

École Polytechnique de Montréal
Département de mathématiques et de génie industriel et département C.G.M.

MTH2302C – Probabilités et statistique

Examen final – Hiver 2009, Jeudi 30 avril 2009 de 13h30 à 16h00

QUESTION 1 (10 points)

On relève la précipitation annuelle maximale (en mm), sur une heure et en un point donné, pour les 100 dernières années. On soupçonne que cette distribution peut être représentée par une loi lognormale. Le tableau suivant donne les fréquences observées et les fréquences théoriques obtenues d'une loi lognormale de moyenne et de variance égales aux moyenne et variance des données ($\bar{x} = 20$ mm, $s^2 = 135$ mm²).

Intervalle i	[0,10)	[10,20)	[20,25)	[25,30)	[30,∞)	Total
Observée (O _i)	20	35	18	8	19	100
Théorique (E _i)	15	45	15	9	16	100

L'hypothèse d'une distribution lognormale semble-t-elle plausible (au niveau $\alpha = 5\%$) ? Faites le test requis.

Réponse

QUESTION 2 (10 points)

Pour améliorer la résistance en compression d'un béton, on envisage incorporer des cendres volantes dans le ciment. Le tableau suivant montre la distribution en fréquence des résistances d'éprouvettes de béton (en MPa) en fonction de la quantité de cendres volantes incorporées au ciment (en %).

	Résistance <30 MPa	≥ 30 MPa	Total
[0%, 1%)	10	15	25
≥ 1%	5	20	25
Total	15	35	50

Selon les résultats obtenus, peut-on affirmer, au niveau $\alpha = 5\%$, que la résistance du béton dépend de la quantité de cendres volantes présentes dans le ciment ? Faites le test requis..

Réponse

QUESTION 3 (10 points)

On veut étudier l'efficacité d'une méthode de traitement des sols contaminés à l'ammoniac (X). Initialement, on a prélevé un échantillon de taille 22 et l'on a obtenu $\bar{x}_1 = 40$ mg/l et $s_1^2 = 350$ (mg/l)². Après un mois d'application du traitement, on a prélevé un second échantillon de taille 10 et l'on a obtenu $\bar{x}_2 = 27$ mg/l et $s_2^2 = 373$ (mg/l)².

- Peut-on conclure que le traitement a permis d'abaisser la concentration du contaminant dans le sol de façon significative (niveau $\alpha = 5\%$) ? On supposera que les variances des deux populations sont égales, mais de valeur inconnue.
- Supposons que l'on considère cette fois la variance connue et $\sigma^2 = 350$ (mg/l)². Quel est alors le risque d'erreur de 2^e espèce (β) si l'on suppose que $\mu_1 - \mu_2 = 10$ mg/l et $\alpha = 5\%$ (Faites le calcul explicite, sans recourir aux graphiques de l'annexe de HM).

Réponses

QUESTION 4 (15 points)

On veut établir une équation de prédiction du logarithme de la conductivité hydraulique ($y = \log(k)$, k en m/s) en fonction du logarithme de la taille du tamis laissant passer 10% du matériau ($x = \log(d_{10})$, d_{10} en mm). On a prélevé sur un site 30 échantillons dans des tranchées et l'on a calculé les quantités suivantes :

$$\bar{x} = -4.64 \quad \bar{y} = -2.31 \quad \sum_{i=1}^{30} (y_i - \bar{y})^2 = 3.30 \quad \sum_{i=1}^{30} (x_i - \bar{x})^2 = 0.78 \quad \sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y}) = 1.31$$

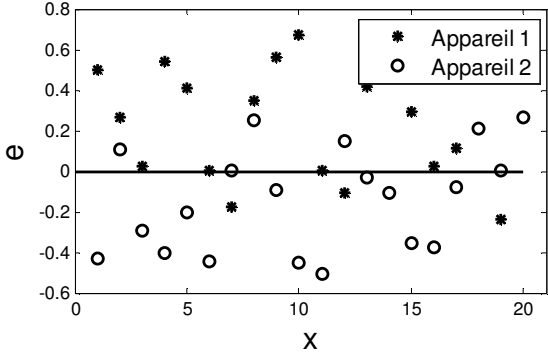
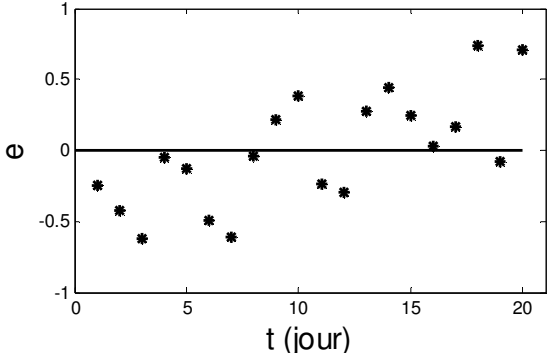
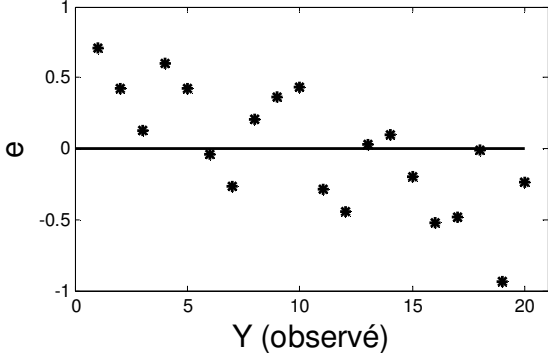
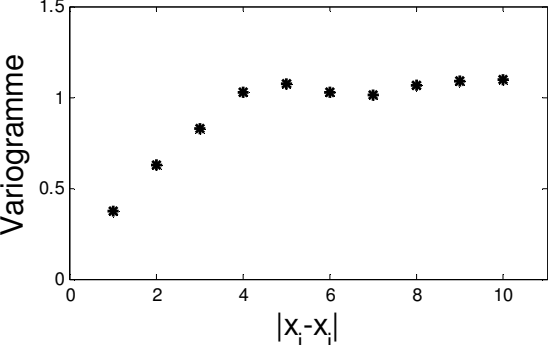
Avec le modèle linéaire de régression : $Y = \beta_0 + \beta_1 x + \varepsilon$, on trouve un R^2 de 0.66.

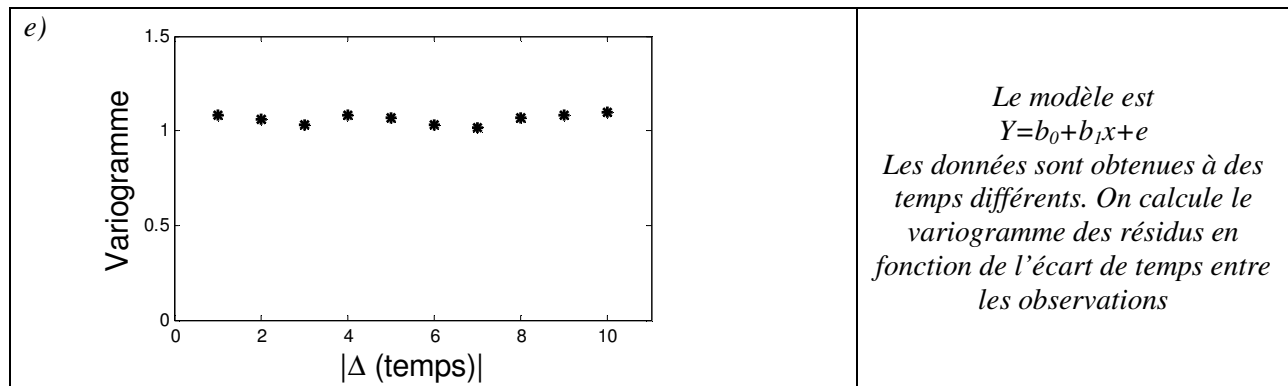
- Estimer les coefficients β_0 et β_1 .
- Calculer la somme du carré des résidus (SS_e).
- Déterminer si le coefficient β_1 est significativement différent de 0 (niveau $\alpha = 5\%$).
- Si l'on ajoute au modèle les variables $x_2 = \log(\text{porosité})$ et $x_3 = \log(\text{saturation})$, on obtient cette fois un $R^2 = 0.85$. Vérifier par le test approprié si l'augmentation du R^2 est significative ($\alpha = 5\%$).
- Vous échantillonnez un 2^e site en 20 emplacements. Indiquez comment on peut utiliser le test d'ajout pour vérifier si les 2 sites ont les mêmes coefficients dans l'équation de régression : $Y = \beta_0 + \beta_1 x + \varepsilon$.

Réponses

QUESTION 5 (15 points)

On vous présente différents graphiques de résidus de régressions.

Graphe des résidus	Remarques
<p>a)</p> 	<p><i>Les données ont été recueillies en utilisant deux appareils différents</i></p>
<p>b)</p> 	<p><i>Le modèle est $Y=b_0+b_1x+e$ Les données ont été prélevées à différents jours</i></p>
<p>c)</p> 	<p><i>Le modèle est $Y=b_0+b_1x+e$</i></p>
<p>d)</p> 	<p><i>Le variogramme des résidus est calculé en fonction de la « distance » selon la variable « x »</i></p>



Pour chaque graphe, indiquez si celui-ci semble correspondre à ce qui est attendu. Si vous identifiez un problème, décrivez ce qui ne va pas et suggérez une modification au modèle susceptible de résoudre le problème.

Réponses

QUESTION 6 (15 points)

On veut relier le débit annuel maximal de 30 rivières (y en m^3/s) à la surface de son bassin hydrographique (x_1 en km^2), la pente moyenne sur le bassin (x_2 sans unité), et la précipitation totale des 48h précédentes (x_3 en mm). On a ajusté le modèle de régression et obtenu :

$$Y=Xb, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 2.1 \\ 33 \\ 4.3 \end{bmatrix}$$

La somme des carrés des erreurs vaut : $2975 (m^3/s)^2$

On a aussi calculé la matrice $(X'X)^{-1}$ et on a obtenu :

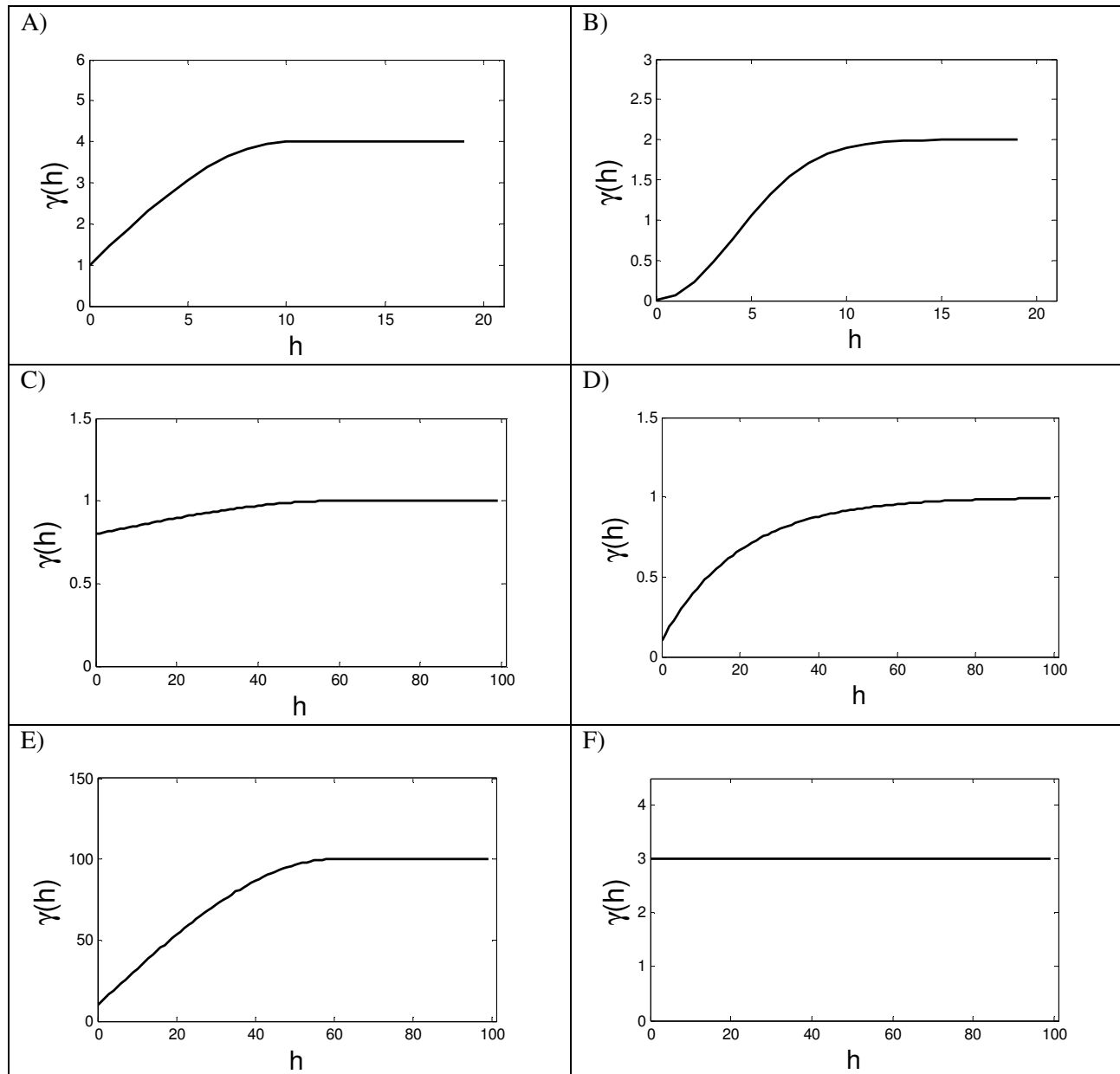
$$(X'X)^{-1} = \begin{pmatrix} 1.49 & -0.02 & -0.05 & -0.06 \\ -0.02 & 0.0008 & 0.005 & 0.0004 \\ -0.05 & 0.005 & 0.053 & -0.0004 \\ -0.06 & 0.0004 & -0.0004 & 0.003 \end{pmatrix}$$

- a) Quelle est la variance du coefficient b_3 ?
- b) Quelle est la corrélation entre les coefficients b_1 et b_2 ?
- c) On a une rivière montrant $x_1=30 km^2$ et $x_2=0.03$. Quelle est la valeur prédite du débit maximal pour cette rivière pour une précipitation de 20 mm sur les dernières 48h ?
- d) Donner l'intervalle de confiance à 95% pour le débit calculé en c) sachant que :
 $x_0(X'X)^{-1}x_0' = 0.296$ avec $x_0=[1 \ 30 \ 0.03 \ 20]$

Réponses

QUESTION 7 (10 points)

Soit la figure suivante illustrant différents modèles de variogramme; h est en m .



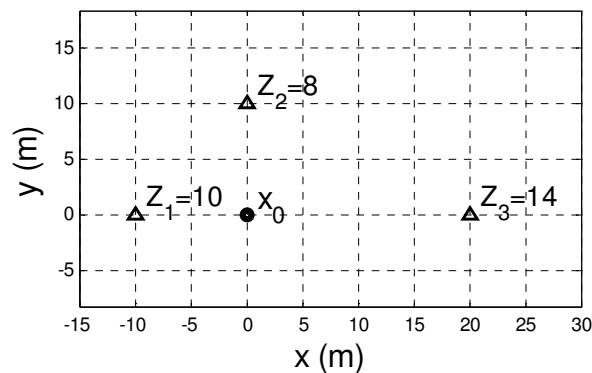
Compléter le tableau en associant une seule figure (lettres A à F) pour chaque énoncé apparaissant dans le tableau. Une même figure peut être associée à plus d'un énoncé de la liste; une figure peut aussi n'être associée à aucun énoncé de la liste).

Réponses (répondez directement dans le tableau)

Énoncé	Figure
1- Utilisé lors du krigeage, ce modèle va retourner une valeur constante partout sauf aux points où l'on a une donnée	
2- Ce modèle va fournir une estimation par krigeage continue aux points où l'on a des données	
3- Deux variables aléatoires espacées de moins de 10 m sont corrélées, alors que si elles sont espacées de plus de 10,m leur corrélation est nulle	
4- Les erreurs d'analyse et de localisation représentent 90% de la variation	
5- La variance est égale à 100	
6- Le modèle présente la plus petite portée	
7- Le modèle est exponentiel	
8- Le modèle est gaussien	
9- Le modèle présente une portée asymptotique et un effet de pépite	
10- Le modèle indique que deux variables aléatoires dont l'espacement est presque zéro devraient montrer une demi-variance de leurs différences égale à 1	

QUESTION 8 (15 points)

La figure suivante donne la localisation de 3 points où l'on a observé l'épaisseur de mort-terrain (épaisseur mesurée en m). On veut estimer par krigeage ordinaire l'épaisseur de mort-terrain au point x_0 . Le modèle de variogramme est sphérique, isotrope, ses paramètres sont : $C_0=1$, $C=2$ et $a=20$ m.



- a) Quelles sont les unités de C_0 et C ?
- b) Construisez, sous forme matricielle, le système de krigeage ordinaire. Ne pas résoudre.

c) On obtient comme vecteur solution :
$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \mu \end{bmatrix} = \begin{bmatrix} 0.39265 \\ 0.39265 \\ 0.21471 \\ -0.64412 \end{bmatrix} .$$
 Quelle est l'épaisseur estimée et la variance de krigeage au point x_0 ?

d) Calculer la variance de $Z_1 + 2Z_2$

Réponses

Corrigé

Barème

1- H_0 la loi est lognormale, vs H_1 elle n'est pas lognormale 10

$$\chi_0^2 = (20-15)^2/15 + (45-35)^2/35 + \dots = 5.1625 \text{ comparé à } \chi_{5-1-2,0.05}^2 = 5.99$$

On accepte H_0 que la loi lognormale décrit bien les fréquences observées

3pts pour le calcul, 4 pts pour les degrés de liberté, 3pts pour la bonne formulation et interprétation du test

2- H_0 : résistance et quantité de cendres volantes sont indépendantes 10

Les fréquences théoriques sous hypothèse d'indépendance sont données par $E_{ij} = N_i N_j / N$. On calcule :

$$E = \begin{matrix} 7.5 & 17.5 \\ 7.5 & 17.5 \end{matrix}$$

$$\chi_0^2 = (10-7.5)^2/7.5 + (15-17.5)^2/17.5 + \dots = 2.381$$

comparé à une $\chi_{1,0.05}^2 = 3.84$

Donc, on ne peut pas rejeter l'hypothèse d'indépendance.

3pts pour le calcul, 4 pts pour les degrés de liberté, 3pts pour la bonne formulation et interprétation du test

3- a) $H_0 \mu_1 = \mu_2$ vs $H_1 \mu_1 > \mu_2$ 5

On calcule $s_p^2 = (21 \cdot 350 + 9 \cdot 373) / 30 = 356.9$

$$t_0 = (\bar{x}_1 - \bar{x}_2) / (s_p \sqrt{1/n_1 + 1/n_2})^{0.5} = (40 - 27) / 7.205 = 1.80$$

comme t_0 est $> t_{30, 5\%} = 1.697$, on rejette H_0 , donc il y a eu diminution significative

Note : -1 si test bilatéral; -2 si test avec les variances inégales

b) Soit $H_0: \delta = 0$, $H_1: \delta = 10$. 5

Le seuil de rejet pour $Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{0.5}}$ est $Z_\alpha = 1.645$.

Pour $\bar{X}_1 - \bar{X}_2$ c'est donc : $1.645 \cdot (350 \cdot (1/22 + 1/10))^{0.5} = 11.737$

On cherche : $P(\bar{X}_1 - \bar{X}_2 < 11.737 \mid H_1 \text{ vraie})$

$$= P(Z < (11.737 - 10) / (350 \cdot (1/22 + 1/10))^{0.5}) = P(Z < 1.737 / 7.1351) = P(Z < 0.24) = 0.596$$

On peut aussi utiliser directement : $P(Z < Z_\alpha - \frac{\mu_1 - \mu_2}{\sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{0.5}}) = P(Z < 0.24) = 0.596$

Note : on ne peut pas utiliser les graphes car « n » n'est pas égal dans les deux groupes

4- a) $b_1=1.31/0.78=1.6795$ $b_0=-2.31-1.6795*(-4.64)=5.4829$ 3

b) $R^2=SCR_m/SCT_m \Rightarrow SCR_m=0.66*3.30=2.178$ et $SCE= SCT_m-SCR_m=3.3-2.178=1.122$ 3

c) $H_0 : \beta_1 = 0$, vs $H_1 \beta_1 \neq 0$. 3

Le test est $F_0 = (SS_r/1) / (SS_e/28) = (3.3-1.12)/(1.12/28) = 54.5 > F_{1,28,5\%} = 4.196$, On rejette H_0 , le coefficient $\beta_1 \neq 0$

On peut aussi calculer l'intervalle de confiance pour β_1 et vérifier que celui-ci n'inclut pas le 0.

d) Le SS_r complet vaut $0.85*3.30=2.805$ ($SS_e=3.3-2.805=0.495$) comparé à $0.66*3.30=2.178$ pour le modèle réduit. 3

On calcule $F_0 = [(2.805-2.178)/2] / [0.495/(30-3-1)] = 16.47 > F_{2,26,.05} = 3.369$ Donc oui l'ajout est significatif.

e) 3

On compare le modèle $Y = \beta_0 + \beta_1 x + \varepsilon$ au modèle $Y = \beta_0 + \beta_1 x + \beta_2 I + \beta_3 Ix + \varepsilon$. où I est une variable indicatrice du site prenant la valeur 0 pour un des deux sites et 1 pour l'autre site. Ix est le produit $I*x$. On teste $H_0 = \beta_2 = \beta_3 = 0$. Si l'on ne peut pas rejeter H_0 , on est obligé de conclure que les deux sites ont la même régression.

note : -1 si la suggestion ne porte que sur l'indicatrice I et n'inclut pas aussi Ix

5) 3 pts par graphe

a) Les résidus de l'appareil 1 ont une moyenne supérieure à 0, ceux de l'appareil 2 ont une moyenne inférieure à 0. Il faudrait permettre une différence de niveau pour les deux appareils par l'inclusion d'une variable indicatrice donnant 0 pour l'un des appareils et 1 pour l'autre.

b) Une tendance dans le temps est présente. On devrait chercher à inclure cette tendance dans le modèle.

c) Deux réponses possibles : i. on s'attendrait plutôt à une pente positive entre résidus et valeur observée ou ii. on ne devrait pas s'intéresser à ce graphe mais plutôt au graphe e vs $y_{prédit}$.

d) Le variogramme montre une corrélation des résidus en fonction de « x ». Ceci n'est pas souhaitable. Il faut revoir le modèle (e.g. nouvelles variables, transformations de « x ») pour chercher à obtenir un variogramme « effet de pépite pur ».

e) C'est ce que l'on souhaite obtenir pour le variogramme des résidus.

Notes : pour d) et e), très peu ont vu que le variogramme était calculé sur les résidus et que l'on doit s'attendre à un effet de pépite pur puisque les résidus doivent être non-corrélés.

6- a) On calcule $CME=2975 / 26 = 114.42 \Rightarrow V(b_3)=114.42*0.003= 0.343$ 4

b) $r(b_1,b_2)= 0.005 / (0.053*0.0008)^{0.5} = 0.768$ 3

c) $y_{prédit} = 20+2.1*30+33*0.03+4.3*20 = 169.9 \text{ m}^3/\text{s}$ 4

d) $169.9 \pm t_{26,0.025} s_e (1+x_0(X'X)^{-1}x_0')^{0.5} = 169.9 \pm 2.056 (114.42)^{0.5} (1+0.296)^{0.5} = 169.9 \pm 25.03 = [144.9, 194.9]$ 4

Note : Accepter aussi la réponse avec l'intervalle pour la moyenne. $169.9 \pm t_{26,0.025} s_e (x_0 (X' X)^{-1} x_0')^{0.5}$

7- 1 pt par bonne réponse; si plus d'une figure pour un énoncé 0. Plusieurs ont mis A à la question 6. C'est bien F puisqu'un effet de pépité a une portée epsilon par la définition de portée.

Énoncé	Figure
1- Utilisé lors du krigeage, ce modèle va retourner une valeur constante partout sauf aux points où l'on a une donnée	F
2- Ce modèle va fournir une estimation par krigeage continue aux points où l'on a des données	B
3- Deux variables aléatoires espacées de moins de 10 m sont corrélées, alors que si elles sont espacées de plus de 10,m leur corrélation est nulle	A
4- Les erreurs d'analyse et de localisation et les variations de très petite échelle représentent 90% de la variation	C
5- La variance de la variable aléatoire est égale à 100	E
6- Le modèle présente la plus petite portée	F
7- Le modèle est exponentiel	D
8- Le modèle est gaussien	B
9- Le modèle présente une portée asymptotique et un effet de pépité	D
10- Le modèle indique que deux variables aléatoires dont l'espacement est presque zéro devraient montrer une demi-variance de leurs différences égale à 1	A

8- a) m^2 3

b) On doit calculer la covariance à $h=10$ et à $h=10*2^{0.5}$ on trouve : 4
 $C(10)=2*(1-(1.5*10/20-0.5(10/20)^3)) = 0.625$ et
 $C(10*2^{0.5})= 2*(1-(1.5*10*2^{0.5}/20-0.5(10*2^{0.5}/20)^3)=0.232$

$$\begin{bmatrix} 3 & 0.232 & 0 & 1 \\ 0.232 & 3 & 0 & 1 \\ 0 & 0 & 3 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \mu \end{bmatrix} = \begin{bmatrix} 0.625 \\ 0.625 \\ 0 \\ 1 \end{bmatrix}$$

c) $Z_0^* = 0.39265*10+0.39265*8+0.21471*14=10.074$ 4

variance de krigeage : $3-0.39265*0.625-0.39265*0.625-(-0.64412)=3.1533$

d) $V(Z_1+2Z_2)=V(Z_1)+4V(Z_2)+4Cov(Z_1,Z_2)$ 4

$\Rightarrow V(Z_1+2Z_2)= (3+4*3+4*0.232)=15.928$