— MFEE - 1ère année —

Optimisation TDM : Moindres carrés et calage optimal de modèles

Objectifs du TDM :

1. Formulation d'un problème de minimisation au sens des moindres carrés.
2. Mise en œuvre de la méthode des moindres carrés pour caler les paramètres d'un modèle dans un cas concret à l'aide de Python/Matlab.
3. Analyser des résultats et critiquer un modèle.

Énoncé du TDM

Lorsque l'on souhaite décrire, ou prédire, un phénomène, on utilise souvent un **modèle**, c'est-à-dire une relation mathématique entre une entrée et une sortie : $y = f(x)$. Bien souvent, un tel modèle comporte des **paramètres** dont il convient de fixer la valeur. Le lien avec le cours d'optimisation est clair : comment déterminer la valeur des paramètres pour que le modèle décrive le phénomène de la façon la plus juste possible ? Lorsque l'on dispose de données de références, issues typiquement d'observations expérimentales, on cherchera la valeur des paramètres tel que l'écart entre ces données de références et notre modèle soit minimal. Ce projet vise à vous familiariser avec la **méthode des moindres carrés** qui comme son nom l'indique consiste à trouver la valeur des paramètres minimisant la somme des carrés des écarts entre données de références et modèle.

Dans la première partie, vous commencerez par traiter un exemple simple vous permettant de formaliser le problème et de réaliser que généralement la détermination des valeurs optimales des paramètres s'accompagne de la résolution d'un système algébrique non linéaire. La résolution d'un tel système étant généralement compliquée, vous mettrez en place un script **Python** ou **Matlab** s'appuyant sur des fonctions prédéfinie, pour déterminer les paramètres optimaux.

Dans la seconde partie, vous pourrez mettre en œuvre cette technique de calage de paramètres pour obtenir un modèle concret de prévision du débit d'un cours d'eau en fonction des précipitations ayant eu lieu dans le bassin versant.

1 Régression linéaire et non linéaire

On s'intéresse à la croissance d'une population de bactérie. On reproduit des photos à différent instant d'une colonie de bactéries issue de [2] sur la figure 1.

Pour éviter de compter, le nombre de bactéries à ces quatre instants est donné dans le tableau ci-dessous

Temps (minutes)	Population N_t (milliers)
0	1
100	5
198	48
306	505

1. Représenter les données sous forme de graphique.

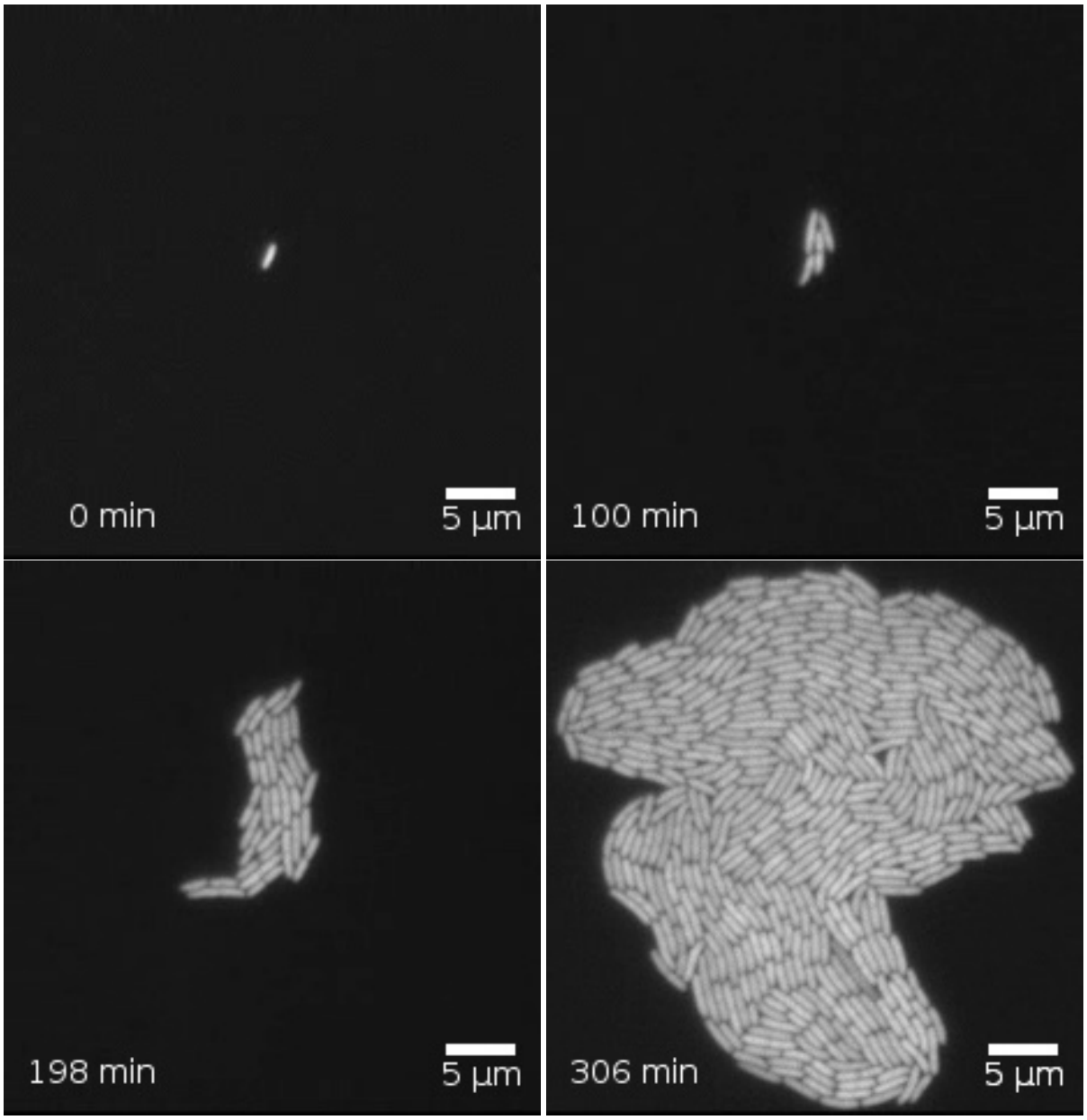

FIGURE 1 – Photos d'une colonie de bactéries à 4 instants successifs d'après [2]. N. B. Le film entier de la croissance peut être trouvé ici : `https://doi.org/10.1371/journal.pbio.0030045.sv001`

2. On suppose que l'on dispose d'un modèle à un paramètre décrivant l'évolution des bactéries : $y_{model}(t; p_0)$ (ici p_0 est le paramètre). On définit le "résidu" ϵ comme la somme sur toutes les observations du carré des écarts entre le modèle et les observations.

   Donner la définition formelle, mathématique, de ϵ.

3. On cherche la valeur de p_0 pour laquelle ϵ est minimum. Donner la condition nécessaire que doit satisfaire la valeur optimale du paramètre p_0 (c.-à-d. condition de 1er ordre).

4. On propose un premier modèle de croissance exponentielle : $y_{model}(t; p_0) = p_0 e^{+t/50}$ où p_0 est le paramètre à optimiser. Ici le temps t est donné en minutes est on suppose pour commencer que le temps caractéristique de la multiplication des bactéries est de 50 minutes. Le problème à résoudre pour trouver la valeur optimale de p_0 est-il linéaire ou non linéaire ?

5. Écrire le système algébrique correspondant à la condition optimal, trouver la valeur optimale de p_0, et comparer graphiquement l'évolution du modèle avec celle des données.

6. Afin de mieux décrire l'évolution de la population, on cherche maintenant à ajuster le temps caractéristique τ : $y_{model}(t; p_0, \tau) = p_0 e^{+t/\tau}$. Formulez de nouveau le problème des moindres carrés pour optimiser les deux paramètres p_0 et τ.
7. Pouvez-vous trouver analytiquement les paramètres optimaux ?
8. Ecrire un script en Matlab ou Python permettant de déterminer la valeur optimale des paramètres. On vous conseil de vous appuyer sur les fonctions prédéfinies, par exemple, avec Matlab vous pourrez utiliser la fonction "`lsqnonlin`" et avec Python vous pouvez utiliser la fonction "`curve_fit`" du package "scipy". A titre de validation, vérifiez que dans le cas du modèle à 1 paramètre vous retrouvez bien la valeur trouvée précédemment. N. B. vous pouvez vous inspirer des programmes mis à votre disposition sur moodle.
9. Donnez les valeurs optimales des paramètres obtenues avec votre programme et comparez graphiquement le modèle avec les données observées.
10. Commenter l'influence de la valeur initiale des paramètres du programme.

2 Prévision du débit d'un cours d'eau et calage de modèle

Dans bien des situations, il est intéressant de prévoir le débit d'un cours d'eau. On pensera par exemple aux problématiques d'irrigation pour l'agriculture, à la gestion des ouvrages hydroélectriques, ou à la prévention des crues. Il parait évident que l'un des paramètres clés pour une telle prévision concerne les précipitations captées par le bassin versant en amont de la position où l'on est intéressé au débit : de très fortes pluies étant susceptibles d'entrainer des débits importants.

2.1 Données

Vous avez à disposition sur Moodle des observations de l'évolution du débit de la source karstique pyrénéenne d'Aliou, $Q_{obs}(t)$, ainsi que des observations des précipitations dans la zone. Ces observations (réelles) couvrent environ 1 semestre avec une fréquence d'enregistrement de $\Delta t = 0.5h$ ($N = 8158$ observations). On remarquera par exemple un événement de crue/orage à l'échelle d'une semaine environ ($\sim 300\Delta t$) qui correspond aux observations $i \in [3551, 3850]$.

— Représentez les débits observés en fonction du temps (hydrogrammes) que vous comparerez avec une représentation de la pluie observée (hyétogramme).

— Quelles observations pouvez-vous faire ?

2.2 Modèle pluies / débits

On propose de modéliser le débit du cours d'eau par un de type Input/Output (entrée/sortie) en représentant toute la complexité du bassin versant par un réservoir effectif. Comme illustré sur le schéma suivant, le réservoir stocke une certaine quantité d'eau, variable, les précipitions, $P(t)$, sont les entrées et le débit du cours d'eau, $Q_{model}(t)$, est la sortie.

On cherche maintenant à relier $Q_{model}(t)$ à $P(t)$.

1. On considère que le réservoir a une section constante A et une hauteur d'eau variable $h(t)$. A partir de la conservation de la masse, donnez la variation de h en fonction de Q_{model} et P.
2. On considère de plus que la relation entre la hauteur d'eau et le débit est linéaire : $Q_{model}(t) = kAh(t)$. On appelle la constante k la conductance spécifique. Quelle est l'unité de k ?

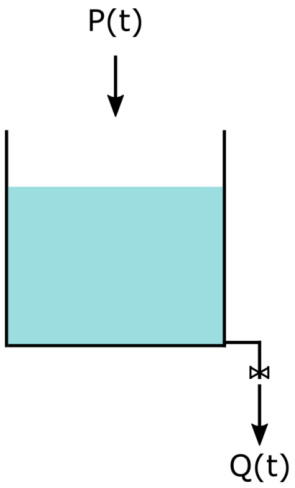

FIGURE 2 – Schéma du modèle entrée / sortie.

3. Proposer une interprétation à la constante $1/k$.
4. Obtenez une équation différentielle pour Q_{model}.
5. Résoudre cette équation différentielle par la méthode de la variation de la constante. Pour cela, poser $Q_p(t) = z(t)Q_h(t)$ où $Q_h(t)$ est la solution de l'équation différentielle homogène. On posera $Q(0) = Q_0$ le débit initial. Montrer alors que :

$$Q(t) = Q_p(t) + Q_h(t) = Q_0 e^{-kt} + \int_0^t k \exp\left(-k(t-t')\right) P(t')dt' \tag{1}$$

6. Proposer une interprétation des différents termes du modèle.
7. Pour le calcul numérique de ce modèle, il conviendra d'approximer l'intégrale sur t' par une somme discrète. Donnez cette approximation.
8. Le temps t sera aussi discrétisé : $t = i\Delta t$ avec le même pas de temps que celui des enregistrements des données. Donnez la formulation du modèle discret.
9. Quels sont les paramètres à ajuster du modèle ?

2.3 Calage des paramètres du modèle

Afin d'utiliser ce modèle pour prévoir le débit du cours d'eau, on cherche à déterminer les paramètres optimaux du modèle au sens des moindres carrés. On va donc chercher à comparer les débits observés et les débits obtenus par le modèle.

1. Donnez l'expression du résidu $epsilon$ correspondant à ce problème.
2. Décrivez l'algorithme général pour optimiser les paramètres et obtenir une prévision du débit.
3. Adapter le programme Python/Matlab de la première partie pour obtenir la valeur optimale des paramètres.
4. On propose d'examiner en 1er lieu un évènement orage/crue à l'échelle d'une semaine environ ($300\Delta t$) : $i \in [3551, 3850]$. Donnez les coefficients optimaux.
5. Proposez une comparaison graphique du débit observé et le débit obtenu par le modèle (hydrogrammes).
6. Discuter des résultats (critiques du modèle, influence de l'input, valeurs et signes des paramètres, etc ...).
7. Calculer la norme totale des résidus :

8. Répétez les étapes précédentes (calages des paramètres, discussions des résultats et calcul des résidus) pour deux autres cas : Q_0 fixé sur $Q_0 = Q_{obs}(0)$ et optimisation sous contrainte $Q_0 > 0$. Comparez les différents cas entre eux.
9. Concluez sur la pertinence du paramètre Q_0. Aurions-nous pu prévoir ce résultat à l'aide du modèle ?

2.3.1 Étude de la fenêtre d'observation

Testez l'influence de la fenêtre d'observation sur les résultats. Vous pourrez choisir des évènements particuliers sur des durées comprises entre quelques jours et quelques semaines (ne pas dépasser ~ $1000\Delta t$).
On pourra comparer les différentes fenêtres entre elles (à l'aide des résidus ou par des remarques sur les résultats). Les événements des données d'Aliou permettent de vérifier à quel point le modèle peut être pertinent.

2.4 Modification du modèle de base

On propose de modifier le modèle linéaire pour assurer de meilleurs résultats par rapport aux données observées. Pour cela, on peut introduire un retard ou bien modifier directement la loi de vidange linéaire en utilisant plusieurs réservoirs en cascade.

2.4.1 Paramètre de délai τ_0

Il est possible d'introduire un retard dans le noyau du modèle de convolution de l'équation (1) sous la forme :

$$k\exp\big(-k(t-t')\big) \to k\exp\big(-k(t-t'-\tau_0)\big) \tag{2}$$

Réitérer l'étude sur ce nouveau paramètre (comparaisons des modèles, fenêtres, résidus, valeur des paramètres ...)

2.4.2 Modèle à "M" réservoirs disposés en série (cascade de Nash)

Le modèle peut être amélioré avec un nouveau mode de couplage en disposant cette fois plusieurs réservoirs (identique au 1er) en série. Cette modélisation s'appelle la Cascade de Nash, combinant M réservoirs en cascade. Ainsi le débit sortant d'un réservoir i se retrouve dans le réservoir i+1 (voir le schéma de la figure si dessous). On a alors une atténuation de l'effet de la pluie sur la sortie finale due au passage à travers tous les réservoirs. Cela revient à remplacer le noyau de l'intégrale de l'équation (1) $h_1(t-t') = k\exp\big(-k(t-t')\big)$ par $h_M(t-t')$:

$$h_M(t-t') = \frac{1}{\Gamma(M)}\big(k(t-t')\big)^{M-1}\,k\exp\big(-k(t-t')\big) \tag{3}$$

où Γ est la fonction Gamma. Plus de détails sur la dérivation de ce modèle sont disponibles dans cette référence [1]

Réitérer l'étude avec ce nouveau modèle en discutant aussi de l'effet du nombre de réservoirs M.

Références

[1] Rachid Ababou, Denis Dartus, and Jean-Michel Tanguy. Model Coupling in Mathematical Models, chapter 15, pages 445–492. John Wiley Sons, Ltd, 2010.

[2] Eric J Stewart, Richard Madden, Gregory Paul, and François Taddei. Aging and death in an organism that reproduces by morphologically symmetric division. PLOS Biology, 3(2) :null, 02 2005.

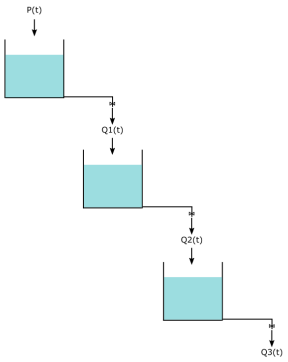

FIGURE 3 – Schéma du modèle de 3 réservoirs en cascade.