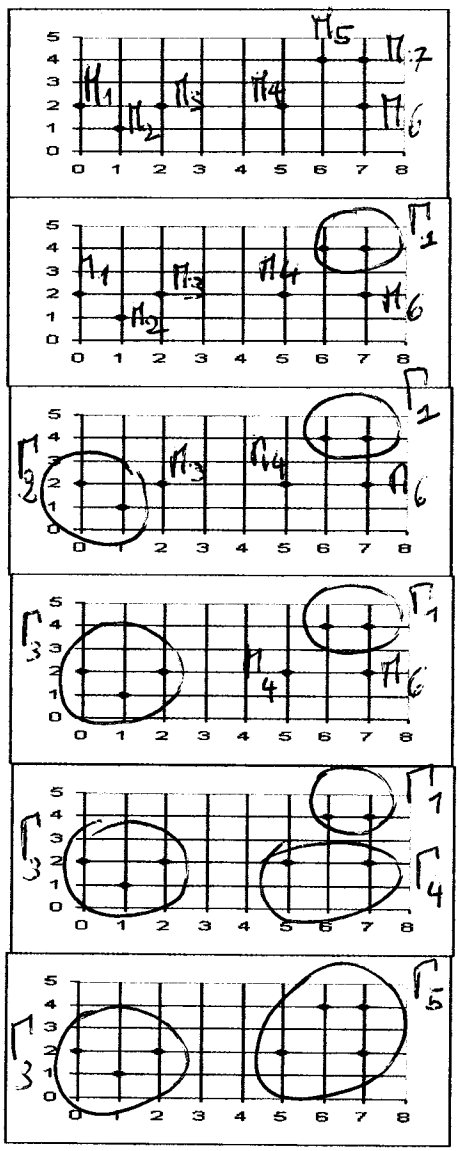
NOM : Date :
PRENOM : Groupe :

Mathématiques pour la Biologie (2009/2010, semestre 2) : Feuille-réponses du TD 7
Classification hierarchique ascendante

Exercice 1. : On se propose de réaliser une classification des 7 points représentés à gauche en utilisant la méthode d'aglomération au plus proche voisin.



	M1	M2	M3	M4	M5	M6	M7
M1	0	2	4	2,5	40	49	53
M2		0	2	1,7	34	37	45
M3			0	5	20	25	29
M4				0	5	4	8
M5					0	5	1
M6						0	4
M7							0

plus courte distance $(\Pi_5, \Pi_7) \to \Gamma_1$

	M1	M2	M3	M4	M6	Γ_1
M1	0	2	4	2,5	49	40
M2		0	2	1,7	37	34
M3			0	5	25	20
M4				0	4	5
M6					0	4
Γ_1						0

(Première) plus courte distance $(\Pi_1, \Pi_2) \to \Gamma_2$

	Γ_2	M3	M4	M6	Γ_1
Γ_2	0	2	1,7	37	34
M3		0	5	25	20
M4			0	4	5
M6				0	4
Γ_1					0

plus courte distance $(\Gamma_2, \Pi_3) \to \Gamma_3$

	Γ_3	M4	M6	Γ_1
Γ_3	0	5	25	20
M4		0	4	5
M6			0	4
Γ_1				0

(Première) plus courte distance $(\Pi_4, \Pi_6) \to \Gamma_4$

	Γ_3	Γ_4	Γ_1
Γ_3	0	5	20
Γ_4		0	4
Γ_1			0

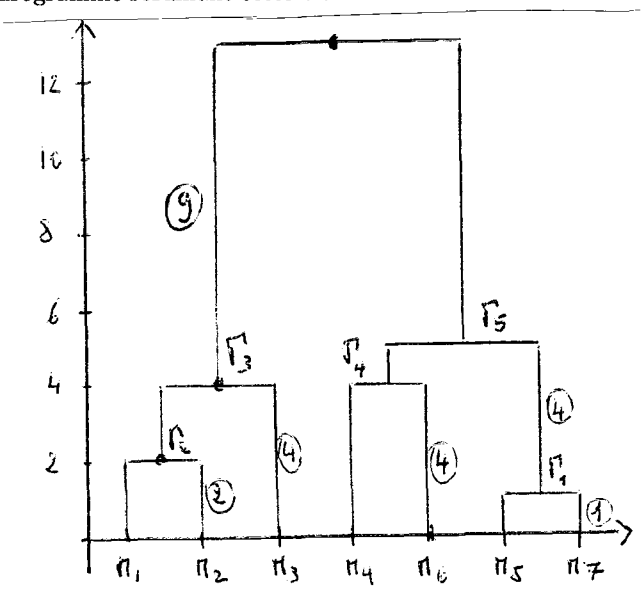
plus courte distance $(\Gamma_4, \Gamma_1) \to \Gamma_5$

	Γ_3	Γ_5
Γ_3	0	5
Γ_5		0

1. Compléter le premier tableau à droite représentant la matrice des distance des points tracés à gauche, en utilisant le carré de la distance euclidienne. Quels sont les deux points les plus proches, les deux points les plus éloignés ?

Dans la première matrice de distance on voit que Π_5 et Π_7 sont les deux points les plus proches et que Π_1 et Π_7 sont les plus éloignés.

2. 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.
3. Poursuivre la classification en complétant les tableaux suivants et en cerclant les classes, $\Gamma_2, \ldots$ créées au fur et à mesure.
4. 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 :

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	0	0,462	0,666	0,538	0,478	0,334	0,666	0,692	0,74
$R2$	0,462	0	0,428	O,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	0,572	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	O,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

1. Compléter les coefficients manquants de cette matrice en explicitant vos calculs.

$d(R3,R5) = \frac{13+12-2(7)}{13+12} = \frac{11}{25} = 0,44 = d(R5,R3)$ ← symétrie

$d(R4,R4) = d(R5,R5) = 0$

$d(R4,R2) = d(R2,R4) = 0,466$ $\quad d(R4,R6) = d(R6,R4) = 0,52$

$d(R4,R5) = \frac{15+12-2(7)}{15+12} = \frac{13}{27} = 0,481 = d(R5,R4)$ ← symétrie

2. Préciser quels sont les deux sites les plus proches ainsi que les deux sites les plus éloignés.

Les deux sites les plus proches sont les sites E2 et E3 car $d(E2,E3) = 0,185$. Les deux sites les plus éloignés sont E3 et R2 car $d(E3,R2) = 0,852$.

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

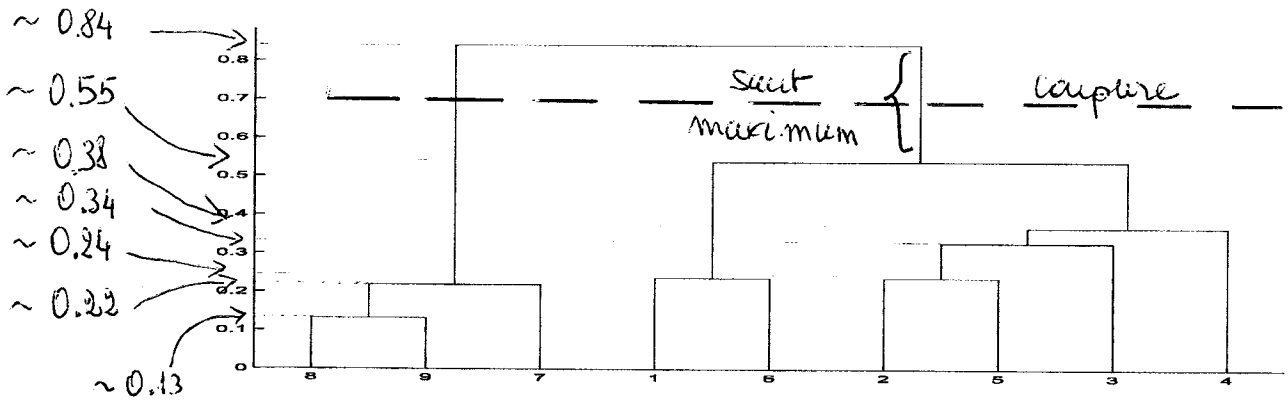


FIG. 1 – Classification des 9 sites

4. Indiquer quelle sont approximativement les ordonnées des regroupements sur le dendrogramme. Pensez-vous qu'elles sont proportionnelles à la distance au plus proche voisin ?

Le premier regroupement R8 et R9 (ou plutôt E2 et E3) est placé à une ordonnée 0.13 (environ), le second regroupement entre le précédent et E1 est placé à 0.22 (environ)
Les deux regroupements suivants entre R1 et R6 et entre R2 et R5 sont placés à 0.24.
Il y a ensuite 4 regroupements à 0.34, 0.38, 0.55 et 0.84.
Les ordonnées ne sont pas proportionnelles à la distance des plus proche voisin car on a utilisé ici une autre distance.

5. Décrire la composition des classes de la partition qui vous semble la plus appropriée. Comment avez-vous fait votre choix ?

Entre les différents regroupements, le plus important (de 0.55 à 0.84) est le dernier.
Il est naturel de couper le dendrogramme juste avant le dernier regroupement : on fait apparaître alors 2 classes : R1, R2, R3, R4, R5 et R6 d'une part et E1, E2 et E3 d'autre part (notés 8, 9 et 10). On retrouve les 2 classes "naturelles" de sites.

6. Pensez-vous qu'un autre choix de distance entre les sites aurait pu conduire à une partition différente ?

Si l'on change de distance, les points les plus proches ne sont plus les mêmes et la classification hiérarchique ascendante peut être donc différente.

7. Pourquoi n'a-t-on pas choisi la distance euclidienne comme distance entre les sites ?

Pour calculer la distance euclidienne entre deux points, il faut connaître les coordonnées cartésiennes de ces points (par exemple $M\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $N\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ $d(M,N) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$)
Ici, on n'a pas de coordonnées pour les sites. On doit donc choisir une autre distance.