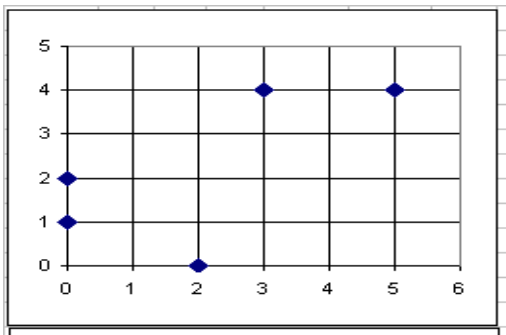
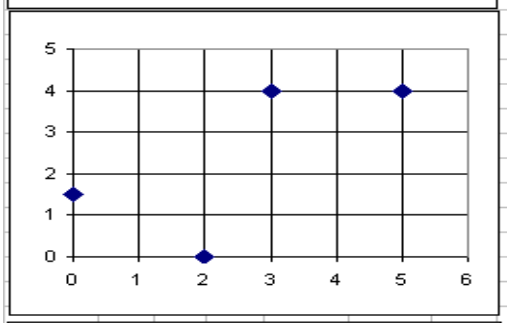
NOM : Date : .
PRENOM : Groupe : .

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

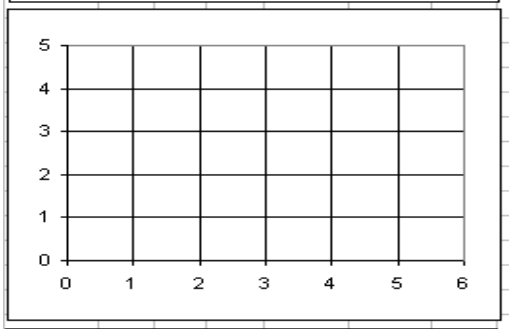
Exercice 1. : La succession des quatres dessins suivants correspond aux étapes successives d'une classification hierarchique ascendantes des cinq points $M_1(2,0)$, $M_2(0,1)$, $M_3(0,2)$, $M_4(3,4)$ et $M_5(5,4)$ progressivement regroupées en classes de deux ou trois points dont les centres de gravité sont notés G_6, G_7 et G_8. On suppose que les cinq points initiaux sont tous affectés du poids 1. La distance choisie pour cette classification, qui apparait dans les quatres matrices de distance, est l'écart de Ward.



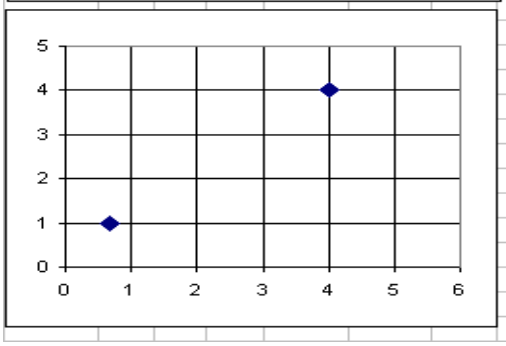
	M1	M2	M3	M4	M5
M1	0,00	2,50	4,00	8,50	12,50
M2	2,50	0,00	0,50	9,00	17,00
M3	4,00	0,50	0,00	6,50	14,50
M4	8,50	9,00	6,50		
M5	12,50			2,00	0,00



	M1	G6	M4	M5
M1	0,00	4,17	8,50	12,50
G6	4,17	0,00		20,83
M4	8,50		0,00	2,00
M5	12,50	20,83	2,00	0,00



	M1	G6	G7
M1	0,00	4,17	13,33
G6	4,17	0,00	22,25
G7	13,33	22,25	0,00



	G8	G7
G8	0,00	24,13
G7	24,13	0,00

1. Compléter le troisième dessin en y plaçant les trois points devant y figurer et indiquer sur les quatres dessins le nom des points.
2. Compléter les six distances manquantes dans les matrices de distances.

3. Préciser les coordonnées des points G_6, G_7 et G_8

4. Calculer les coordonnées du centre de gravité G_9 des cinq points.

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 :

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	0	$0,462$	$0,666$	$0,538$	$0,478$	$0,334$	$0,666$	$0,692$	$0,74$
$R2$	$0,462$	0	$0,428$	.....	$0,334$	$0,52$	$0,786$	$0,8$	$0,852$
$R3$	$0,666$	$0,428$	.....	.....	$0,44$	$0,652$	$0,846$	$0,786$	$0,84$
$R4$	$0,538$	$0,466$	.....	0	$0,482$	$0,52$	$0,572$	$0,466$	$0,556$
$R5$	$0,478$	$0,334$	$0,44$	$0,482$	0	$0,636$	$0,76$	$0,63$	$0,666$
$R6$	$0,334$	$0,52$	$0,652$	$0,52$	$0,636$	0	.....	.....	$0,546$
$E1$	$0,666$	$0,786$	$0,846$	$0,572$	$0,76$	$0,478$	.....	.....	$0,28$
$E2$	$0,692$	$0,8$	$0,786$	$0,466$	$0,63$	$0,6$	.....	.....	$0,186$
$E3$	$0,74$	$0,852$	$0,84$	$0,556$	$0,666$	$0,546$	$0,28$	$0,186$	0

1. Compléter les coefficients manquants de cette matrice.
2. Préciser quels sont les deux sites les plus proches ainsi que les deux sites les plus éloignés.

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

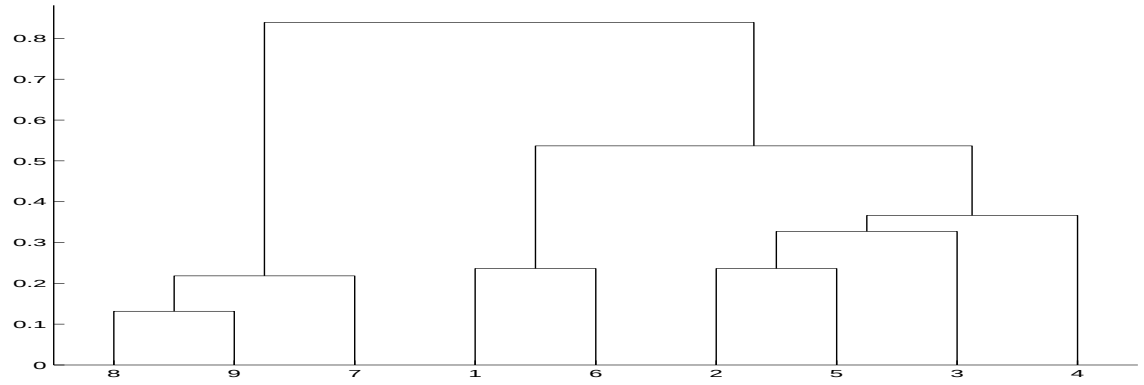


FIG. 1 – Classification des 9 sites

Décrire la composition des classes de la partition qui vous semble la plus appropriée.

4. Un autre choix de distance entre les sites aurait-il pu conduire à une partition différente ?

Pourquoi n'a-t-on pas choisi la distance euclidienne ?