

Du modèle déterministe au modèle probabiliste

Jean-Jacques Boreux

ULg, Département des Sciences et Gestion de l'Environnement.

CEREGE, ECCOREV, le 19 mai 2008

Abstract

La non linéarité du modèle déterministe et la rareté des données sont des situations assez courantes en science de l'environnement. Dans ces conditions, l'estimation ponctuelle du paramètre d'un modèle déterministe ne saurait suffire, surtout dans une perspective d'aide à la décision. C'est pourquoi nous proposons de considérer la réponse du modèle déterministe comme l'espérance conditionnelle de l'observable. Cette stratégie, très conforme à l'esprit bayésien, permet de quantifier les incertitudes et ainsi de fournir une aide à la décision.

Part I

La chambre d'émission

0.1 Mise en situation

L'air ambiant des habitations contient des *composés organiques volatiles (COV)* émis par des objets courants. Par exemple, un morceau de bois traité émet un ou plusieurs polluants au taux R (masse de polluant émise par unité de surface et par unité de temps). C'est le profil temporel de cette émission qui nous intéresse, notamment l'instant à partir duquel cette émission est jugée inoffensive pour la santé.

Le modèle le plus courant suppose que la variation du taux d'émission dans un intervalle de temps infinitésimal est proportionnelle au taux d'émission et que la constante de proportionnalité est invariante dans le temps

$$\frac{dR(t)}{dt} = -\eta R(t) \quad (1)$$

- $R \equiv$ le taux d'émission : $[R] = ML^{-2}T^{-1}$
- $\eta \equiv$ la constante de proportionnalité : $[\eta] = T^{-1}$

L'intégration de (1) sur l'intervalle réel $[0, t]$ donne

$$R(t) = R_0 \exp(-\eta t) \quad (2)$$

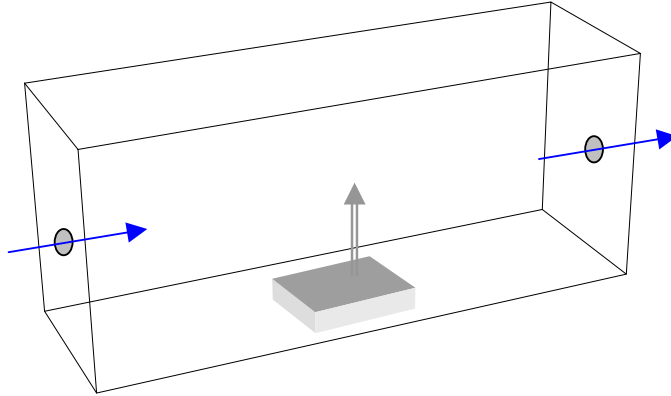
où $R_0 \equiv R(t=0)$ représente le taux d'émission initial.

Remarque 1 *Le temps t_α au bout duquel le taux d'émission n'est plus que la fraction α du taux initial R_0 est une quantité d'un grand intérêt opérationnel. Par exemple, pour $\alpha = 0.05$:*

$$t_\alpha = \frac{-\ln \alpha}{\eta} \simeq \frac{3.0}{\eta} \quad (3)$$

Remarque 2 *Du point de vue de l'émission de formaldéhydes dans l'air ambiant, le modèle (2) montre qu'une petite valeur de R_0 et une grand valeur de η constituent un scénario favorable pour celui qui met le matériau testé sur le marché. Nous y reviendrons plus loin.*

Une mesure directe du taux d'émission n'est pas possible et c'est pourquoi on utilise une *chambre d'émission*. Il s'agit d'une enceinte ventilée dans laquelle on place un échantillon de l'objet émetteur (Figure ??).



Représentation schématique d'une chambre d'émission.

La concentration instantanée en polluant résulte d'un bilan de masse. $R(t)$ étant le taux d'émission instantané (2), la variation de la masse du polluant dans l'enceinte dans l'intervalle dt est la différence entre la masse de polluant émise par le matériau et la masse perdue par ventilation pendant ce même laps de temps :

$$V \frac{dC(t)}{dt} = AR(t) - qVC(t) \quad (4)$$

- $A \equiv$ l'aire de l'objet émetteur : $[A] = L^2$
- $C(t) \equiv$ la concentration au temps t : $[C] = ML^{-3}$
- $V \equiv$ le volume de l'enceinte : $[V] = L^3$
- $q \equiv$ le renouvellement de l'air : $[q] = T^{-1}$

On reconnaît une équation différentielle linéaire du premier ordre dont la solution sur l'intervalle réel $[0, t]$ est (v. annexe)

$$C(t) = \frac{aR_0}{q - \eta} [\exp(-\eta t) - \exp(-qt)] \quad (5)$$

où $a \equiv A/V$, $[a] = L^{-1}$.

Ce modèle déterministe est donc caractérisé par un paramètre bidimensionnel qu'il est commode de noter θ :

$$\theta \equiv (R_0, \eta)$$

Pour obtenir des informations sur θ , on mesure la concentration en des temps discrets à la sortie de la chambre, i.e. n couples C_i, t_i .

Remarque 3 *La durée typique d'une telle expérience est d'une dizaine de jours pendant laquelle la ventilation, q , doit être maintenue rigoureusement constante. On conçoit aisément que la répétition de l'expérience dans les mêmes conditions est malaisée et ... coûteuse.*

0.1.1 Exemple numérique

Cet exemple est extrait de (Hayter et Dowling)¹. Il traite de l'émission de formaldéhydes (COV toxiques) par des tapis. Un morceau de tapis d'aire $A = 210 \text{ cm}^2$ est placé dans une chambre d'émission de volume $V = 53 \text{ dm}^3$ ventilée au taux de $q = 1.01 \text{ h}^{-1}$ (Table 1).

t	0.5	1.5	2.5	4.5	8.5	24.5	48.5
C	0.219	0.397	0.410	0.549	0.333	0.243	0.163
t	72.5	144.5	168.5	196.5	216.5	240.5	
C	0.132	0.019	0.031	0.027	0.023	0.018	

Table 1: Concentration en formaldéhydes ($\mu\text{g}/\text{m}^3$) en fonction du temps (h).

On rappelle que $a = A/V$, i.e. $a = \frac{210 \times 10^{-4}}{53 \times 10^{-3}} \simeq 0.40 \text{ m}^{-1}$.

0.2 Première solution : minimiser la somme des carrés des écarts

Le modèle déterministe (??) donne la concentration du polluant dans la chambre d'émission à l'instant t . Il est commode de la noter $F(\theta, t)$:

$$F(\theta, t) = \frac{aR_0}{q - \eta} [\exp(-\eta t) - \exp(-qt)]$$

Soit SCE la somme des carrés des écarts entre l'observation C_i réalisée au temps t_i et la valeur $F(\theta, t_i)$

$$SCE = \sum_{i=1}^n (C_i - F(\theta, t_i))^2$$

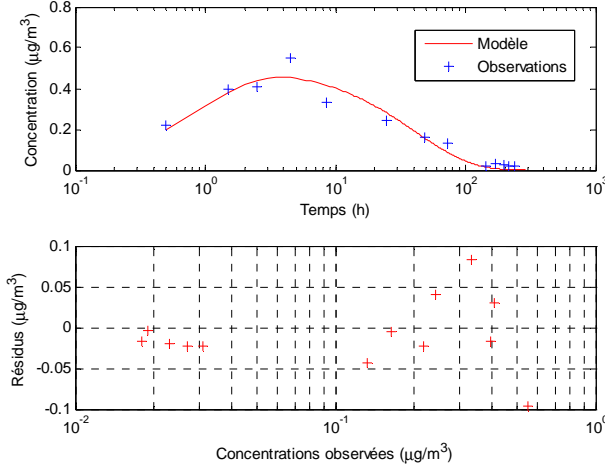
L'estimation de θ par la méthode des moindres carrés n'a pas de solution analytique car le modèle (4) est non linéaire et aucune transformation ne peut le linéariser. La fonction *lsqcurvefit* de Matlab fournit une solution ponctuelle

$$\hat{R}_0 \simeq 1.27 \mu\text{g}/\text{m}^2/\text{h} \quad \hat{\eta} \simeq 0.024 \text{ h}^{-1} \quad (6)$$

En injectant $\hat{\eta}$ dans la relation (3), on voit qu'il faut un peu plus de cinq jours pour que le taux d'émission perde 95% de sa valeur initiale \hat{R}_0 .

La figure (??) montre l'ajustement. On remarque que les résidus ne sont pas uniformément distribués.

¹Hayter, T. and M. Dowling (1993). Experimental Designs and Emission Rate Modelling for Chamber Experiments. Atmospheric Environment, **27A** (14), pp. 2225-2234.



Estimation ponctuelle des paramètres et résidus correspondants.

0.3 Seconde solution : bruiteur le modèle déterministe

0.3.1 Résoudre les équations différentielles entre deux temps quelconques

Pour $i = 1, \dots, n$, posons

$$n_i \equiv t_i - t_{i-1}, \quad R_i \equiv R(t_i), \quad C_i \equiv C(t_i)$$

La résolution des équations différentielles (1) et (4) entre les temps t_{i-1} et t_i conduit aux résultats suivants :

$$R_i = R_{i-1} \exp(-\eta n_i) \quad (7)$$

$$C_i = \frac{aR_{i-1}}{q - \eta} \{ \exp(-\eta n_i) - \exp(-qn_i) \} \quad (8)$$

0.3.2 Choisir une stratégie

Pour bruiteur le modèle déterministe, plusieurs stratégies sont possibles.

On va assumer que (7) donne la *vraie* valeur du taux d'émission à l'instant t_i . Pourquoi? C'est un choix de modélisation qui révèle une certaine confiance dans le modèle déterministe de décroissance exponentielle (modèle 1). Par récurrence on obtient

$$R_i = R_0 \exp(-\eta t_i)$$

En tenant compte de ce résultat, la relation (8) devient

$$C_i = \frac{aR_0}{q - \eta} \{ \exp(-\eta t_i) - \exp(-qn_i - \eta t_{i-1}) \} \quad (9)$$

car

$$\exp(-\eta t_{i-1}) \exp(-\eta n_i) = \exp(-\eta t_i)$$

D'un point de vue bayésien, la concentration C_i est une variable aléatoire (v.a.) car elle dépend des paramètres R_0 et η , inconnus et incertains (rappelons que les temps t_i , a et q sont connus). Bien sûr, conditionnellement à la connaissance des paramètres et des temps t_i et t_{i-1} , la relation (9) fournit une valeur particulière, c'est-à-dire $C_i = c_i$.

0.3.3 Du modèle déterministe au modèle probabiliste

Soit Y_i l'observable, c.-à-d. la concentration en polluant mesurée à la sortie de la chambre d'émission. Avant la mesure c'est une v.a. réelle Y_i ; après la mesure c'est le nombre réel y_i . La v.a. Y_i n'est pas la v.a. C_i car d'inévitables fluctuations des conditions de l'expérience et les erreurs de mesure perturbent les résultats. Par exemple, de très légères variations du taux d'aération de la chambre, q , agissent sur C_i et, de toute façon, sa répartition homogène dans la chambre est une vue de l'esprit (pensons à la concentration au point de mesure). Bref, l'observation sera tantôt au-dessus de la prévision déterministe, tantôt en-dessous.

Un modèle probabiliste réaliste et très simple consiste à ajouter un bruit blanc à C_i

$$Y_i = C_i + \varepsilon_i, \quad \varepsilon_i \sim \text{dnorm}(0, \tau) \quad (10)$$

Ce modèle est caractérisé par un paramètre tridimensionnel

$$\theta = (R_0, \eta, \tau)$$

En d'autres termes, on a considéré que la réponse du modèle déterministe est l'espérance conditionnelle de l'observable :

$$E(Y_i | \theta, t_i, t_{i-1}, q, a) = C_i$$

Remarque 4 *D'autres modèles sont possibles. Par exemple*

$$Y_i = C_i \exp(\varepsilon_i), \quad \varepsilon_i \sim \text{dnorm}\left(0, \frac{\tau}{t_i - t_{i-1}}\right)$$

Comme notre propos est d'illustrer une procédure, nous nous en tiendrons au modèle (10).

0.3.4 Inférence bayésienne sous WinBUGS

A priori, les trois composantes du paramètre θ sont indépendantes

$$R_0 \perp \eta \perp \tau \Rightarrow [\theta] = [R_0] [\eta] [\tau]$$

Pour R_0 , on a choisi un prior normal limité aux réels positifs (11, figure 1 gauche); pour les paramètres η et τ , un prior gamma est un choix classique (12, figure 1 droite).

$$[R_0|m, p] = \sqrt{\frac{p}{2\pi}} \exp\left(-\frac{p}{2}(R_0 - m)^2\right) \times 1_{[0, \infty)}(R_0) \quad (11)$$

$$[\eta|a, b] = \frac{b^a}{\Gamma(a)} \eta^{a-1} \exp(-b\eta) \quad (12)$$

Un prior non informatif est obtenu en faisant tendre la précision p et les hyperparamètres a et b vers zéro (Figure, 1).

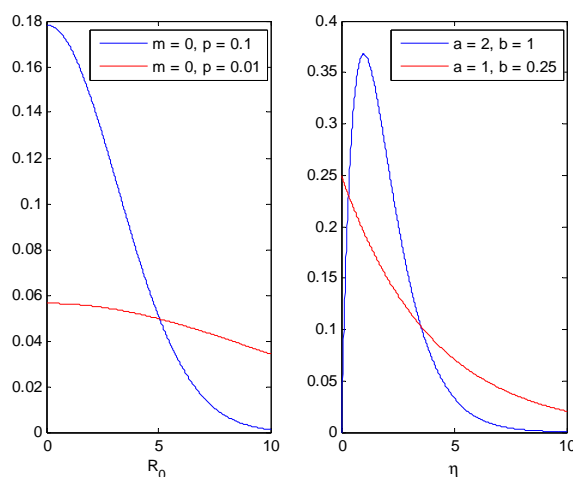


Figure 1: les priors

Le code WinBUGS est donné en annexe. Après 6000 itérations dont 1000 pour la période de chauffe, on obtient les résultats suivants (figures 2 et 3, table 2).

0.3.5 Commentaires

- Les estimations ponctuelles sont dans l'intervalle de crédibilité.
- Les paramètres R_0 et η sont linéairement corrélés ($p = 0.68$). Ils ne l'étaient pas *a priori*!

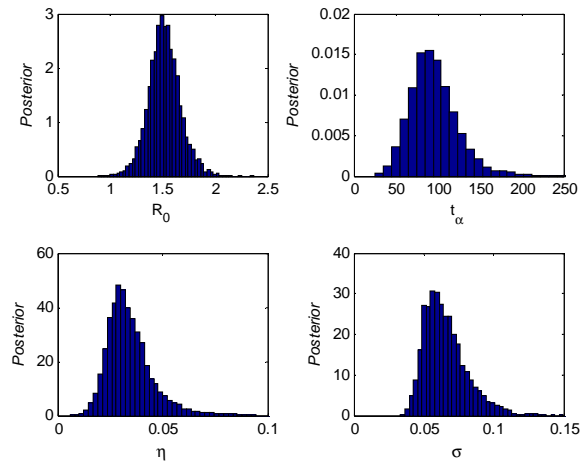


Figure 2: Les distributions *a posteriori*.

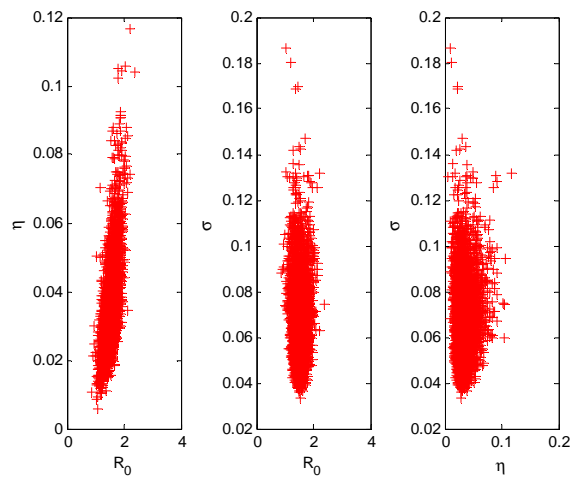


Figure 3: Corrélations linéaires entre les paramètres : $\rho(\eta, R_0) = 0.68$.

Paramètre	Unité	2.5 ^e	50 ^e	97.5 ^e	SCE
R_0	$\mu g/m^2/h$	1.23	1.53	1.88	1.27
η	h^{-1}	0.018	0.033	0.064	0.024
σ	-	0.043	0.063	0.102	-
$t_{0.05}$	h	47	91	165	124

Table 2: Intervalle de crédibilité à 95 pourcent.

<i>Scenarii</i>	$R_0 < \hat{R}_0$	$R_0 > \hat{R}_0$
$\eta > \hat{\eta}$	0.01	0.86
$\eta < \hat{\eta}$	0.03	0.10

Table 3: Table Caption

- Il y a 14 chances sur cent que t_α excèdent cinq jours ($\alpha = 0.05$) (Figure 2, b).
- Si on disposait de valeurs guides r_0^* et η^* , l'événement $R_0 < r_0^*$ et $\eta > \eta^*$ définirait un scénario favorable au marchand de tapis. En l'absence de telles valeurs guides, les références peuvent être les estimations ponctuelles (6). La table (3) montre quatre *scenarii* et leur probabilité respective calculée à partir des chaînes de Markov (au-delà de la période de chauffe).

0.4 Conclusion

Un modèle est déterministe s'il ne fait pas appel au calcul de probabilités. Un modèle est stochastique s'il fait appel au calcul de probabilités. Le premier est une construction intellectuelle basée sur des lois physiques, par exemple un bilan de masse. Le second est une construction intellectuelle qui imagine un mécanisme aléatoire générateur des observations. Tous les modèles, qu'ils soient déterministes (expliquer les observations) ou probabilistes (interpréter les observations), sont des caricatures de la réalité. Cette simplification est le prix qu'il faut payer pour la rendre intelligible. Or la non linéarité et la rareté des données sont des situations assez courantes en science de l'environnement. Dans ces conditions, l'estimation ponctuelle du paramètre d'un modèle déterministe ne saurait suffire, surtout dans une perspective d'aide à la décision qui ne peut jamais faire l'impasse sur les incertitudes. C'est pourquoi nous proposons de considérer la réponse du modèle déterministe comme l'espérance conditionnelle de l'observable. Cette stratégie, très conforme à l'esprit bayésien, permet de quantifier les incertitudes et ainsi de fournir une aide à la décision crédible.

0.5 Annexe 1

0.6 Equation différentielle du premier ordre à variables séparées

On la résout dans l'intervalle réel $[a, t]$

$$\frac{dR(t)}{dt} = -\eta R(t)$$

On a trivialement

$$\int_{R_a}^{R(t)} \frac{dR(t)}{R(t)} = -\eta \int_a^t dt \Rightarrow \ln \frac{R(t)}{R_a} = -\eta(t-a)$$

Finalement

$$R(t) = R_a \exp\{-\eta(t-a)\}$$

0.6.1 Equation différentielle linéaire du premier ordre

On la résout dans l'intervalle réel $[a, t]$

Soit à résoudre

$$\frac{dC(t)}{dt} + qC(t) = \frac{A}{V} R_a \exp\{-\eta(t-a)\} \quad (13)$$

On résout d'abord l'équation réduite

$$\frac{dC(t)}{dt} + qC(t) = 0 \Rightarrow C(t) = C_a \exp\{-q(t-a)\}$$

Ensuite on fait varier la constante d'intégration C_a de telle sorte que

$$C(t) = C_a(t) \exp\{-q(t-a)\}$$

soit la solution générale de l'équation différentielle initiale

$$\frac{dC(t)}{dt} = \frac{dC_a(t)}{dt} \exp\{-q(t-a)\} - \underbrace{qC_a(t) \exp\{-q(t-a)\}}_{C(t)}$$

c.-à-d.

$$\frac{dC(t)}{dt} + qC(t) = \frac{dC_a(t)}{dt} \exp\{-q(t-a)\} \quad (14)$$

L'identité des relations (13) et (14) implique

$$\frac{dC_a(t)}{dt} \exp\{-q(t-a)\} = \frac{A}{V} R_a \exp\{-\eta(t-a)\}$$

$$\frac{dC_a(t)}{dt} = \frac{A}{V} R_a \exp\{(q-\eta)(t-a)\}$$

On en déduit

$$C_a(t) = \frac{AR_a}{V(q-\eta)} [\exp\{(q-\eta)(t-a)\} - 1]$$

La solution générale est donc

$$C(t) = \frac{AR_a}{V(q-\eta)} [\exp\{-\eta(t-a)\} - \exp\{-q(t-a)\}] \quad (15)$$

Quand $a = 0$:

$$C(t) = \frac{AR_0}{V(q-\eta)} [\exp(-\eta t) - \exp(-qt)]$$

0.7 Annexe 2

```
model;
{
  for(i in 2 : N) {
    n[i] <- t[i] - t[i - 1]
    C[i] <- kappa * (exp((-eta) * t[i]) - exp((-q) * n[i] - eta * t[i - 1]))
    y[i] ~ dnorm(C[i], tau)
  }
  tau ~ dgamma(0.001, 0.001)
  R0 ~ dnorm(0.0, 1.0E-6) I(0, )
  eta ~ dgamma(0.001, 0.001)
  kappa <- (a * R0) / (q - eta)
  sigma <- 1/sqrt(tau)
  T <- (-log(0.05))/eta
}

list(t=c(0,0.5, 1.5, 2.5, 4.5, 8.5, 24.5, 48.5, 72.5, 144.5, 168.5, 196.5, 216.5, 240.5), y=c(0, 0.219,
0.397, 0.410, 0.549, 0.333, 0.243, 0.163, 0.132, 0.019, 0.031, 0.027, 0.023, 0.018), N=14,
q=1.01, a=0.4)

list(tau=1, R0=1, eta=0.02)
list(tau=1, R0=2, eta=0.01)
```

Figure 4: Code WinBUGS.