Banque "Agro-Véto"
A - 0717

MATHÉMATIQUES

Modélisation mathématique et informatique

Durée : 3 heures 30 minutes

Chaque candidat est responsable de la vérification de son sujet d'épreuve : pagination et impression de chaque page. Ce contrôle doit être fait en début d'épreuve : en cas de doute, il doit alerter au plus tôt le chef de centre qui contrôlera et éventuellement remplacera le sujet.

Si, au cours de l'épreuve, un candidat repère ce qui lui semble être une erreur d'énoncé, il le signale sur sa copie et poursuit sa composition en expliquant les raisons des initiatives qu'il a été amené à prendre.

L'usage d'une calculatrice est autorisé pour cette épreuve. Les questions d'informatique devront être rédigées en langage Python exclusivement.

Le sujet est composé de quatre parties indépendantes.

La question 5. de la partie 1 et les questions 3 et 5.c) de la partie 4 sont des questions d'informatique. Elles peuvent être traitées indépendamment des questions mathématiques qui les précèdent.

Dans le domaine de l'écophysiologie végétale on s'intéresse aux différentes phases du développement des plantes afin de mieux contrôler les rendements de culture. Plus précisément, on s'intéresse à l'évolution du nombre de feuilles d'une plante en fonction du "temps thermique". Ce dernier correspond à la somme cumulée des températures au cours des différents jours de l'expérience. De telles mesures permettent de mettre en évidence trois phases de développement appelées phases de rosette, d'élongation et de floraison. La modélisation la plus souvent utilisée consiste à supposer que l'évolution du nombre de feuilles en fonction du temps thermique se comporte comme une fonction affine au cours de chacune des trois phases précédentes (cf. Figure 1).

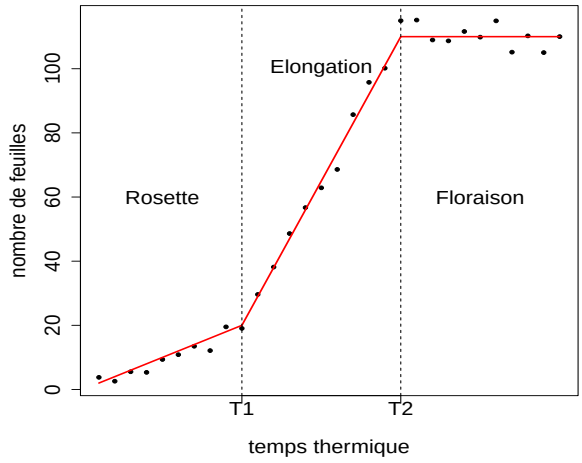

FIGURE 1 – Schéma illustrant l'évolution du nombre de feuilles d'une plante en fonction du temps thermique ; y sont représentés les valeurs expérimentales ('•') et les ajustements affines dans chacune des trois phases.

L'objectif est alors de mettre en place des méthodes automatiques permettant de détecter les instants séparant d'une part les phases de rosette et d'élongation (T1) et d'autre part les phases d'élongation et de floraison (T2). Il est, en effet, intéressant de voir si ces instants ont tendance à changer en fonction de certaines conditions expérimentales auxquelles les plantes pourraient être soumises.

PARTIE 1

Dans cette partie, on **se focalise sur l'une des trois phases précédentes** et on s'intéresse à l'ajustement affine que l'on obtient à partir du nuage de points $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, n étant un entier naturel strictement plus grand que 1, en utilisant le critère des moindres carrés où l'on supposera que les x_i sont distincts. Dans la notation (x_i, y_i), x_i correspond au temps thermique de la $i$ème observation et y_i correspond au nombre de feuilles de la $i$ème observation (cf. Figure 2).

1. On note $\overline{x} = (\sum_{i=1}^{n} x_i)/n$. Montrer que

$$\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

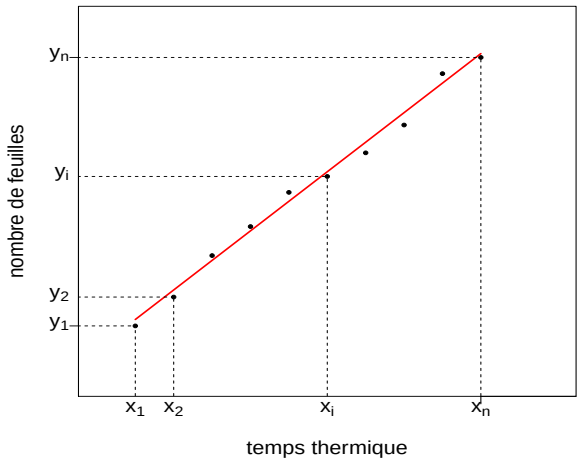

FIGURE 2 – Schéma illustrant l'évolution du nombre de feuilles en fonction du temps thermique durant l'une des trois phases.

2. On définit la fonction $F : \mathbb{R}^2 \longrightarrow \mathbb{R}$ par :

$$\forall (a,b) \in \mathbb{R}^2, \quad F(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

Montrer que le système suivant

$$\begin{cases} \frac{\partial F}{\partial a}(a,b) &= 0, \\ \frac{\partial F}{\partial b}(a,b) &= 0. \end{cases}$$

admet une unique solution $(\widehat{a}, \widehat{b}) \in \mathbb{R}^2$ où

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x},$$

avec $\overline{y} = (\sum_{i=1}^{n} y_i)/n$ et

$$\widehat{b} = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

3. On admet que le couple $(\widehat{a}, \widehat{b})$ obtenu réalise un minimum global. Interpréter le résultat obtenu.
4. À partir des valeurs de (x_i, y_i) données dans le tableau ci-dessous, calculer à l'aide de la calculatrice les valeurs correspondantes de $\widehat{a}$ et de $\widehat{b}$ pour cet exemple.

x_i	2	4	6	8	10	12	14	16	18	20
y_i	8	13	20	27	32	38	44	50	56	62

5. On souhaite automatiser les calculs de $\hat{a}$ et $\hat{b}$ lorsqu'un grand nombre de valeurs (x_i, y_i) sont à disposition. On utilise Python, en considérant que les valeurs $x_1, \ldots, x_n$ sont stockées dans une liste `x` et que les valeurs $y_1, \ldots, y_n$ sont stockées dans une liste `y`.

   Consigne. Chaque algorithme, à écrire en Python, doit être précédé d'une phrase expliquant le raisonnement suivi pour l'écrire.

   (a) Écrire une fonction `moy` qui prend en entrée `L`, une liste de nombres (non vide), et qui renvoie la moyenne des valeurs de `L`.

   (b) Soit `L` une liste de longueur $n \geq 1$. Indiquer, sans justification, la ou les réponse(s) correcte(s) :

      i. Les éléments de `L` sont numérotés de 0 à $n-1$ (inclus) ;

      ii. Les éléments de `L` sont numérotés de 1 à n (inclus) ;

      iii. Les éléments de `L` sont numérotés de 0 à $n+1$ (inclus) ;

      iv. Il y a n éléments dans `L`.

      v. Il y a $n+1$ éléments dans `L`

   (c) Écrire une fonction `Bchap` qui prend en entrée `x` et `y`, deux listes de nombres de même longueur $n \geq 2$, et qui renvoie la valeur de $\hat{b}$ associée.

   (d) Écrire une fonction `Achap` qui prend en entrée `x` et `y`, deux listes de nombres de même longueur $n \geq 2$, et qui renvoie la valeur de $\hat{a}$ associée.

PARTIE 2

Dans cette partie, pour tenir compte de la variabilité des expérimentations, nous allons modéliser les y_i comme des réalisations de variables aléatoires Y_i définies par les n équations suivantes, n étant un entier naturel strictement plus grand que 1 :

$$Y_i = a + bx_i + \varepsilon_i \ , \ 1 \leq i \leq n,$$

où a et b sont des paramètres réels inconnus, x_i est le temps thermique associé à la $i$ème observation (il sera considéré comme déterministe ici c'est-à-dire non aléatoire) et les ε_i sont des variables aléatoires indépendantes d'espérance nulle et de variance σ^2. Ceci correspond à un modèle de régression linéaire simple.
En utilisant la même démarche que celle proposée dans la partie 1, nous définissons :

$$\widehat{a} = \overline{Y} - \widehat{b}\overline{x},$$

où $\overline{Y} = (\sum_{i=1}^n Y_i)/n$, $\overline{x} = (\sum_{i=1}^n x_i)/n$ et

$$\widehat{b} = \frac{\sum_{i=1}^n Y_i(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Le but de cette partie est d'étudier les propriétés des variables aléatoires $\widehat{a}$ et $\widehat{b}$. Ces variables aléatoires sont appelées des estimateurs de a et b.

1. On note $\mathbb{E}(X)$ l'espérance d'une variable aléatoire X. Calculer $\mathbb{E}(Y_i)$ où $i \in \{1, \ldots, n\}$.
2. Calculer $\mathbb{E}\left(\sum_{i=1}^n Y_i(x_i - \overline{x})\right)$.

3. En déduire que $\widehat{b}$ est un estimateur sans biais de b où on dira qu'un estimateur est sans biais lorsque $\mathbb{E}(\widehat{b}-b)=0$. On pourra utiliser le résultat de la question 1 de la partie 1 ici admis :
$$\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (x_i-\overline{x})^2 .$$
4. Calculer $\mathbb{E}(\overline{Y})$. En déduire que $\widehat{a}$ est un estimateur sans biais de a.
5. Montrer que
$$\widehat{b}-b = \frac{\sum_{i=1}^{n} \varepsilon_i (x_i-\overline{x})}{\sum_{i=1}^{n} (x_i-\overline{x})^2}.$$
6. On note $\mathrm{Var}(X)$ la variance de la variable aléatoire X. Montrer que :
$$\mathrm{Var}(\widehat{b}) = \mathrm{Var}(\widehat{b}-b).$$
7. En déduire que :
$$\mathrm{Var}(\widehat{b}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$
8. Dans cette question, nous nous intéressons au comportement de l'estimateur $\widehat{b}$ en fonction du nombre d'observations n et nous le notons $\widehat{b}_n$. On dit que $\widehat{b}_n$ converge en probabilité vers b lorsque :
$$\forall \varepsilon > 0,\ \mathbb{P}\left(|\widehat{b}_n - b| \geq \varepsilon\right) \to 0, \quad \text{quand } n \text{ tend vers l'infini.}$$
   (a) Calculer $\sum_{i=1}^{n}(x_i-\overline{x})^2$ lorsque $x_i = i$, où $i \in \{1,\dots,n\}$.
   (b) Déduire en utilisant la question 7, que lorsque $x_i = i$, $\widehat{b}_n$ converge en probabilité vers b.
9. Soit $\widetilde{b}$ une variable aléatoire définie comme une combinaison linéaire des Y_i par $\widetilde{b} = \sum_{i=1}^{n} \mu_i Y_i$ où $\sum_{i=1}^{n} \mu_i = 0$ et $\sum_{i=1}^{n} \mu_i x_i = 1$.
   (a) Montrer que :
$$\mathrm{Var}(\widetilde{b}) = \mathrm{Var}(\widetilde{b}-\widehat{b}) + \mathrm{Var}(\widehat{b}).$$
   (b) En déduire que :
$$\mathrm{Var}(\widetilde{b}) \geq \mathrm{Var}(\widehat{b}).$$
10. Parmi les estimateurs sans biais et définis comme une combinaison linéaire des Y_i, $\widehat{b}$ est de variance minimale. Justifier.

PARTIE 3

On se place dans cette partie dans le même cadre que celui de la partie 2 mais on suppose de plus que les variables aléatoires ε_i sont indépendantes et identiquement distribuées de loi gaussienne d'espérance nulle et de variance σ^2 notée $\mathcal{N}(0,\sigma^2)$.

1. Donner la loi de probabilité de $\widehat{b}$ défini par
$$\widehat{b} = \frac{\sum_{i=1}^{n} Y_i (x_i-\overline{x})}{\sum_{i=1}^{n} (x_i-\overline{x})^2}$$
en utilisant les résultats des questions 3 et 7 de la partie 2 ici admis.

2. Dans la phase de floraison et **uniquement dans celle-ci**, une façon usuelle d'estimer σ^2 est de considérer l'estimateur suivant :

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

(a) Montrer que $\widehat{\sigma}^2$ peut se réécrire comme suit :

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon})^2,$$

où $\overline{\varepsilon} = (\sum_{i=1}^{n}\varepsilon_i)/n$, les ε_i étant des variables aléatoires gaussiennes centrées de variance σ^2.

Dans les questions suivantes on supposera que $n = 4$.

(b) Soient

$$\widetilde{E} = \begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\end{pmatrix},\ E = \begin{pmatrix}\varepsilon_1 - \overline{\varepsilon}\\ \varepsilon_2 - \overline{\varepsilon}\\ \varepsilon_3 - \overline{\varepsilon}\\ \varepsilon_4 - \overline{\varepsilon}\end{pmatrix},\ \text{et } \Gamma = \begin{pmatrix}(1-1/4) & -1/4 & -1/4 & -1/4\\ -1/4 & (1-1/4) & -1/4 & -1/4\\ -1/4 & -1/4 & (1-1/4) & -1/4\\ -1/4 & -1/4 & -1/4 & (1-1/4)\end{pmatrix}.$$

Montrer que :

$$E = \Gamma\widetilde{E}. \tag{1}$$

(c) Soit J la matrice de taille 4×4 ne contenant que des 1. Quel est le rang de J ?

(d) Montrer que le vecteur de taille 4 ne contenant que des 1 est vecteur propre de J. Quelle est la valeur propre associée ?

(e) En déduire les valeurs propres de J.

(f) En écrivant Γ sous la forme :

$$\Gamma = \mathrm{I}_4 - J/4,$$

où I_4 désigne la matrice identité de $\mathbb{R}^4$, trouver les valeurs propres de la matrice Γ.

(g) Montrer que

$$\Gamma = P\ D\ {}^tP,$$

où ${}^tP$ désigne la transposée de P, D est une matrice diagonale. Préciser la matrice D et les propriétés de P.

(h) Soit z un vecteur colonne de $\mathbb{R}^4$:

$$z = \begin{pmatrix}z_1\\ z_2\\ z_3\\ z_4\end{pmatrix}.$$

En observant que :

$$\sum_{i=1}^{4} z_i^2 = {}^tzz,$$

et en utilisant l'équation (1), montrer que :

$$\sum_{i=1}^{4}(\varepsilon_i - \bar{\varepsilon})^2 = \sum_{i=1}^{4}\lambda_i^2 V_i^2,$$

où V_i désigne la $i$ème composante du vecteur $V = {}^tP\widetilde{E}$ et les λ_i désignent les valeurs propres de Γ.

(i) On dit que la variable aléatoire Z suit une loi du $\chi^2(p)$ lorsque $Z = \sum_{i=1}^{p} W_i^2$ où les W_i sont des variables aléatoires indépendantes, de loi normale d'espérance nulle et de variance 1. En admettant que les V_i sont des variables aléatoires indépendantes, gaussiennes d'espérance nulle et de variance σ^2, déduire la loi de probabilité de $3\widehat{\sigma}^2/\sigma^2$.

3. Supposons maintenant que l'on ait à notre disposition les observations des deux premières phases : "rosette" et "élongation". En s'inspirant du critère des moindres carrés, proposer un critère pour estimer T1 correspondant au temps thermique séparant les phases de rosette et d'élongation, défini dans la Figure 1.

4. Supposons plus généralement que l'on ait à notre disposition l'ensemble des observations c'est-à-dire celles correspondant aux trois phases. Proposer un critère pour estimer T1 et T2 correspondant aux temps thermiques séparant les différentes phases de développement définis dans la Figure 1.

PARTIE 4

1. Soit u une fonction connue dans $\mathcal{C}^4(\mathbb{R})$ que l'on observe aux n temps thermiques : x_1, x_2,..., x_n que l'on supposera deux à deux distincts. On notera u_i la valeur de u en x_i :

$$u_i = u(x_i)\ ,\ 1 \leq i \leq n.$$

On supposera de plus que les données sont récoltées à des intervalles de temps réguliers $h > 0$ avec $h = x_{i+1} - x_i\ ,\ 1 \leq i \leq n-1$.

(a) Écrire un développement de Taylor-Young de u à l'ordre 4 en x_i de la forme

$$u(x) = u(x_i) + \cdots$$

(b) En appliquant le développement précédent à $x = x_{i+1}$ et $x = x_{i-1}$ déduire que pour $2 \leq i \leq n-1$,

$$u''(x_i) = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + e(h),$$

et expliciter le terme $e(h)$.

2. À partir de maintenant on va travailler sur un modèle simplifié où les valeurs expérimentales $y_i\ ,\ 1 \leq i \leq n$, sont portées par la fonction linéaire par morceaux de la Figure 1 et récoltées à des intervalles de temps réguliers h ($h = x_{i+1} - x_i\ ,\ 1 \leq i \leq n-1$). On aura donc :

$$\begin{cases} y_i = a_1 + b_1 x_i & \text{si } x_i \leq T_1, \\ y_i = a_2 + b_2 x_i & \text{si } T_1 \leq x_i \leq T_2, \\ y_i = a_3 + b_3 x_i & \text{si } T_2 \leq x_i. \end{cases}$$

Les valeurs des paramètres a_j et b_j, $1 \leq j \leq 3$ sont supposées inconnues.

On notera

$$D_i = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}, \quad \text{pour } 2 \leq i \leq n-1.$$

(a) Trouver la valeur de D_i lorsque $x_{i+1} \leq T_1$ en utilisant la question 1b de la partie 4 sans faire de calcul.

(b) Proposer une méthode pour détecter les deux instants séparant les différentes phases.

3. Cette question ne nécessite pas d'avoir obtenu les résultats qui précèdent pour y répondre.

(a) Que renvoie l'algorithme suivant, appliqué sur une liste `L` non vide :

```
def Mystere(L):
    """ ... (L: liste non vide)."""
    n = len(L)
    p = 0
    m = L[0]
    for k in range(1, n):
        if m < L[k]:
            m = L[k]
            p = k
    return(p)
```

(b) Si `L` est vide, indiquer *avec justification* la ou le(s) réponse(s) correcte(s) :

i. `Mystere(L)` renvoie une erreur.

ii. `Mystere(L)` renvoie le booléen `False`.

iii. `Mystere(L)` renvoie le nombre 0.

iv. `Mystere(L)` ne renvoie rien.

(c) Dans la fonction `Mystere`, on peut éviter l'utilisation de la variable `m` (sans changer le résultat renvoyé lorsque `L` est non vide) : proposer, en expliquant le raisonnement suivi, une version modifiée de `Mystere` en ce sens. Que renvoie la fonction modifiée lorsque la liste entrée est vide ?

(d) On dispose dorénavant d'une liste `L` de nombres, de longueur supérieure ou égale à deux, dont les éléments sont distincts deux à deux. On souhaite renvoyer les positions des deux plus grands éléments de `L`.

i. On suppose que l'on dispose d'un algorithme de tri. Proposer une démarche pour répondre à la question.

ii. On souhaite désormais répondre sans utiliser d'algorithme de tri mais en utilisant la fonction `Mystere`. On commence par calculer `p1=Mystere(L)` et ensuite, on cherche le plus grand élément `L[p2]` de `L` tel que `p2` est différent de `p1`. Compléter les trois lignes manquantes de la fonction suivante, afin qu'elle renvoie la liste `[p1,p2]` des positions des deux plus grands éléments de `L`.

```
def DeuxMax(L):
    """..."""  #à compléter
    n = len(L)
    p1 = Mystere(L)
    p2 = 0
    for k in range(1, n):
        ...   #à compléter
        ...   #à compléter
    return([p1, p2])
```

4. Dans la pratique les valeurs expérimentales sont bruitées et on a :

$$\begin{cases} y_i &= a_1 + b_1 x_i + \epsilon_i \quad \text{si } x_i \leq T_1, \\ y_i &= a_2 + b_2 x_i + \epsilon_i \quad \text{si } T_1 \leq x_i \leq T_2, \\ y_i &= a_3 + b_3 x_i + \epsilon_i \quad \text{si } T_2 \leq x_i. \end{cases}$$

où ϵ_i désigne un terme d'erreur ayant la forme particulière suivante $\epsilon_i = (-1)^i \epsilon$, où $\epsilon > 0$.

   (a) Calculer D_i dans le cas où $x_{i+1} \leq T_1$.

   (b) En supposant que ϵ est "suffisamment petit", proposer une méthode pour détecter les deux instants séparant les différentes phases.

5. Pour éliminer le bruit on peut faire appel à une méthode dite de lissage. Elle consiste à modifier les données initiales qui seront notées $\left(y_i^0\right)_{1 \leq i \leq n}$ en répétant un certain nombre de fois la récurrence suivante pour $k \geq 0$:

$$\begin{cases} y_i^{k+1} &= y_i^k + \alpha \frac{y_{i-1}^k - 2y_i^k + y_{i+1}^k}{h^2} \ , \ 2 \leq i \leq n-1, \\ y_1^{k+1} &= y_1^k, \\ y_n^{k+1} &= y_n^k, \end{cases}$$

où $\alpha \in \mathbb{R}$.

On se limite ici au premier intervalle de temps (avant T_1), cf. Figure 2. Plus précisément,

$$y_i^0 = a_1 + b_1 x_i + \epsilon_i \ , \ 1 \leq i \leq n,$$

où $\epsilon_i = (-1)^i \epsilon$, avec $\epsilon > 0$.

   (a) Montrer qu'une itération de lissage appliquée au vecteur $\left(y_i^0\right)_{1<i<n}$ donne un résultat de la forme :

$$a_1 + b_1 x_i + C(\alpha, h)\epsilon_i,$$

   où $C(\alpha, h)$ dépend de α et de h.

   (b) En déduire un intervalle pour α tel que cette méthode permette de réduire le bruit. Existe-t-il une valeur de α permettant de supprimer ce bruit ?

(c) Écrire en Python une fonction `Lisse` qui applique une itération de lissage. `Lisse` prend en entrée `Y`, `a` et `h` où `Y` est une liste de nombres contenant les valeurs $(y_i^k)_{1\leq i\leq n}$ (pour un certain k), `a` est la valeur de α et `h` est la valeur de h ; elle renvoie les valeurs $(y_i^{k+1})_{1\leq i\leq n}$ dans une liste.

(d) Dans la pratique, sur le problème complet de la Figure 1 *i.e.* lorsque l'on ne se limite pas au premier intervalle de temps, on applique plusieurs itérations de lissage. On obtient alors le graphe de la Figure 3.

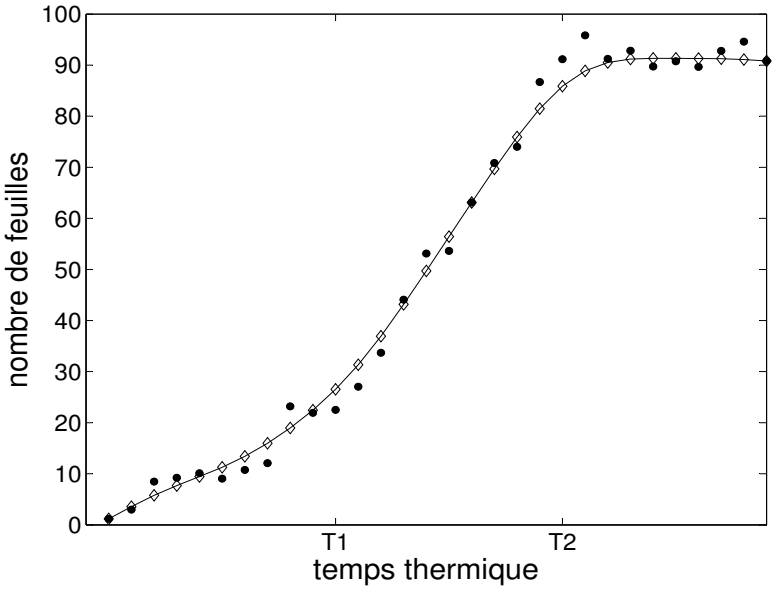

FIGURE 3 – Nombre de feuilles d'une plante en fonction du temps thermique ; y sont représentées les valeurs expérimentales initiales ('•') et les valeurs obtenues après 10 itérations de lissage ('◊').

En s'inspirant des questions précédentes, proposer une méthode pour estimer T1 et T2.

(e) Cette méthode est-elle satisfaisante pour estimer T1 et T2 ?