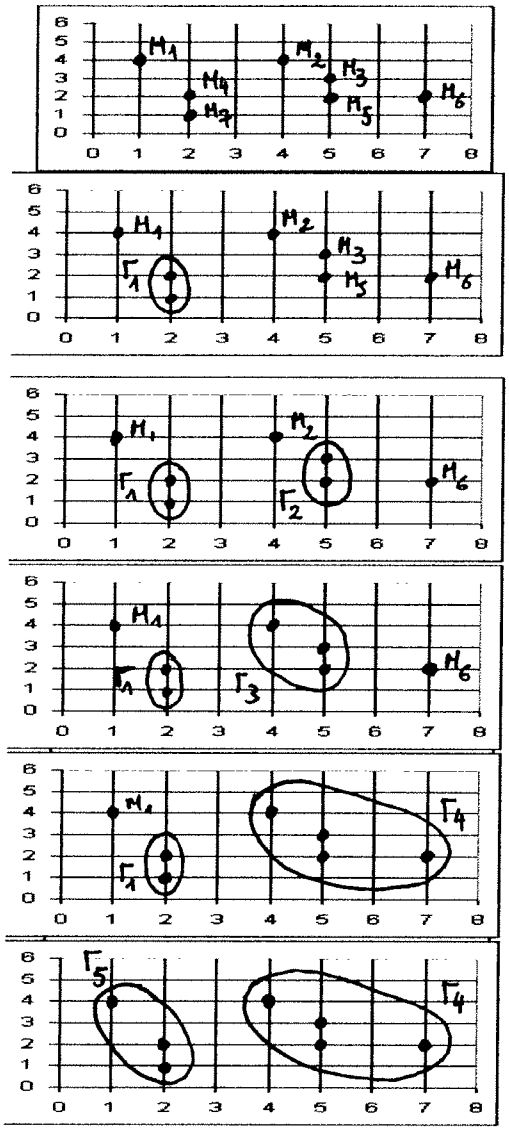
NOM : CORRIGÉ Date :
PRENOM : Groupe :

Mathématiques pour la Biologie (2010/2011, semestre 2) : Feuille-réponses du TD 8
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 : $M_1 = (1\,;4)$, $M_2 = (4\,;4)$, $M_3 = (5\,;3)$, $M_4 = (2\,;2)$, $M_5 = (5\,;2)$, $M_6 = (7\,;2)$ et $M_7 = (2\,;1)$.



	M1	M2	M3	M4	M5	M6	M7
M1	0	9	17	5	20	40	10
M2		0	2	8	5	13	13
M3			0	10	1	5	13
M4				0	9	25	1
M5					0	4	10
M6						0	26
M7							0

deux Choix (M_4, M_7) ou (M_3, M_5) On prend $\Gamma_1 = \{M_4, M_7\}$

	M1	M2	M3	M5	M6	Γ1
M1	0	9	17	20	40	5
M2		0	2	5	13	8
M3			0	1	5	10
M5				0	4	9
M6					0	25
Γ1						0

On regroupe M_3 et M_5 → $\Gamma_2 = \{M_3, M_5\}$

	M1	M2	M6	Γ1	Γ2
M1	0	9	40	5	17
M2		0	13	8	2
M6			0	25	4
Γ1				0	9
Γ2					0

On regroupe M_2 et Γ_2 → $\Gamma_3 = \{M_2, M_3, M_5\}$

	M1	M6	Γ1	Γ3
M1	0	40	5	9
M6		0	25	4
Γ1			0	8
Γ3				0

On regroupe M_6 et Γ_3 → $\Gamma_4 = \{M_2, M_3, M_5, M_6\}$

	M1	Γ1	Γ4
M1	0	5	9
Γ1		0	8
Γ4			0

On regroupe M_1 et Γ_1 → $\Gamma_5 = \{M_1, M_4, M_7\}$

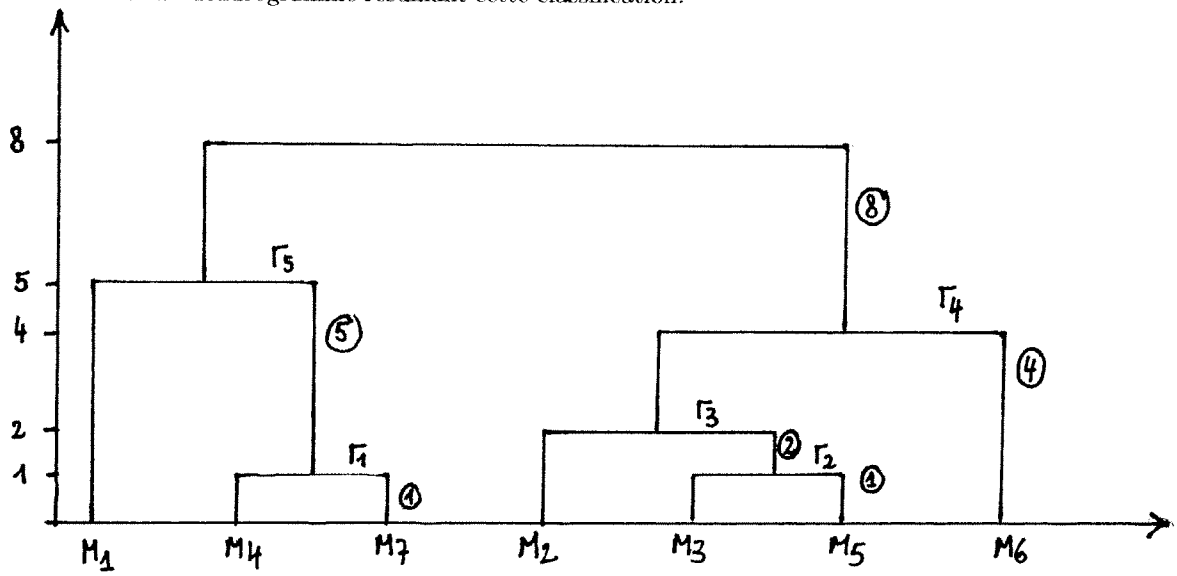
	Γ4	Γ5
Γ4	0	8
Γ5		0

1. Calculer le carré de la distance euclidienne de M_1 à M_4.

$$d_2^2(M_1, M_4) = (x_1 - x_4)^2 + (y_1 - y_4)^2 = (1-2)^2 + (4-2)^2 = (-1)^2 + (2)^2$$
$$= 1 + 4 = 5$$

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. On obtient la matrice des distances suivante :

1. Calculer $d(R_1, R_2)$ puis $d(R_6, R_3)$.

$$d(R_1,R_2) = \frac{(R_1R_1)+(R_2R_2)-2(R_1R_2)}{(R_1R_1)+(R_2R_2)} = \frac{11+15-2\times 7}{11+15} = \frac{12}{26} \simeq 0{,}462$$

$$d(R_6,R_3) = \frac{(R_{66}R_{66})+(R_3R_3)-2(R_6R_3)}{(R_{66}R_{66})+(R_3R_3)} = \frac{10+13-2\times 4}{10+13} = \frac{15}{23} \simeq 0{,}652$$

2. Compléter la colonne manquante de la matrice des distances suivante.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	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

3. Que pensez vous du choix de la distance. Pourquoi n'avoir pas choisi une distance euclidienne ?

La distance euclidienne est calculée à partir de coordonnées qu'on n'a pas et qui n'ont pas d'intérêt pour le problème traité.

La distance choisie est mieux adaptée :
si les sites A et B ont les mêmes espèces on a $n_A = n_B = n_{AB}$ donc $d(A,B) = 0$,
si les sites A et B n'ont pas d'espèce commune on a $n_{AB} = 0$ donc $d(A,B) = 1$ ce qui est le maximum

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

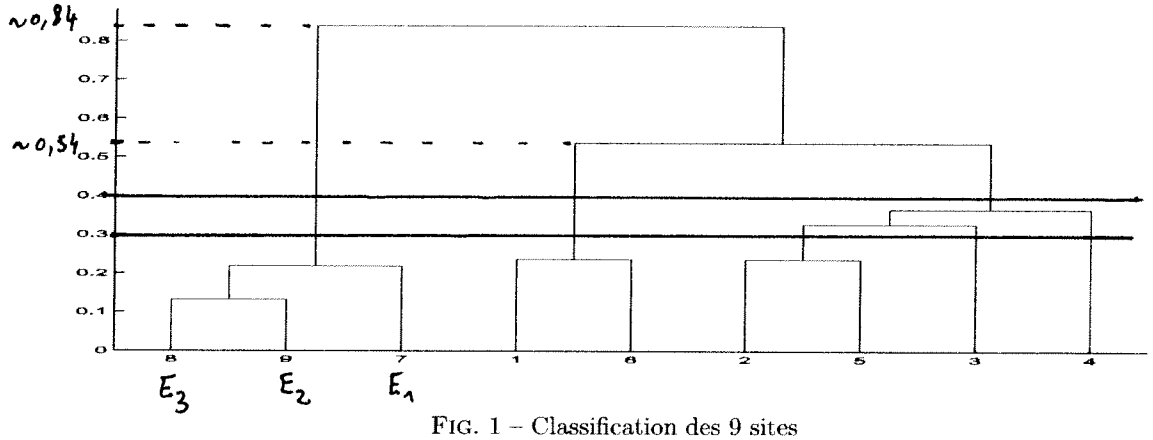


FIG. 1 – Classification des 9 sites

Décrire la composition des classes de la partition obtenue en coupant ce dendrogramme à la hauteur 0.3.

On obtient les classes

$\Gamma_1 = \{7,8,9\} = \{E_1, E_2, E_3\}$ $\Gamma_2 = \{1,6\}$ $\Gamma_3 = \{2,5\}$

$\Gamma_4 = \{3\}$ et $\Gamma_4 = \{4\}$

5. Même question si l'on coupe à 0.4.

On obtient les classes

$\Gamma_1 = \{7,8,9\}$ $\Gamma_2 = \{1,6\}$ et $\Gamma_3 = \{2,3,4,5\}$

6. Décrire la composition des classes de la partition qui vous semble la plus appropriée. Expliquer pourquoi celle-ci plutot qu'une autre.

Le saut maximal en hauteur, de 0,54 à 0,84, est fait lors du dernier regroupement. On peut donc couper à ce niveau. On distingue alors les deux groupes $\{7,8,9\}$ et $\{1,2,3,4,5,6\}$.