Plans d'expériences pour modèles non linéaires: des problématiques et des challenges en biologie

Séminaire de Statistique du CNAM
Paris, 22 Janvier 2015

Jean-Pierre Gauchi

Institut National de la Recherche Agronomique
Unité de Mathématiques et Informatique Appliquées du Génome à l'Environnement
Jouy-en-Josas, France.

Portrait

L'Institut national de la recherche agronomique

- Créé en 1946
- Établissement public à caractère scientifique et technologique
- Placé sous la double tutelle des ministères en charge de l'agriculture et de la recherche
- 2e organisme de recherche publique français avec près de 9000 collaborateurs et un budget de 680 millions d'euros
- 1er organisme européen de recherche agronomique

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT
INRA

Portrait

Trois champs d'intervention majeurs...

Alimentation Agriculture Environnement

... déclinés en six axes de recherche :

- > Environnement et espace rural
- > Alimentation humaine et sécurité des aliments
- > Qualité des produits agricoles
- > Connaissance du vivant
- > Pratiques et systèmes agricoles
- > Sciences sociales

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Plan de l'exposé

- Introduction
- Partie I : Plans d'expériences dans un cadre dynamique
- Partie II : Plans d'expériences dans un cadre statique
- Conclusion

Introduction

- En biologie, dans de nombreux sujets de recherche des expériences **difficiles** et **coûteuses** doivent être conduites.
- On se focalisera ici sur le contexte de la **sécurité alimentaire** dans lequel les microbiologistes sont amenés à construire des modèles mathématiques de **croissance** / **décroissance** des populations bactériennes pour :
  - aider à la **compréhension** des mécanismes de la dynamique bactérienne,
  - fournir aux agences gouvernementales sanitaires (typiquement l'ANSES) des éléments scientifiques pour établir *in fine* des **dates de limite de consommation des aliments**, des **durées maximales de stockage**, des **températures seuils de la chaîne du froid**, etc.

Listeria monocytogenes : une bactérie pathogène en MA

- Partie I : Plans d'expériences dans un cadre dynamique

Quelques problématiques en microbiologie prédictive

- Identification de systèmes dynamiques bactériens
- Inférence et Sélection statistique de modèle
- Prédiction d'évolution-analyse de risque
- Contrôle d'évolution bactérienne
- Détermination des instants optimaux des observations dans les modèle dynamiques
  $\Longrightarrow$ **plans d'expériences dynamiques**

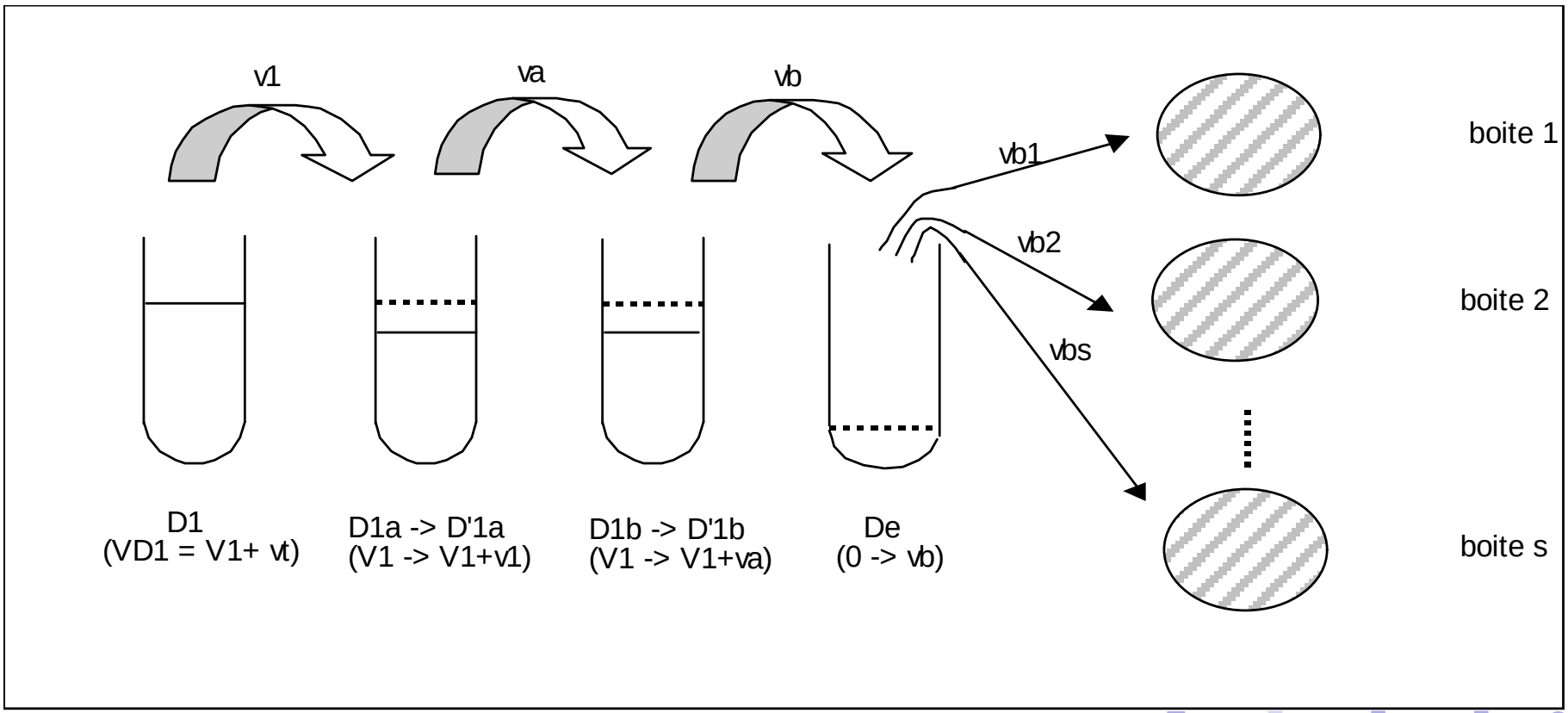
Les données brutes directement observables

Des comptages de colonies de bactéries sur boites de Petri, à différents temps (UFC_t) :

Les données à modéliser

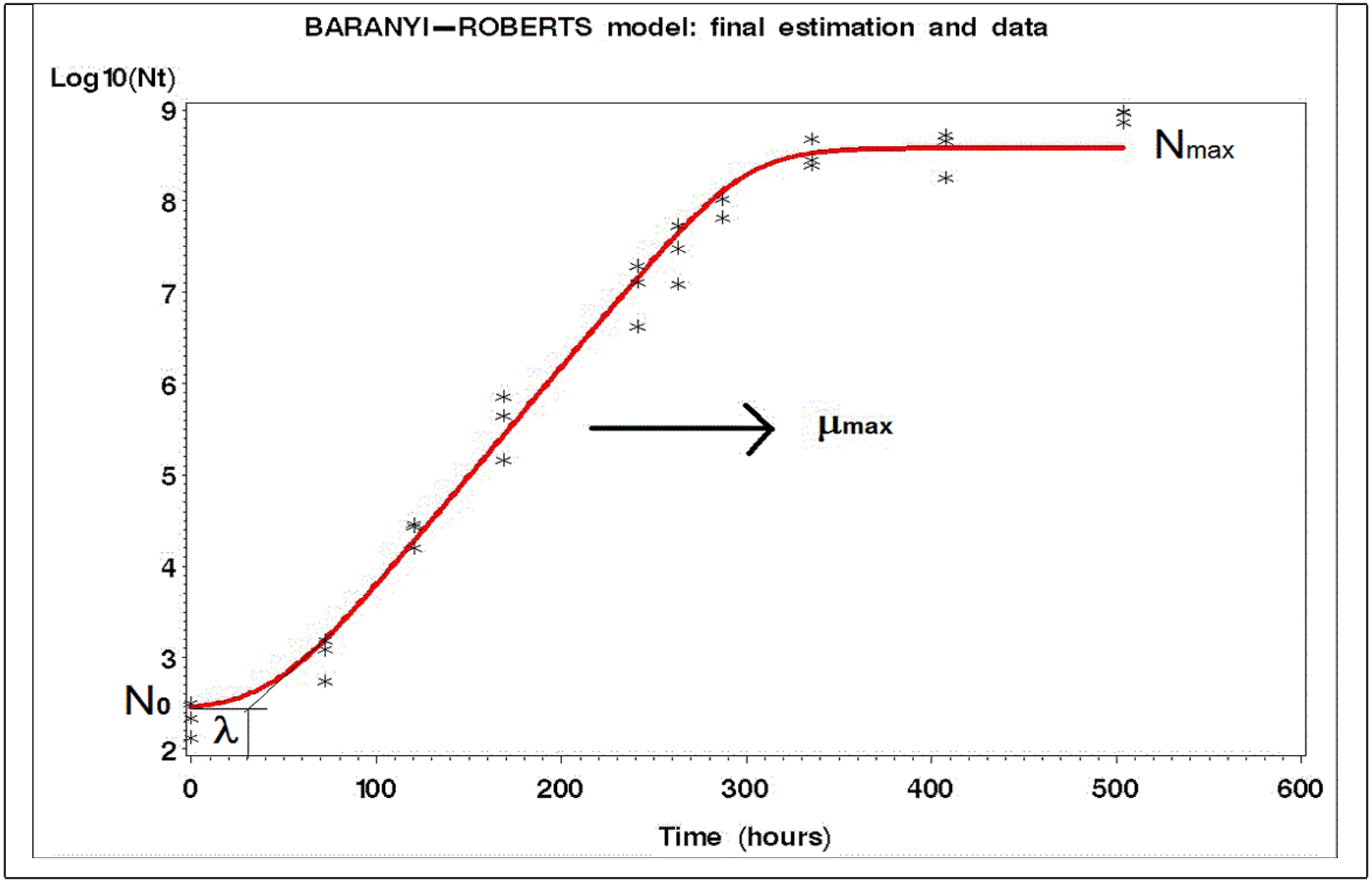
Non directement observables: ce sont les nombres de bactéries N_t dans le milieu primaire où la croissance bactérienne a lieu

$\Longrightarrow$ ces nombres sont obtenus par calcul à partir des UFC_t en tenant compte de la série de prélèvements-dilutions appliquée à un prélèvement primaire dans le milieu primaire:



Un modèle dynamique primaire de croissance bactérienne

Le graphe du modèle de Baranyi & Roberts (1995)



La fonction du modèle de Baranyi & Roberts

$$\eta(\theta, t) = \log_{10} N_t(\theta) = \frac{1}{\ln(10)} \left[\ln(N_0) + \mu_{\max} A - \ln(B)\right]$$

avec :

- $N_t(\theta) =$ nombre de bactéries "calculé" dans le milieu primaire au temps t
- $A = t + \frac{1}{\mu_{\max}} \ln(A_0)$
- $A_0 = \exp(-\mu_{\max} t) + \exp(-\mu_{\max} \lambda) - \exp(-\mu_{\max} t - \mu_{\max} \lambda)$
- $B = 1 + \frac{[\exp(\mu_{\max} A)] - 1}{\frac{N_{\max}}{N_0}}$
- le vecteur des paramètres : $\theta = (N_0, \lambda, \mu_{\max}, N_{\max})^T$

Des difficultés intrinsèques

- La croissance n'est donc **pas directement observable**,
- Les processus de prélèvement, dilution et comptage sont sophistiqués, chaque étape étant entâchée d'erreur expérimentale, rendant **impossible** l'expression d'une fonction de vraisemblance des observations,
- On cherche à estimer les paramètres **naturels** θ, mais aussi les **coefficients de variation** des erreurs précédentes.

$\Longrightarrow$ **l'utilisation des méthodes classiques d'estimation (régression non linéaire paramétrique ou non, ...) est impossible.**

Le contexte de modélisation

Système dynamique à espace d'état non linéaire observé indirectement:

$$\begin{cases} x_t = f_t(x_{t-1}, \theta, \varepsilon_t) \\ \theta_t = \theta_{t-1} \\ y_t \sim \mathcal{L}(x_{t-1}, \theta) \end{cases}$$

où:

- $x_t \in \mathbb{R}^d$, $\theta \in \mathbb{R}^p$, $y_t \in \mathbb{R}^s$,
- f_t est une fonction connue,
- ε_t est un vecteur de bruit blanc,
- $\mathcal{L}$ est la loi de probabilité de y_t (en général de forme analytique inconnue MAIS simulable).
- l'équation d'état $\theta_t = \theta_{t-1}$ est rajoutée pour associer l'estimation des densités conditionnelles des paramètres θ à celle des variables x_t.

Le cadre de l'identification et estimation

- On propose de se placer dans le cadre de méthodes de **filtrage particulaire** pour estimer θ.
- En particulier on a choisi une technique **non paramétrique** de filtrage non linéaire particulaire basée sur une approche à **noyaux de convolution et un rééchantillonnage particulaire** (Rossi & Vila, 2005, 2006).
- Cette technique est programmée dans le **logiciel convivial FILTREX** (en langage Matlab) qui offre à ce jour trois fonctionnalités:
  - l'dentification paramétrique,
  - la comparaison de deux modèles,
  - la simulation de stratégies temporelles optimales d'échantillonnage (plans d'expériences dynamiques).

Filtre à convolution de particules -1

Objectif:

- En tout temps t estimation de:

$$p_t\left(x_t|y_{1:t}\right) \quad \text{et} \quad E\left\{x_t|y_{1:t}\right\}$$
$$p_t\left(\theta_t|y_{1:t}\right) \quad \text{et} \quad E\left\{\theta_t|y_{1:t}\right\}$$

Principes de base:

- Estimations récursives non paramétriques des densités conditionnelles des variables d'état et des paramètres par:

$$\begin{aligned} p_t\left(x_t|y_{1:t}\right) &= p_t\left(x_t, y_{1:t}\right)/p_t\left(y_{1:t}\right) \\ p_t\left(\theta_t|y_{1:t}\right) &= p_t\left(\theta_t, y_{1:t}\right)/p_t\left(y_{1:t}\right) \end{aligned}$$

Filtre à convolution de particules -2

L'algorithme: génération séquentielle de particules

- à $t = 0 : \breve{x}_0^i \sim p_0^x$; $\breve{\theta}_0^i \sim p_0^\theta$; $\tilde{\varepsilon}_0^i \sim \mathcal{L}_{\varepsilon_0}$, $i = 1, ..., n$
- à $t = 1 : \tilde{x}_1^i = f_1\left(\breve{x}_0^i, \breve{\theta}_0^i, \tilde{\varepsilon}_0^i\right)$; $\tilde{\theta}_1^i = \breve{\theta}_0^i$; $\tilde{y}_1^i \sim \mathcal{L}_t\left(\tilde{x}_1^i, \tilde{\theta}_1^i\right)$
- à $t > 1$ (prédiction):

$$\begin{aligned} \breve{x}_{t-1}^i &\sim \hat{p}_{t-1}^n(x_{t-1}|y_{1:t-1}) \; ; \; \tilde{\varepsilon}_{t-1}^i \sim \mathcal{L}_{\varepsilon_{t-1}} \\ \breve{\theta}_{t-1}^i &\sim \hat{p}_{t-1}^n(\theta_{t-1}|y_{1:t-1}) \end{aligned}$$

$$\Longrightarrow \tilde{x}_t^i = f_t\left(\breve{x}_{t-1}^i, \breve{\theta}_{t-1}^i, \tilde{\varepsilon}_{t-1}^i\right); \tilde{\theta}_t^i = \breve{\theta}_{t-1}^i; \tilde{y}_t^i \sim \mathcal{L}_t\left(\tilde{x}_t^i, \tilde{\theta}_t^i\right)$$

- $t = t + 1$

  Rq: la convergence est prouvée.

Filtre à convolution de particules -3

Estimations des densités (correction):
Sur la base de densités *a priori* (typiquement uniformes) en $t = 0$ pour x_t et pour θ, les estimations des densités à tout temps $t > 0$ sont calculées avec les formules à noyaux:

$$\begin{aligned}
\hat{p}_t^n(\theta|y_{1:t}) &= \frac{\sum_{i=1}^n K_{h_n}^y(\tilde{y}_t^i - y_t) \times K_{h_n}^\theta(\tilde{\theta}_t^i - \theta)}{\sum_{i=1}^n K_{h_n}^y(\tilde{y}_t^i - y_t)} \\
\hat{p}_t^n(x|y_{1:t}) &= \frac{\sum_{i=1}^n K_{h_n}^y(\tilde{y}_t^i - y_t) \times K_{h_n}^x(\tilde{x}_t^i - x)}{\sum_{i=1}^n K_{h_n}^y(\tilde{y}_t^i - y_t)}
\end{aligned}$$

où les triplets $(\tilde{x}_t^i, \tilde{\theta}_t^i, \tilde{y}_t^i)$, $i = 1, \ldots, n$, sont les n **particules** échantillonnées selon les densités estimées au temps précédent selon le modèle à espace d'état.

Application au modèle de Baranyi-Roberts -1

Mise sous forme autorégressive de l'équation d'état:

$$
\begin{aligned}
N_{t+1} &= \delta\ N_0\ \exp(\mu_{max}\ A_t)\ \frac{1}{B_t}\ (\mu_{max}\ \frac{dA_t}{dt} - \frac{dB_t}{dt}\ \frac{1}{B_t}) + N_t + \varepsilon_t \\
A_t &= t + \frac{1}{\mu_{\max}} \ln\left(\exp\left(-C_t\right) + \exp\left(\mu_{\max}\lambda\right) - \exp\left(-C_t - \mu_{\max}\lambda\right)\right) \\
B_t &= 1 + \frac{\exp\left(\mu_{\max}A_t\right) - 1}{\left(N_{\max}/N_0\right)} \\
C_t &= \mu_{\max}t
\end{aligned}
$$

où:

- N_0 est le nombre de bactéries à $t_0 = 0$,
- ε_t est une erreur de Poisson,
- δ est le pas de discrétisation (schéma d'Euler).

Application au modèle de Baranyi-Roberts -2

Le modèle d'observation:

$$y_t \sim \mathcal{L}(x_t, \theta_t)$$

où:

- y_t est le comptage qui suit la loi $\mathcal{L}$, conditionnellement à x_t, et à θ_t.
- la loi $\mathcal{L}$ est basée sur le processus d'échantillonnage-dilution: elle est simulable.
- les erreurs sont de type:
  - Poisson (échantillonnage dans les tubes de dilution),
  - Gaussien (lors de la préparation des volumes de dilution),
  - Lognormal (lors du comptage sur les boites de Petri).

Application au modèle de Baranyi-Roberts -3

Le protocole

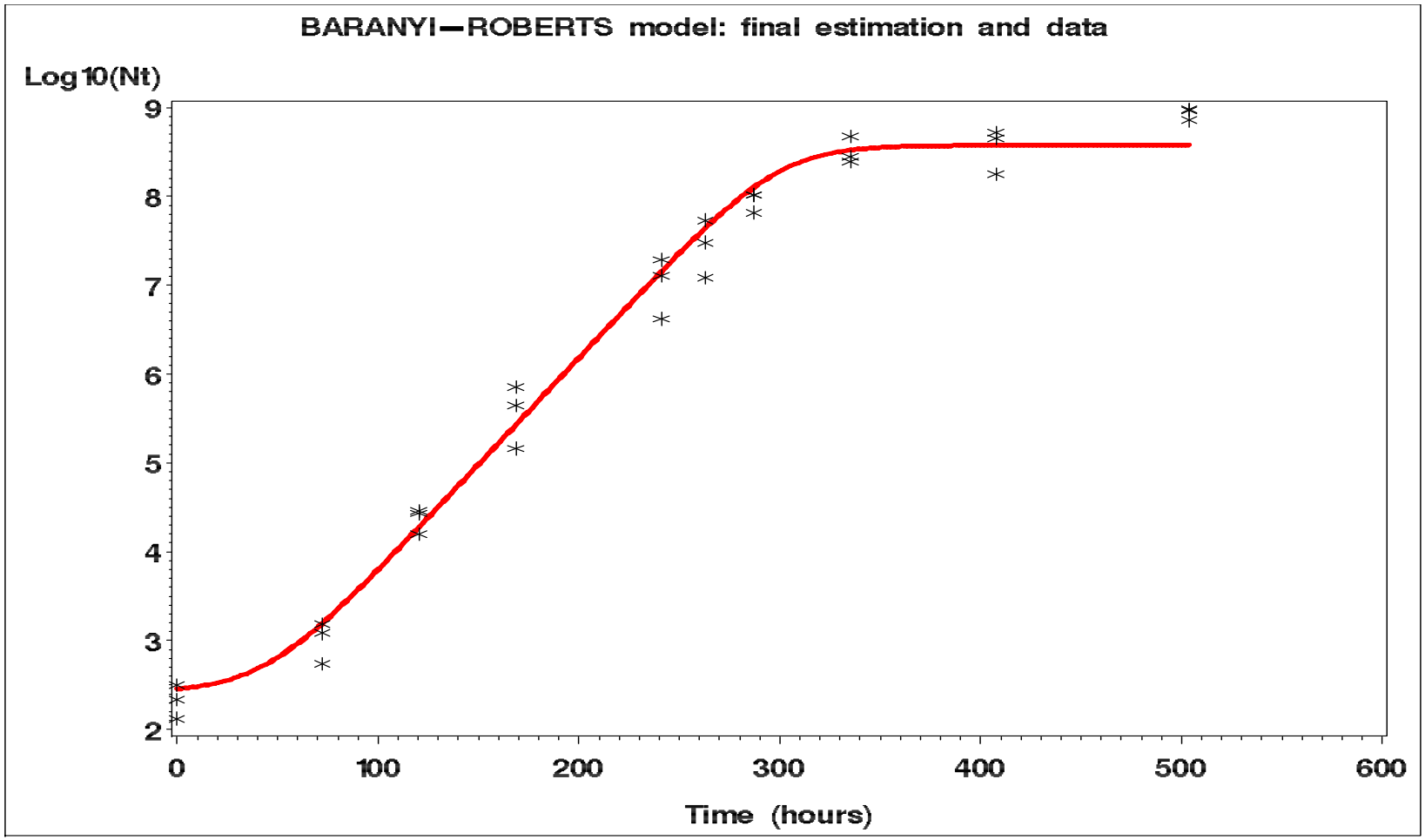
- Les données de comptage obtenues aux 10 temps suivants (en heures) choisis "naivement":
  0, 72, 120, 168, 240, 264, 288, 336, 408, 504.
- Trois boites de Petri préparées à chacun des 10 temps.
- Cinq facteurs de dilution différents selon le temps.
- Les plages de variation postulées pour les paramètres:

$$\mu_{\max} \in [0.01\,;\,2]\,;\lambda \in [10\,;\,70]\,;\,N_0 \in [100\,;\,400]\,;\,N_{\max} \in [10^8;10^9]$$

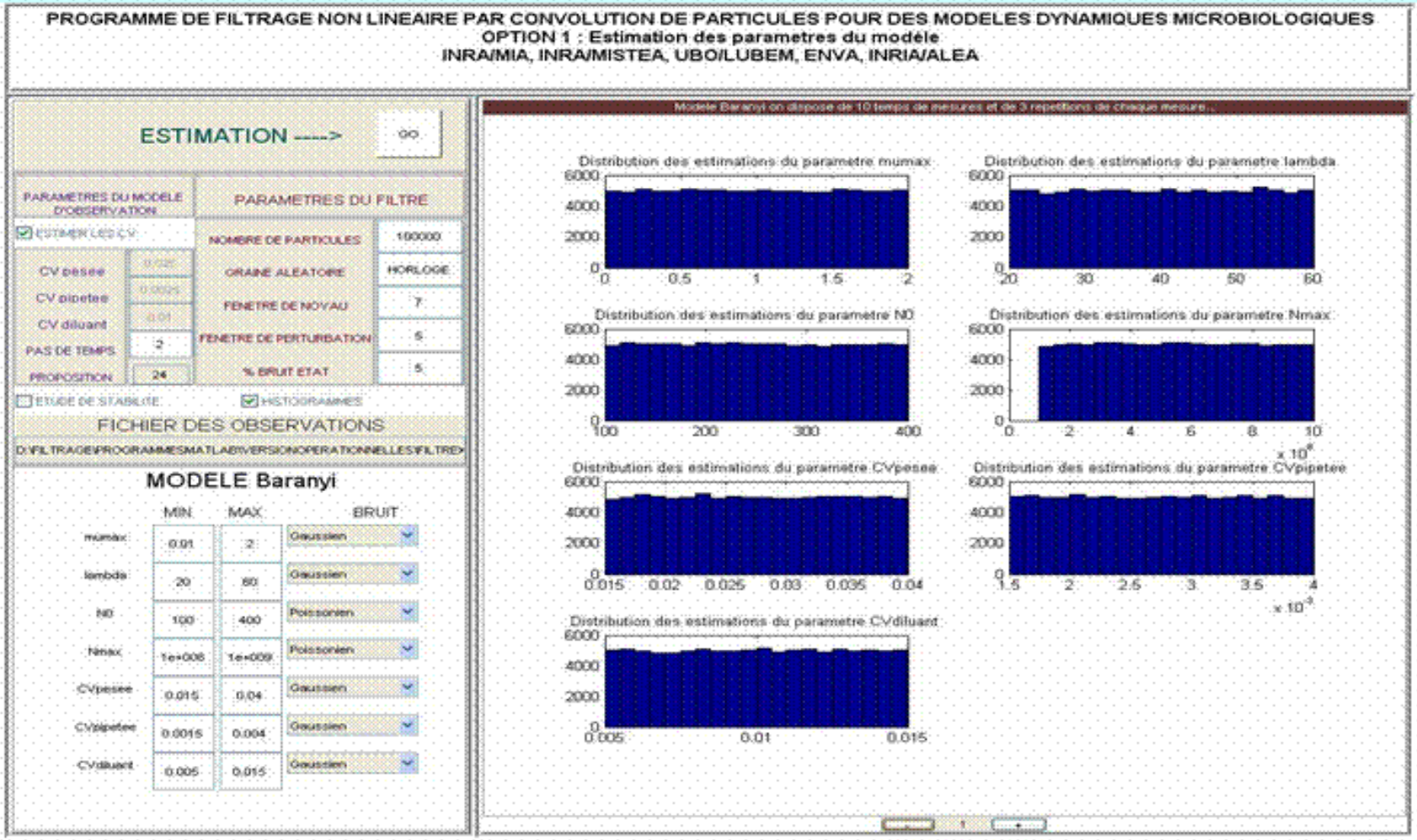
- Pas de discrétisation $\delta = 2$ heures, nombre de particules:
  $n = 10^4 \Longrightarrow$ environ 5 secondes de calcul sur PC.

Application au modèle de Baranyi-Roberts -10

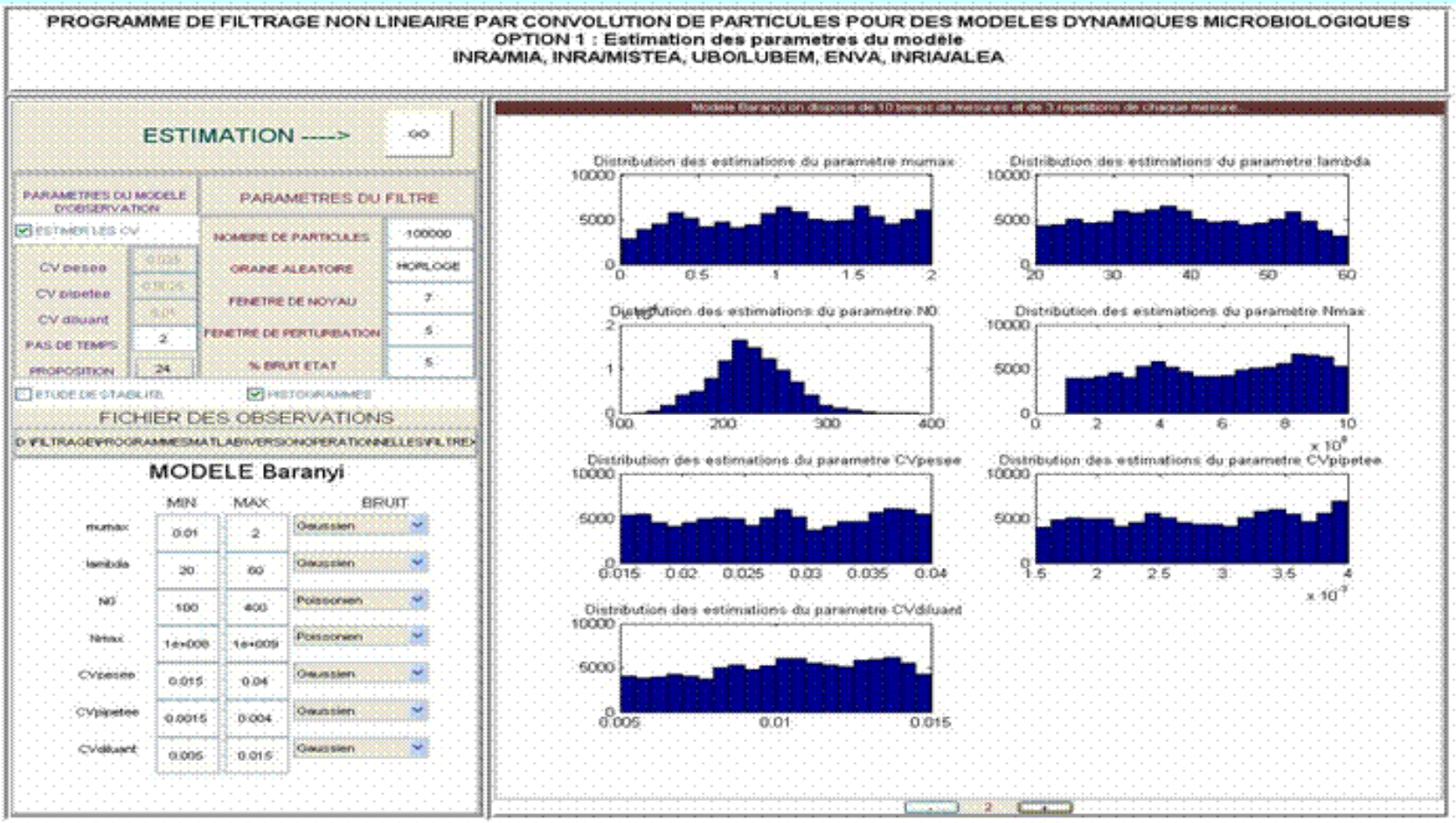
Le modèle estimé



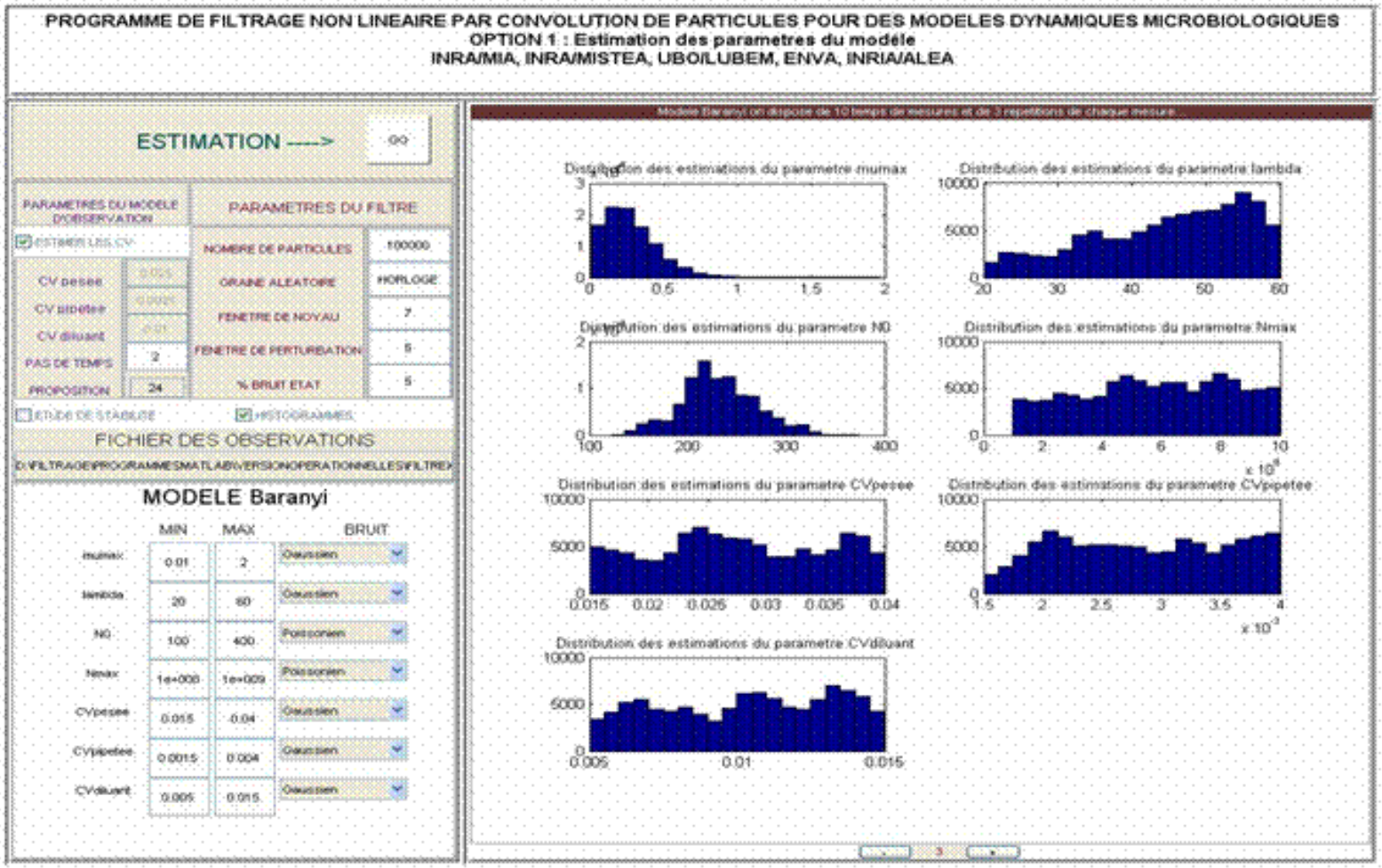
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 0



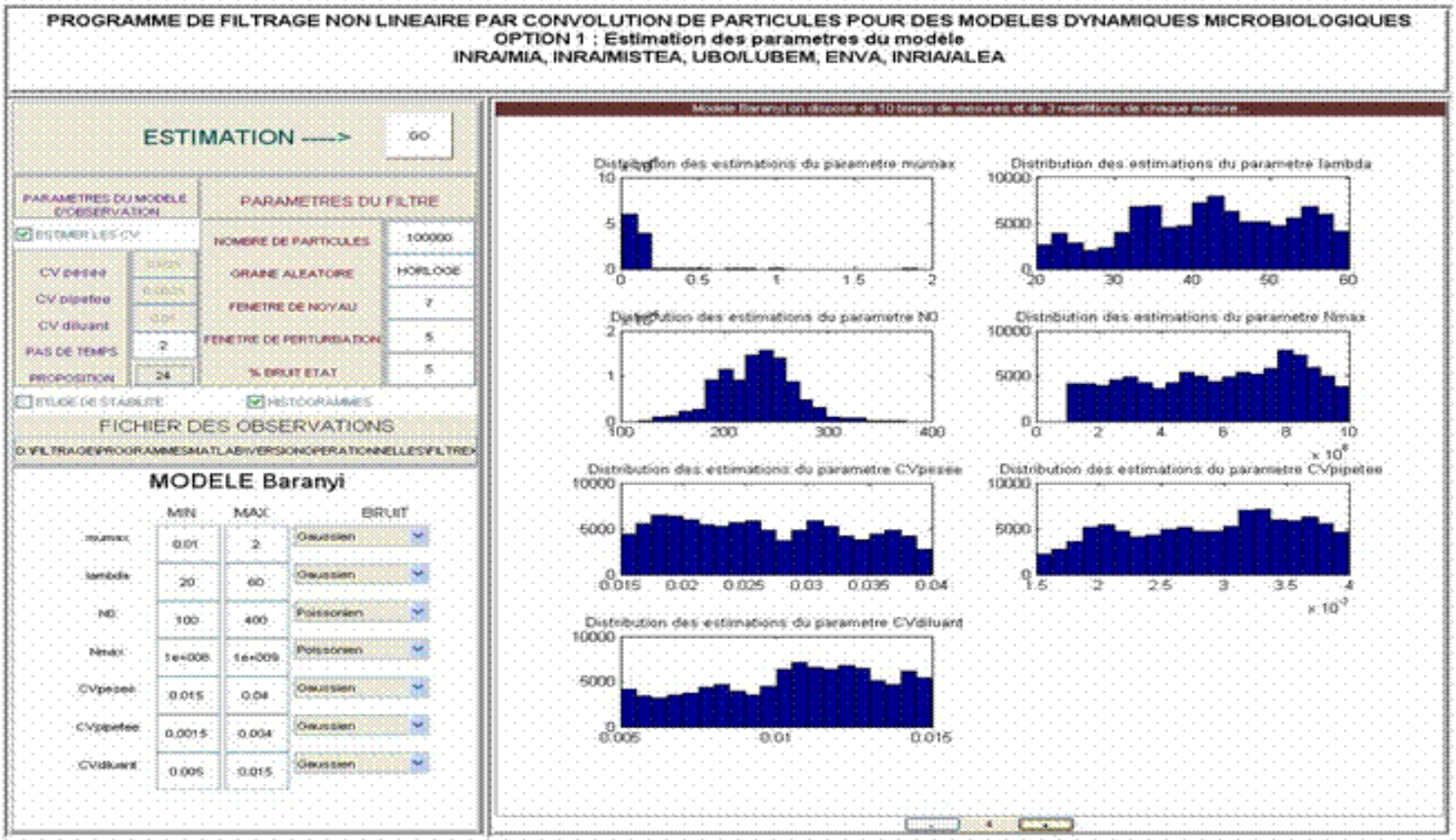
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 1



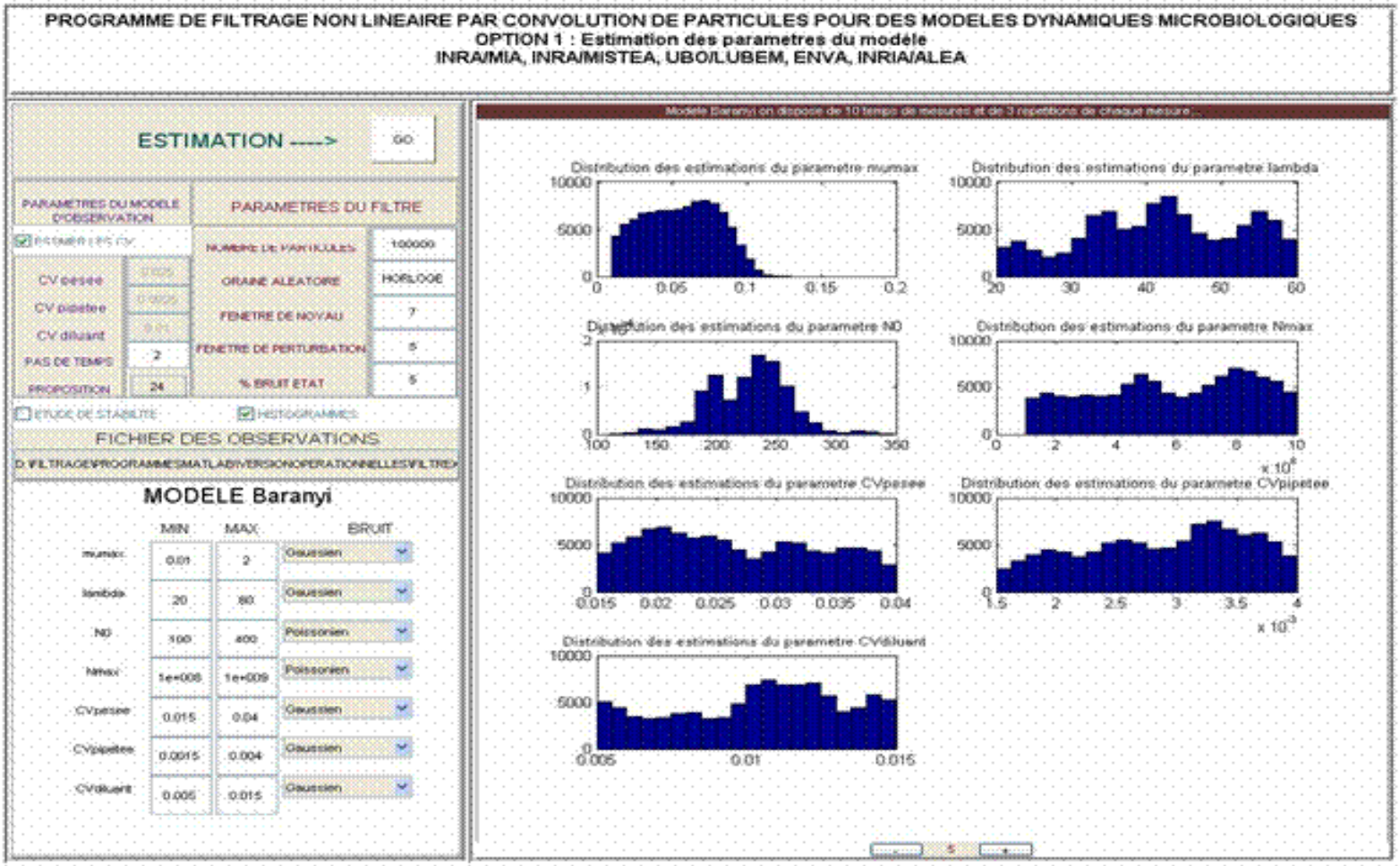
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 2



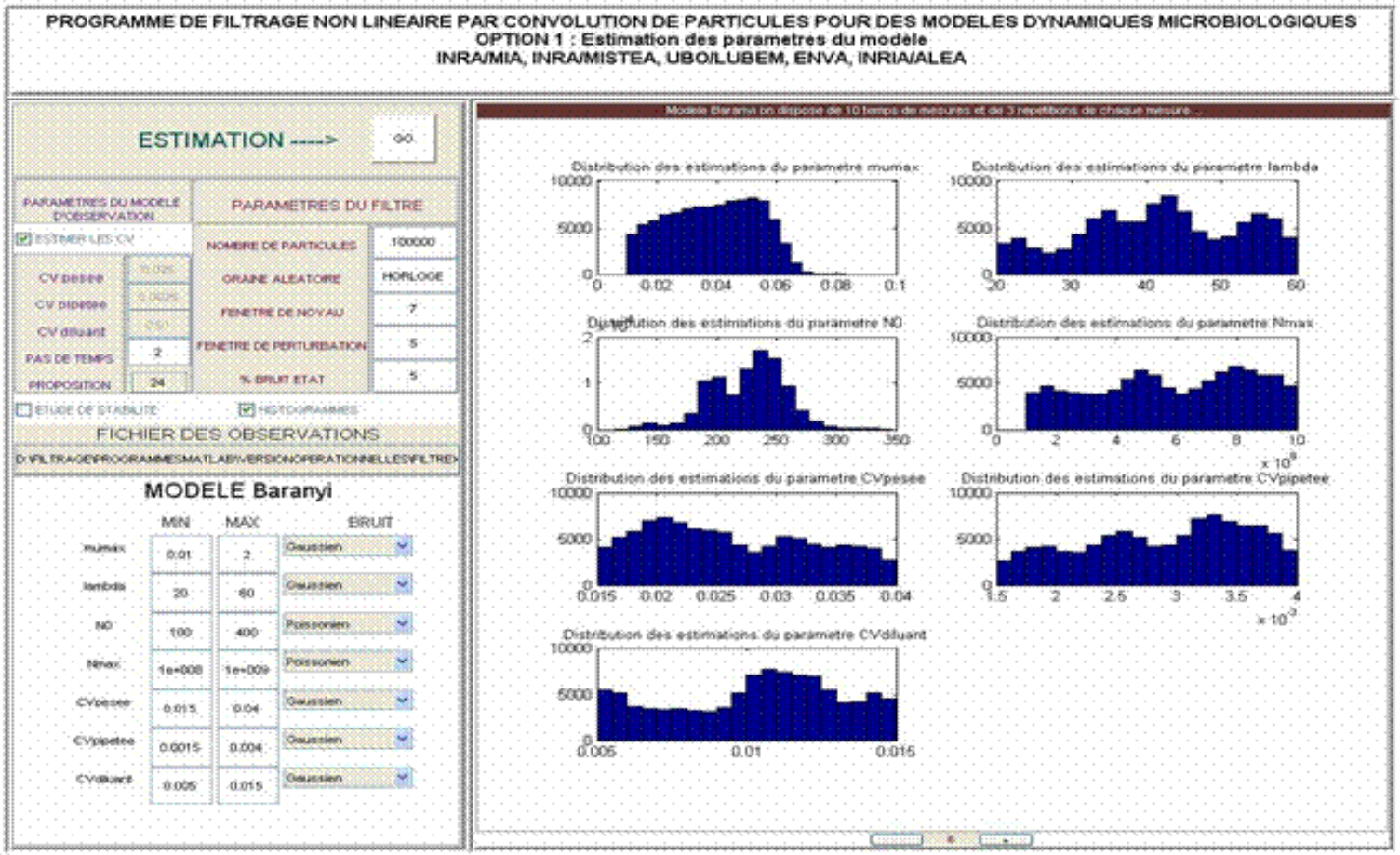
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 3



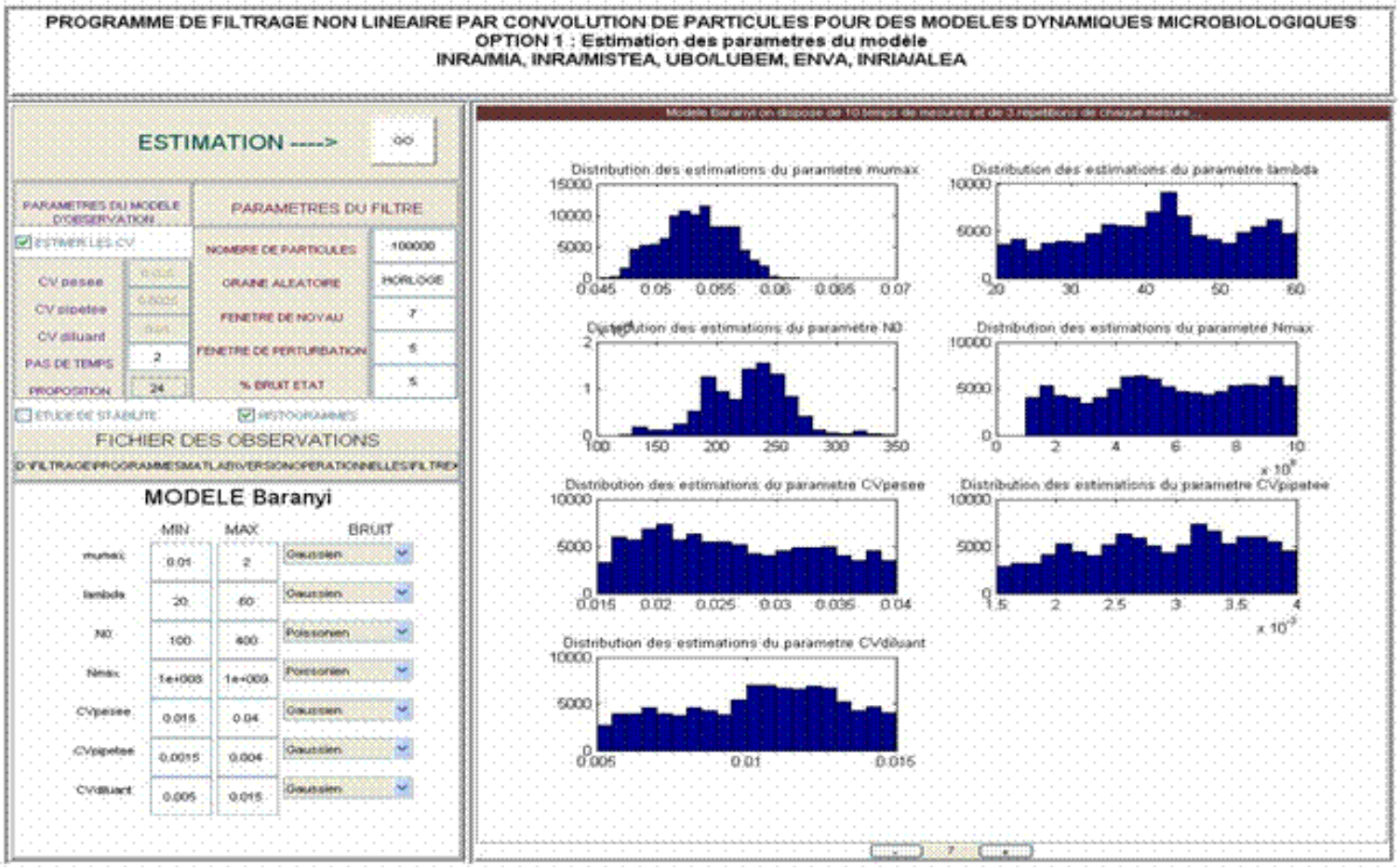
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 4



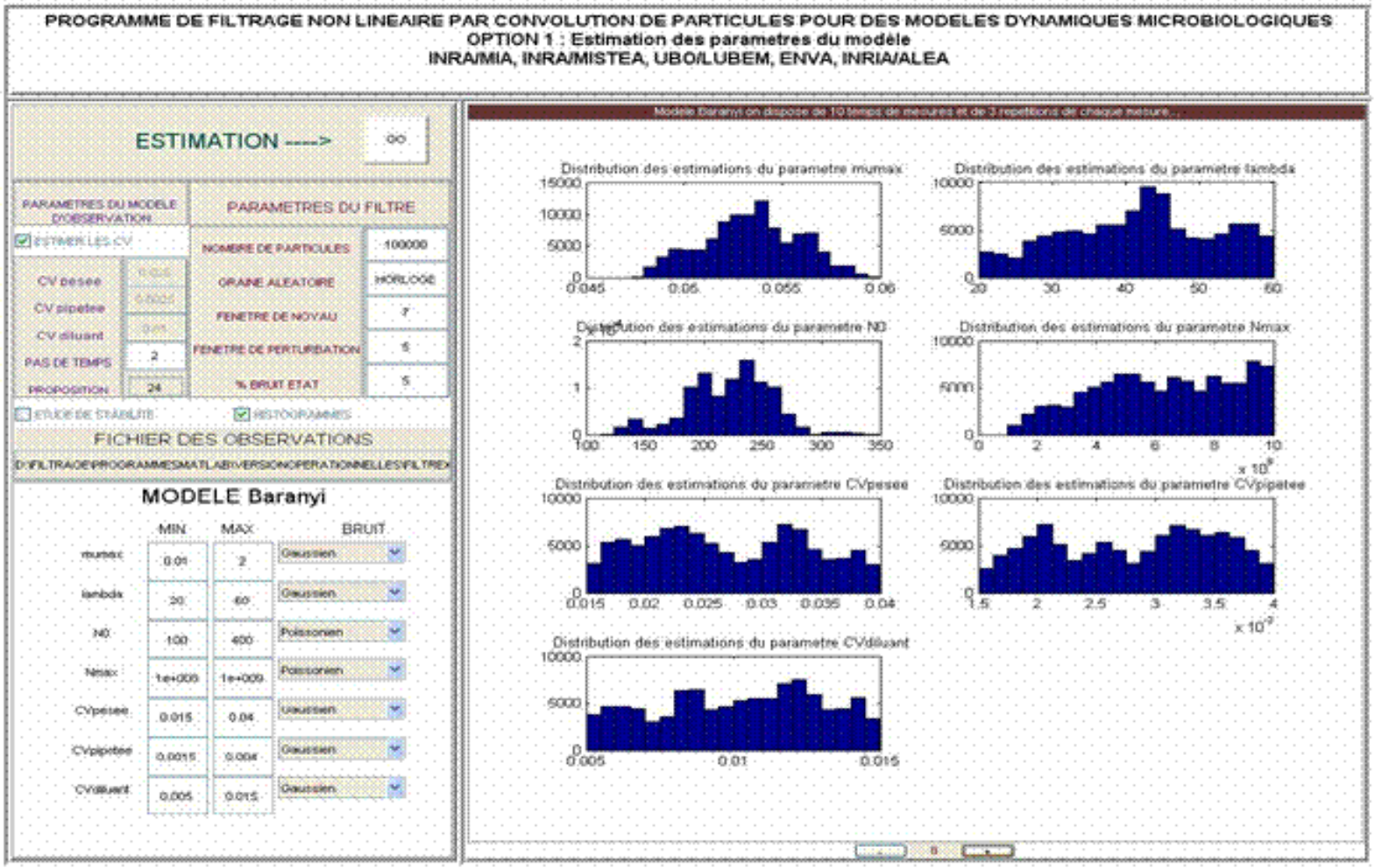
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 5



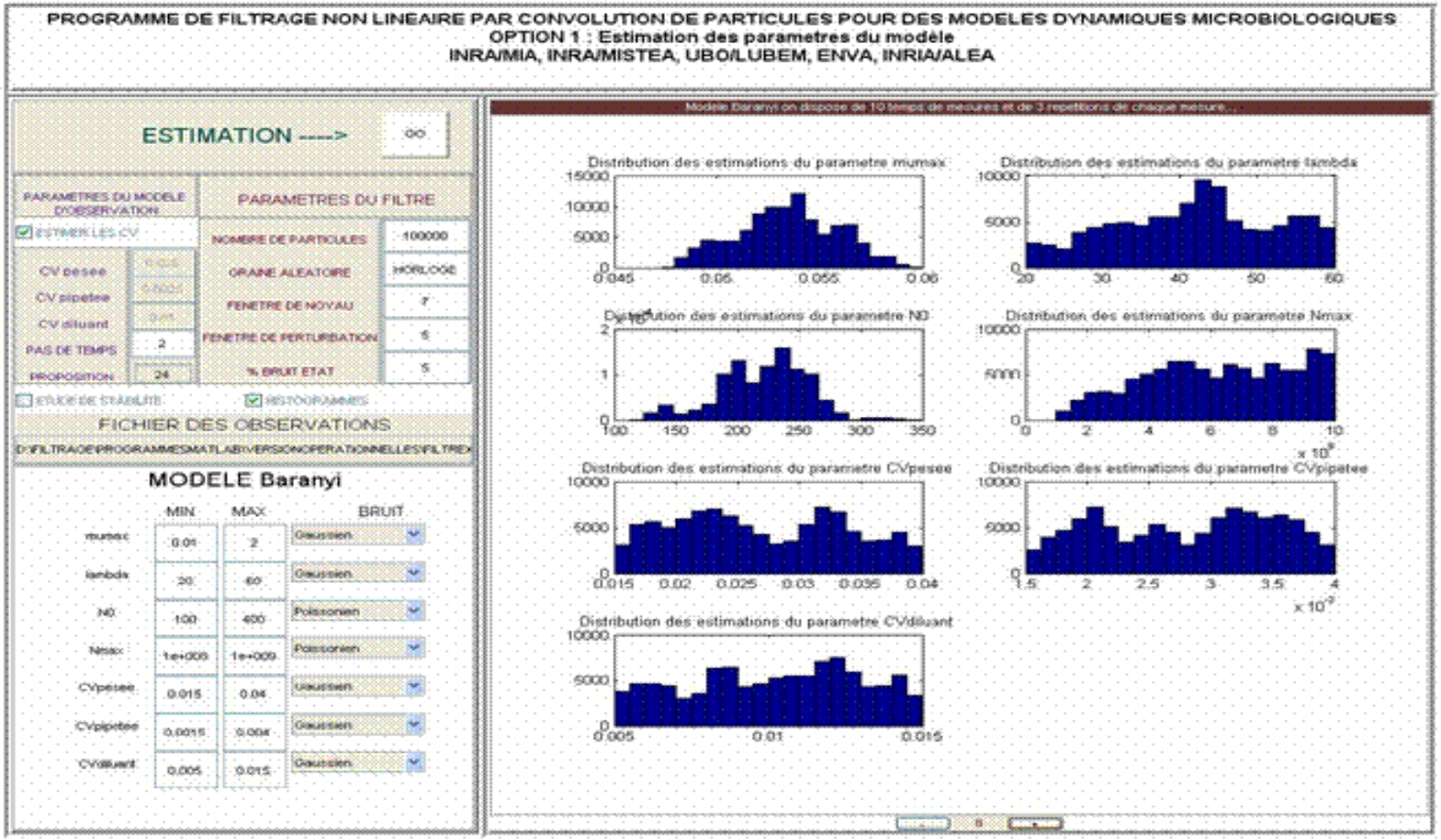
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 6



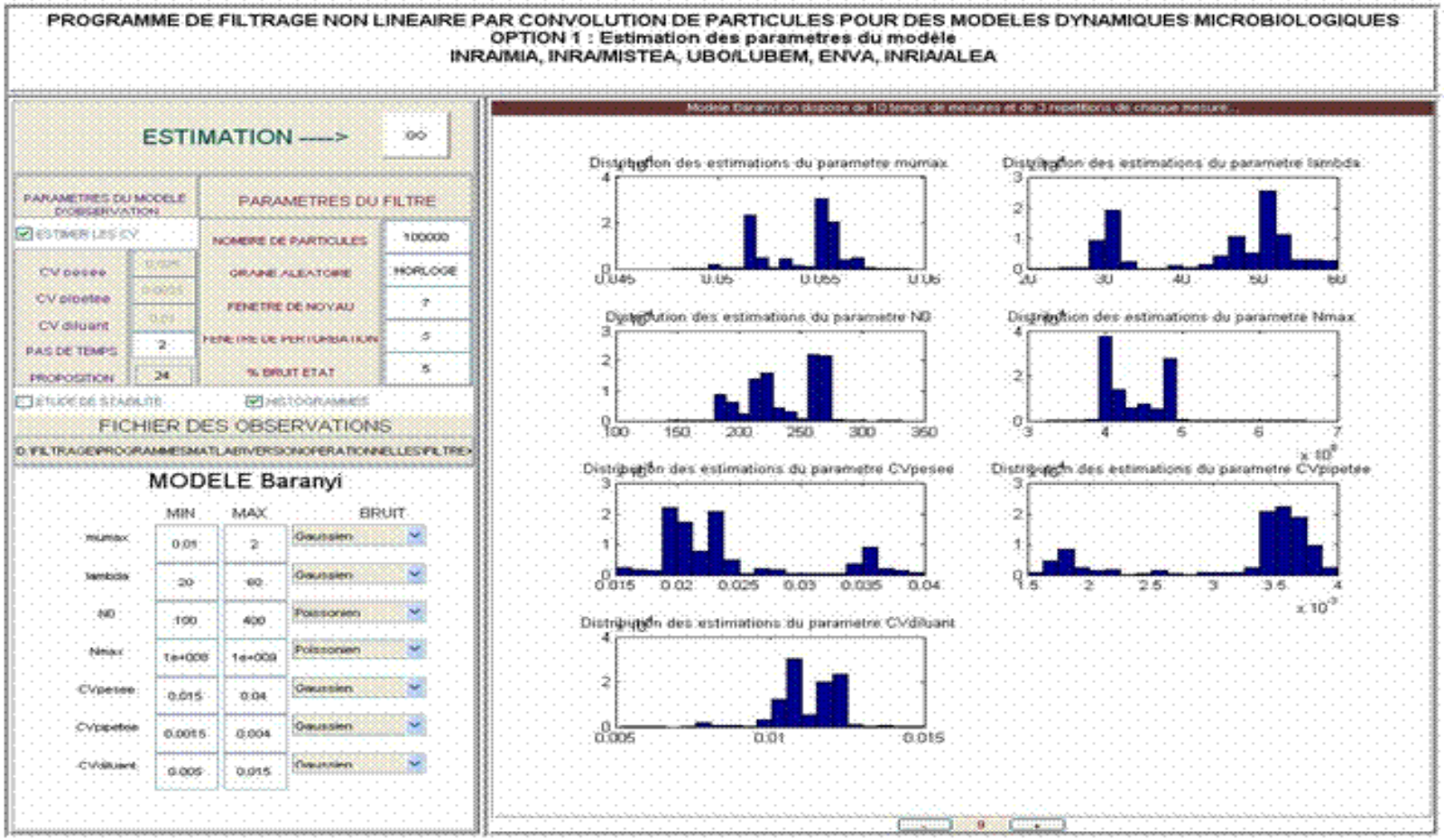
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 7



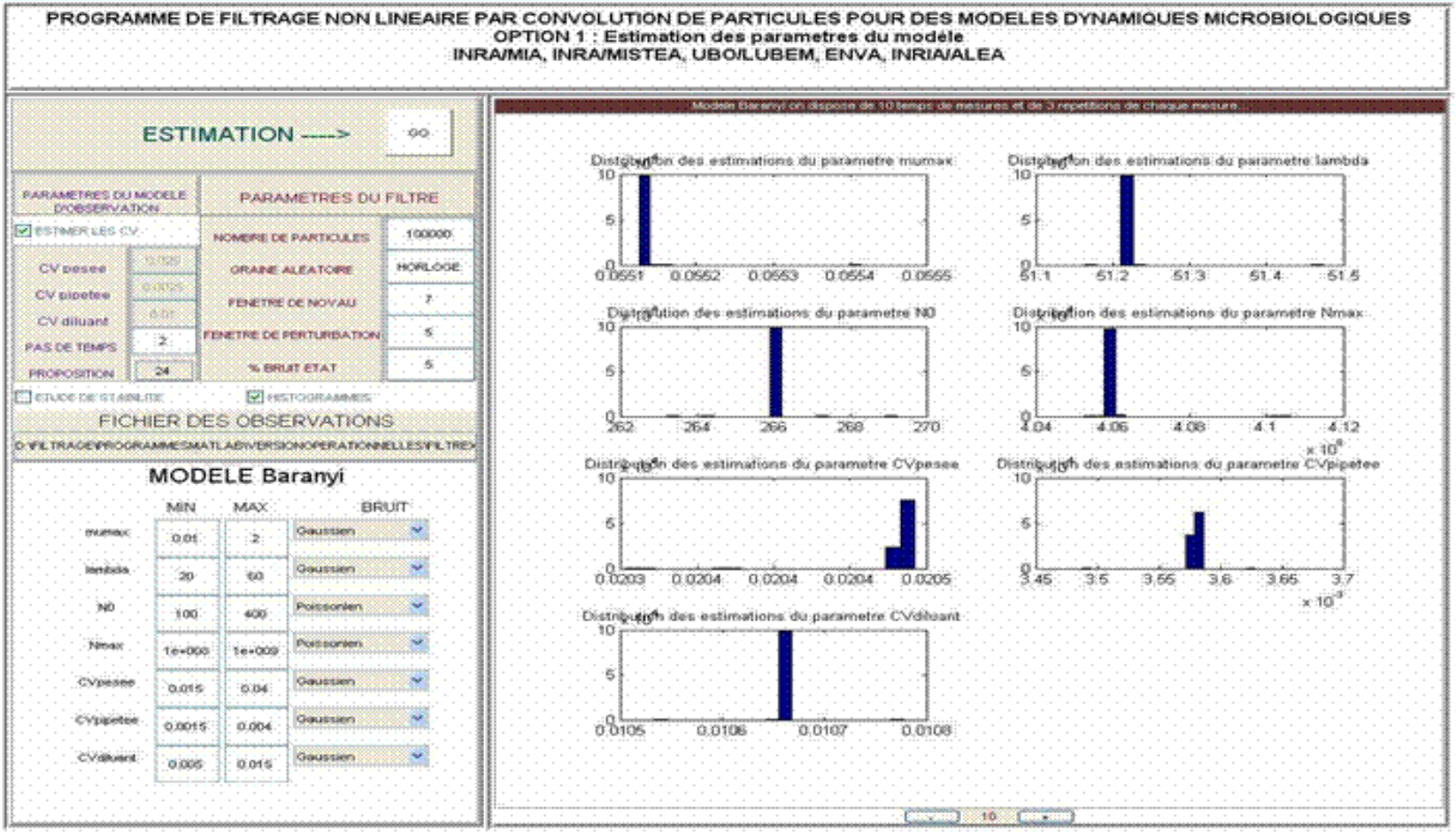
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 7



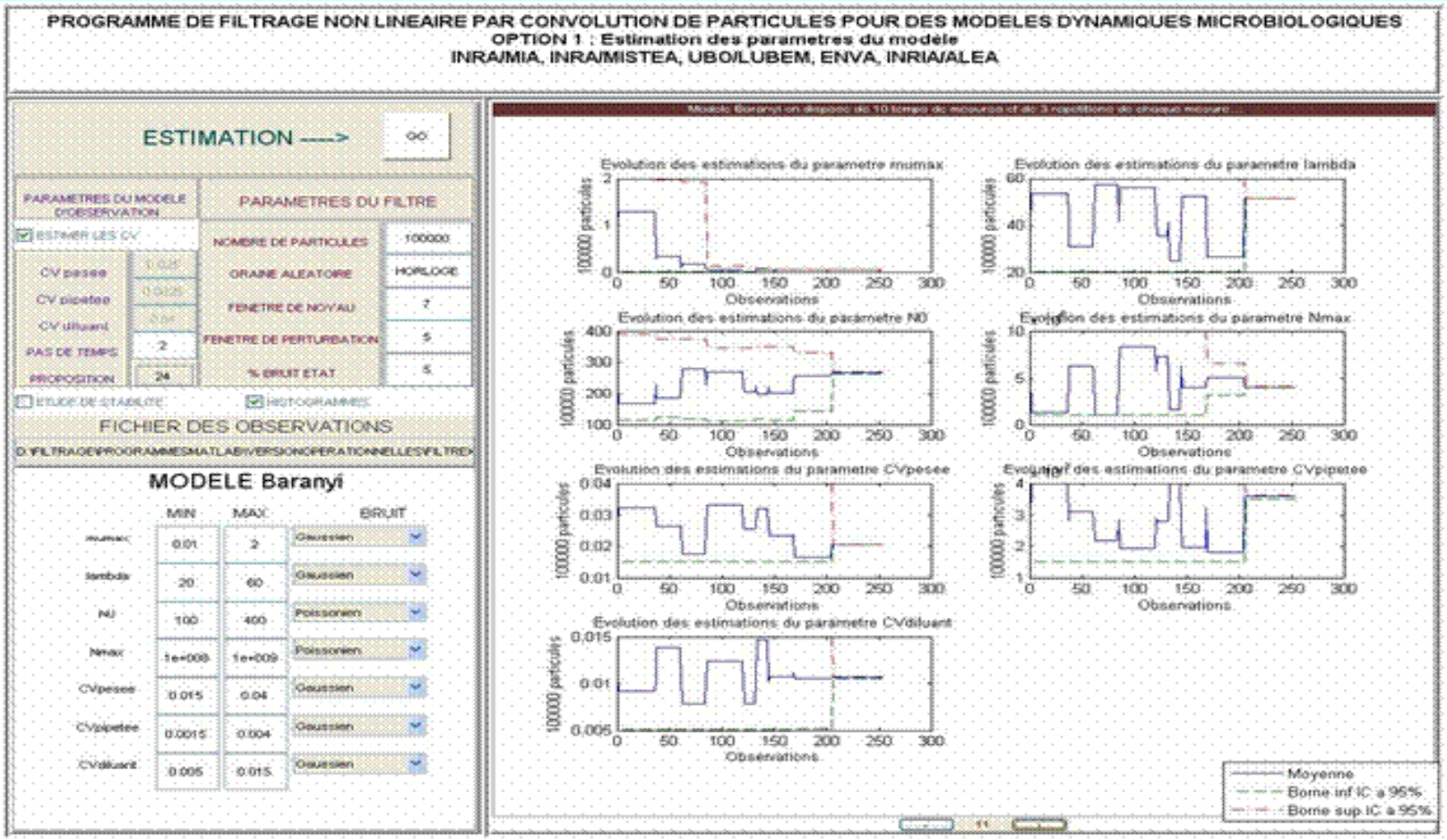
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 8



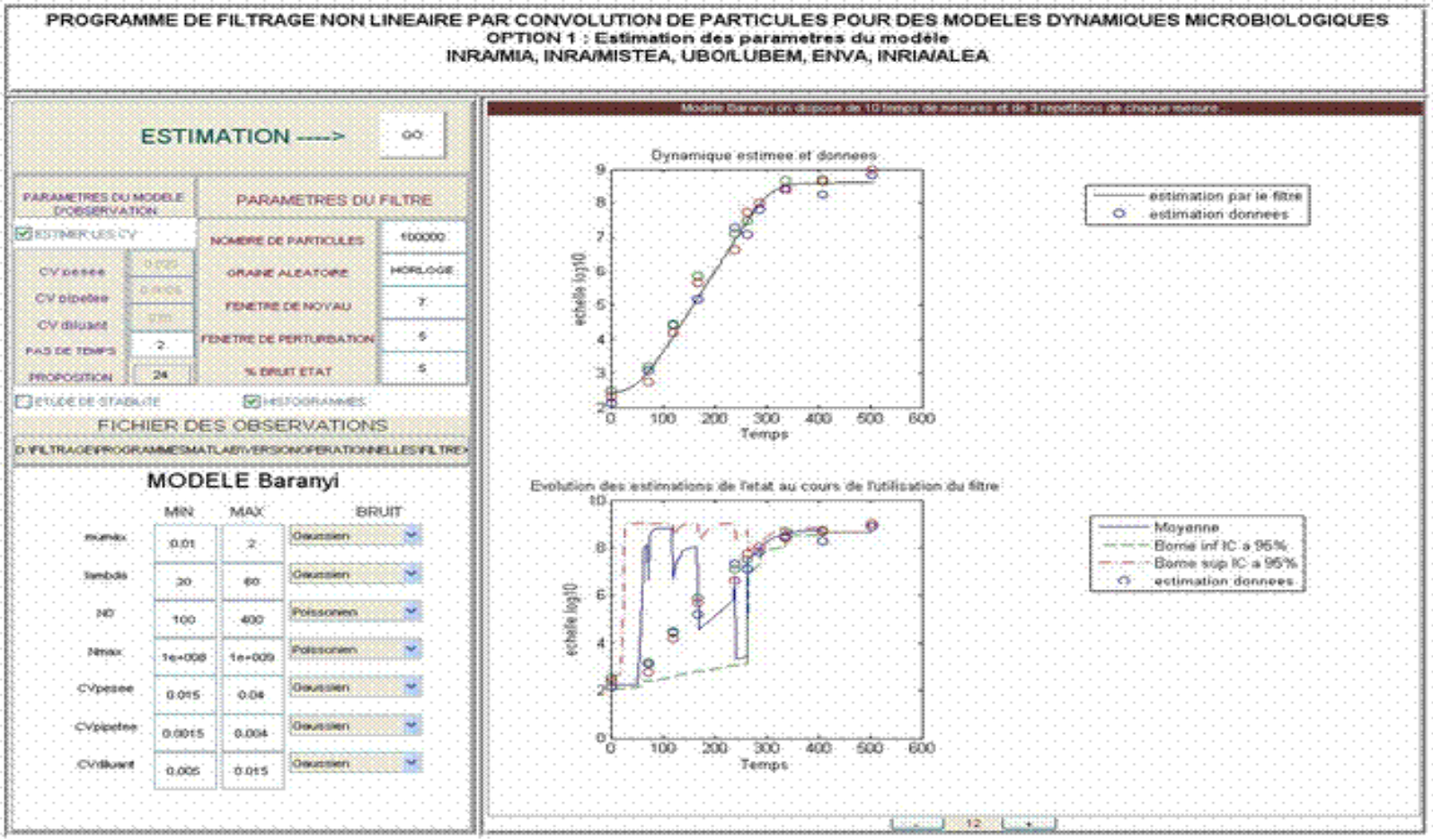
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 9



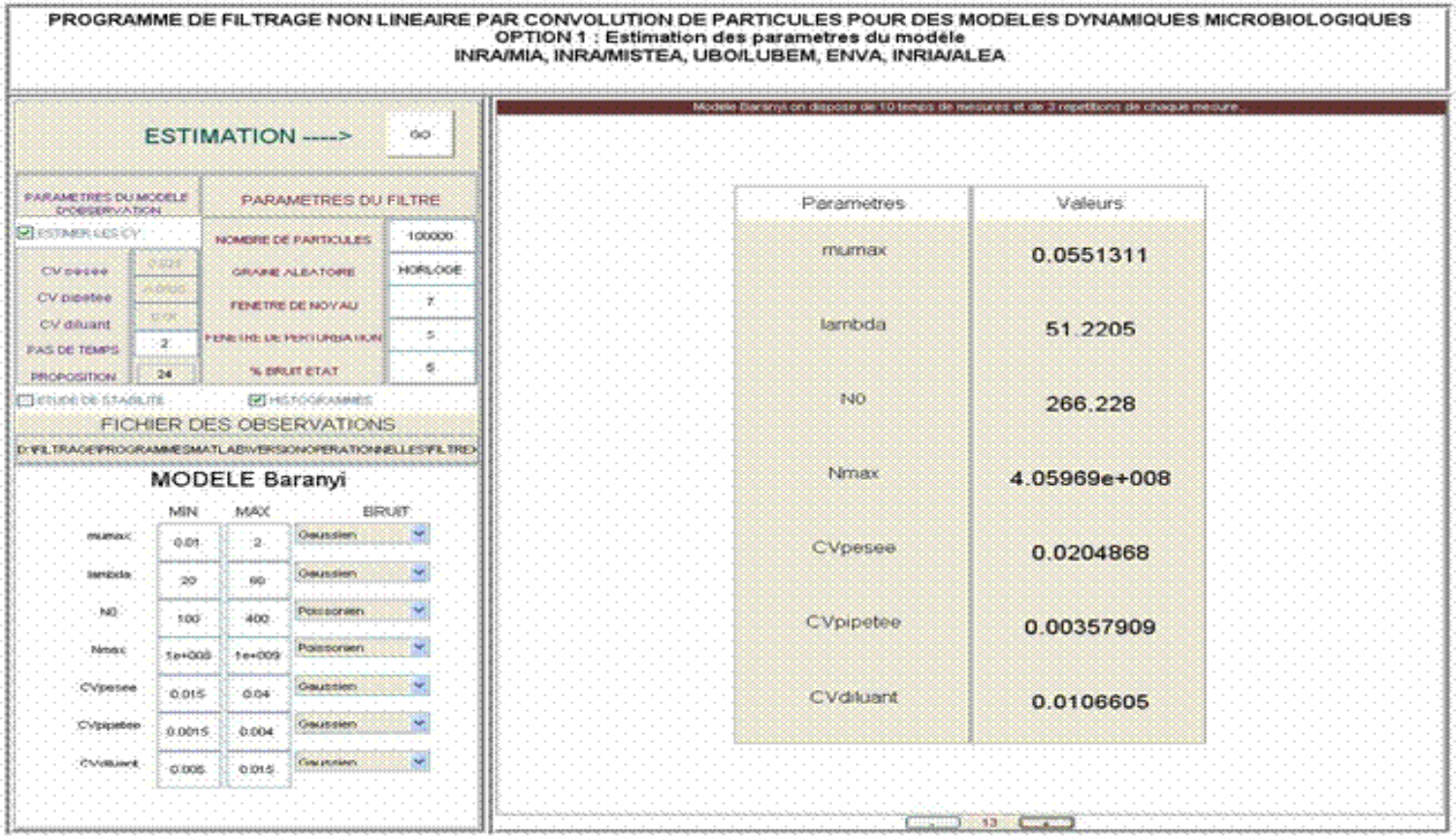
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 10



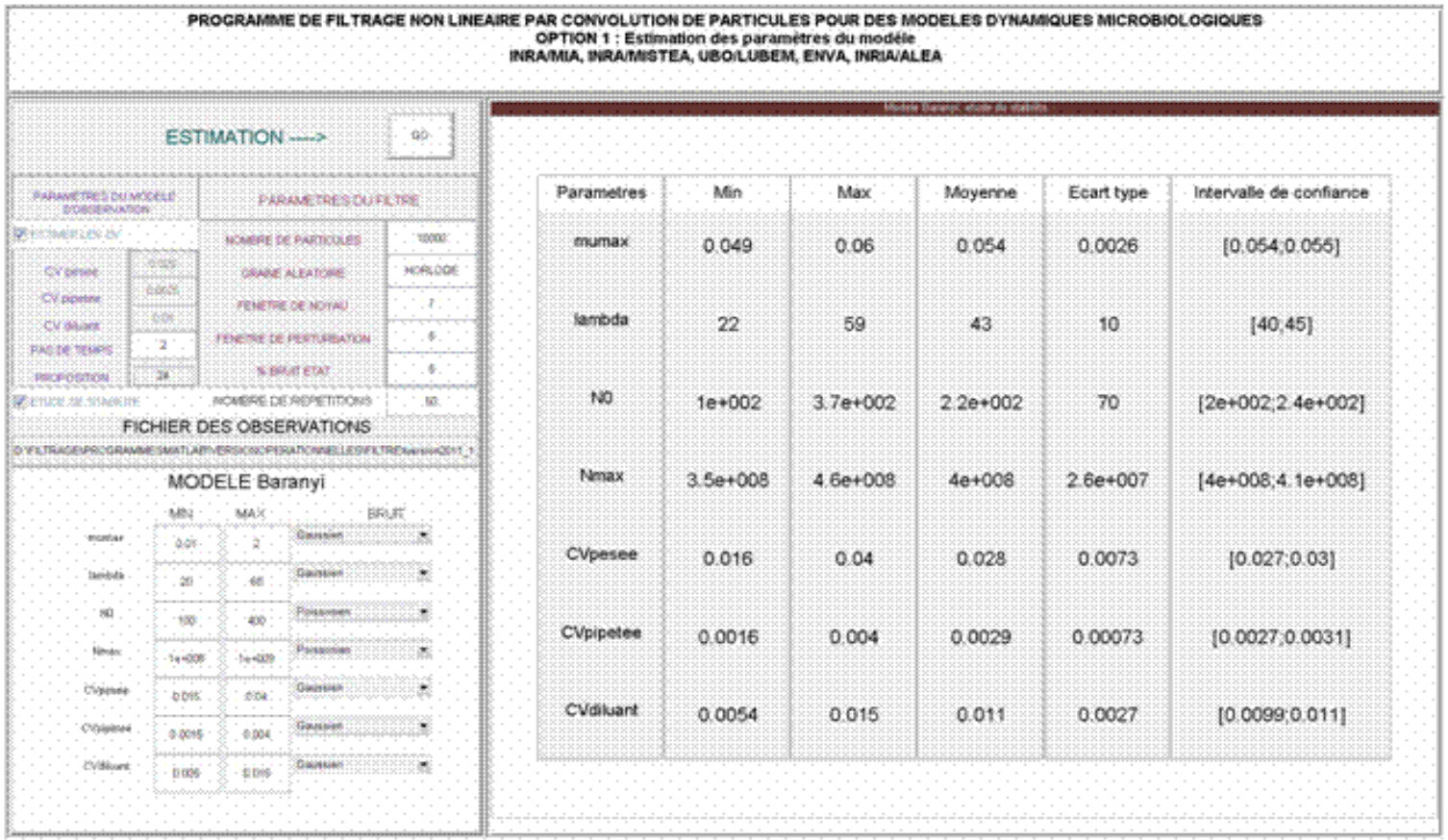
Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 11



Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995
(données Listeria monocytogenes)
Etape 12



Un exemple: Identification paramétrique du modèle primaire de croissance de Baranyi & Roberts,1995 (données Listeria monocytogenes)
Etape 13



Parametres	Min	Max	Moyenne	Ecart type	Intervalle de confiance
mumax	0.049	0.06	0.054	0.0026	[0.054;0.056]
lambda	22	59	43	10	[40;45]
N0	1e+002	3.7e+002	2.2e+002	70	[2e+002;2.4e+002]
Nmax	3.5e+008	4.6e+008	4e+008	2.6e+007	[4e+008;4.1e+008]
CVpesee	0.016	0.04	0.028	0.0073	[0.027;0.03]
CVpipetee	0.0016	0.004	0.0029	0.00073	[0.0027;0.0031]
CVdiluant	0.0054	0.015	0.011	0.0027	[0.0099;0.011]

Plan d'expériences -1

- Une expérience biologique =
  - un prèlèvement dans le milieu primaire à un temps donné,
  - suivi d'un processus de dilution-échantillonnage,
  - puis d'un comptage d'unités formant colonies sur une à cinq boites de Petri.
- Ces expériences biologiques sont coûteuses, longues, délicates, parfois dangereuses (parfois plus 20 expériences sont réalisées)
  $\Longrightarrow$ on souhaite en réaliser le **nombre minimum** mais qui conduisent toutefois à de **bonnes estimations finales**.
  $\Longrightarrow$ on va chercher **séquentiellement les temps optimaux** de prélèvement dans le milieu primaire où se fait la croissance microbienne.
- Un autre souhait est d'accélérer la convergence c-a-d l'accélération du rétrécissement des densités conditionnelles $p_t\left(\theta|y_0, ..., y_t\right)$.

Plan d'expériences -2

- **Une difficulté intrinsèque:**

La démarche des plans optimaux usuels n'est pas possible car on ne dispose pas de valeur a priori pour les paramètres, **seules des plages de variation sont postulables.**

- **Le critère choisi:**
  Déterminer le temps optimal prochain comme le temps auquel **un indice de sensibilité paramétrique total est maximum**.
  $\Longrightarrow$ Choix d'un type d'indice de sensibilité total, appelé $T_d SIVIP$, basé sur **les VIP de la régression PLS** (Tenenhaus, 1998), en se basant sur:
  - l'ajustement d'un modèle polynomial (de degré d, avec interactions) de N_t en fonction des p paramètres pour chaque t, depuis le t courant jusqu'au $t_{\max}$ $\Longrightarrow$ établissement d'une courbe de sensibilité en fonction du temps, pour chaque paramètre,
  - la mise à jour de ces courbes depuis chaque temps optimal.

Plan d'expériences -3

Argumentation

- Etant donnée une durée de filtrage $[0, t_{\max}]$ et un nombre maximum H de temps d'observation, il est alors possible de déterminer la suite des temps d'observation $0 < t_1 < t_2 < \ldots < t_H \leq t_{\max}$ qui optimise l'identification paramétrique en un certain sens, par exemple en fournissant les fonctions de densité conditionnelles les plus rétrécies.
- D'un point de vue de la **théorie de l'information** ce problème de plan optimal peut être reformulé en termes de **sensibilité de la sortie** N_t.
- Dans le cas du filtrage "on-line" qui nous intéresse ici on préfèrera une démarche séquentielle sur $[0, t_{\max}]$, avec un horizon glissant (Gauchi & Vila, 2012)

Plan d'expériences -4

Définition des *VIP*

- Soit la redondance pour une composante PLS t_h:

$$Rd(Y; t_h) = cor^2(Y, t_h)$$

- Soit la redondance pour H composantes (H obtenu par validation croisée):

$$Rd(Y; t_1, \ldots, t_H) = \sum_{h=1}^{H} cor^2(Y, t_h)$$

- Le VIP (Variable Importance in the Projection) d'une entrée (ici une des p composantes de θ) se définit par:

$$VIP_{Hj} = \left[\frac{p}{Rd(Y; t_1, \ldots, t_H)} \sum_{h=1}^{H} Rd(Y; t_h) w_{hj}^2 \right]^{1/2}$$

où les w_{hj} sont les composantes du vecteur propre w_h de la régression PLS à l'étape h.

$\Longrightarrow$ on remarque que $\sum_{j=1}^{p} VIP_{Hj}^2 = p$.

Plan d'expériences -5

Définition des indices $T_d SIVIP$

- Soit l'indice de sensibilité **individuel** défini pour chaque monome du polynome complet de degré d (P monomes pour p paramétres)

$$ISIVIP_k = VIP_k^2 / P \quad , \quad k = 1, ..., P$$

- L'indice de sensibilité **total** pour l'entrée θ_j pour le polynome de degré d est défini par:

$$T_d SIVIP_j = \sum_{u=1}^{J} ISIVIP_{\Omega_{ju}} \quad , \quad j = 1, ..., p$$

où:

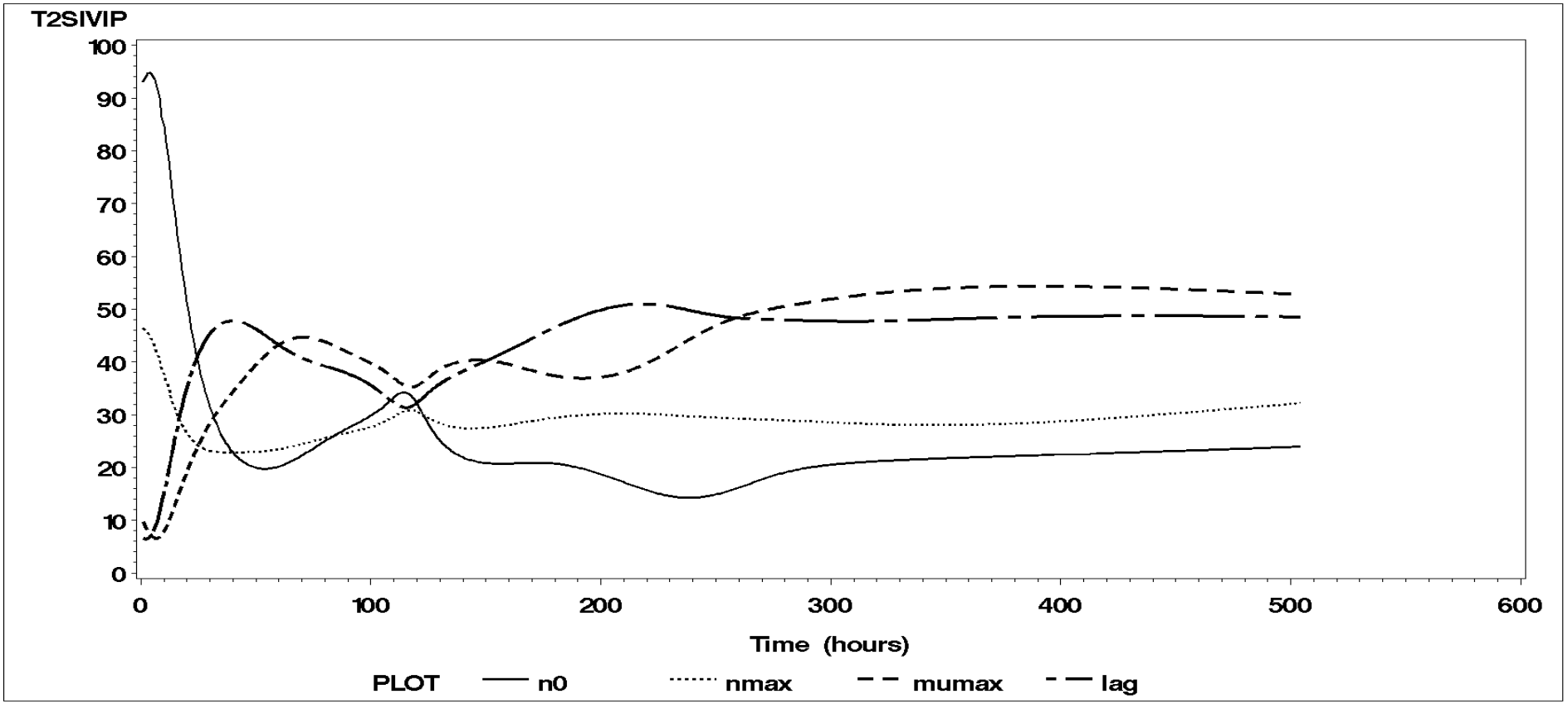
  - Ω_{ju} est le u^{ième} ensemble d'indices où un indice j apparaît, J étant le nombre total de ces Ω_{ju},
  - $ISIVIP_{\Omega_{ju}}$ est l'indice de sensibilité individuel correspondant au monome

- Par exemple, avec 3 entrées, X_1, X_2, X_3, and $d = 2$, on a: $T_2 SIVIP(X_1) = ISIVIP(X_1) + ISIVIP(X_1^2) + ISIVIP(X_1 X_2) + ISIVIP(X_1 X_3)$; on ramène enfin les $T_d SIVIP_j$ entre 0 et 100%.

Plan d'expériences -6

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t = 0$.

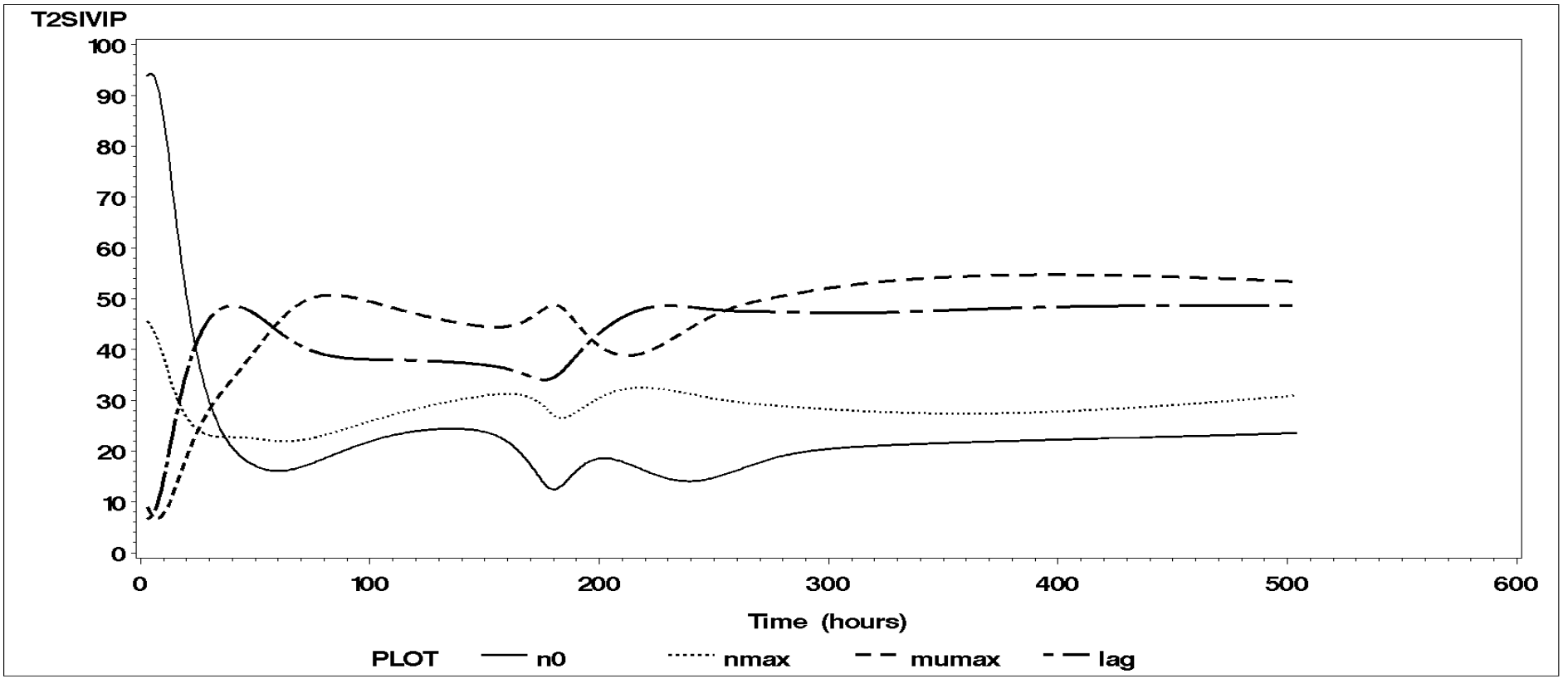
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 2$, sur la courbe N_0.**



Plan d'expériences -7

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 2$.

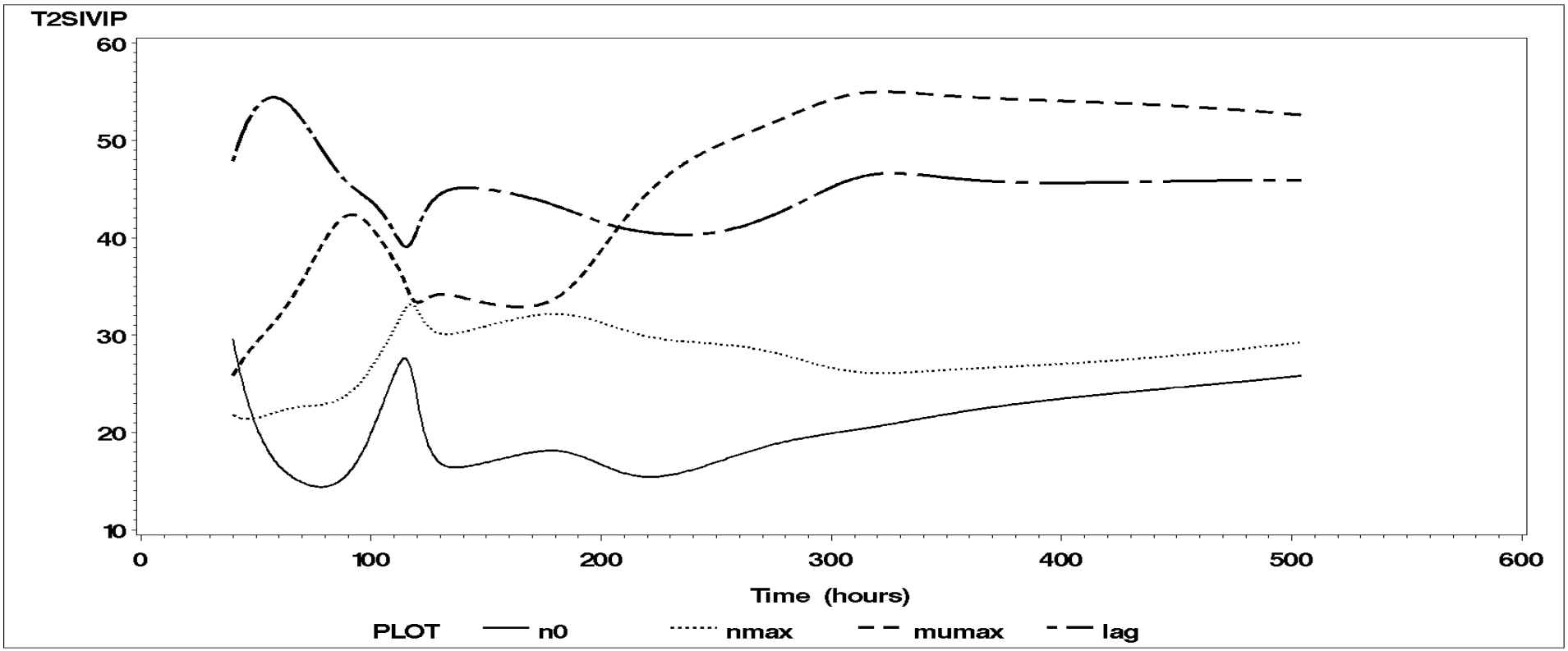
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 39$, sur la courbe λ.**



Plan d'expériences -8

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 39$.

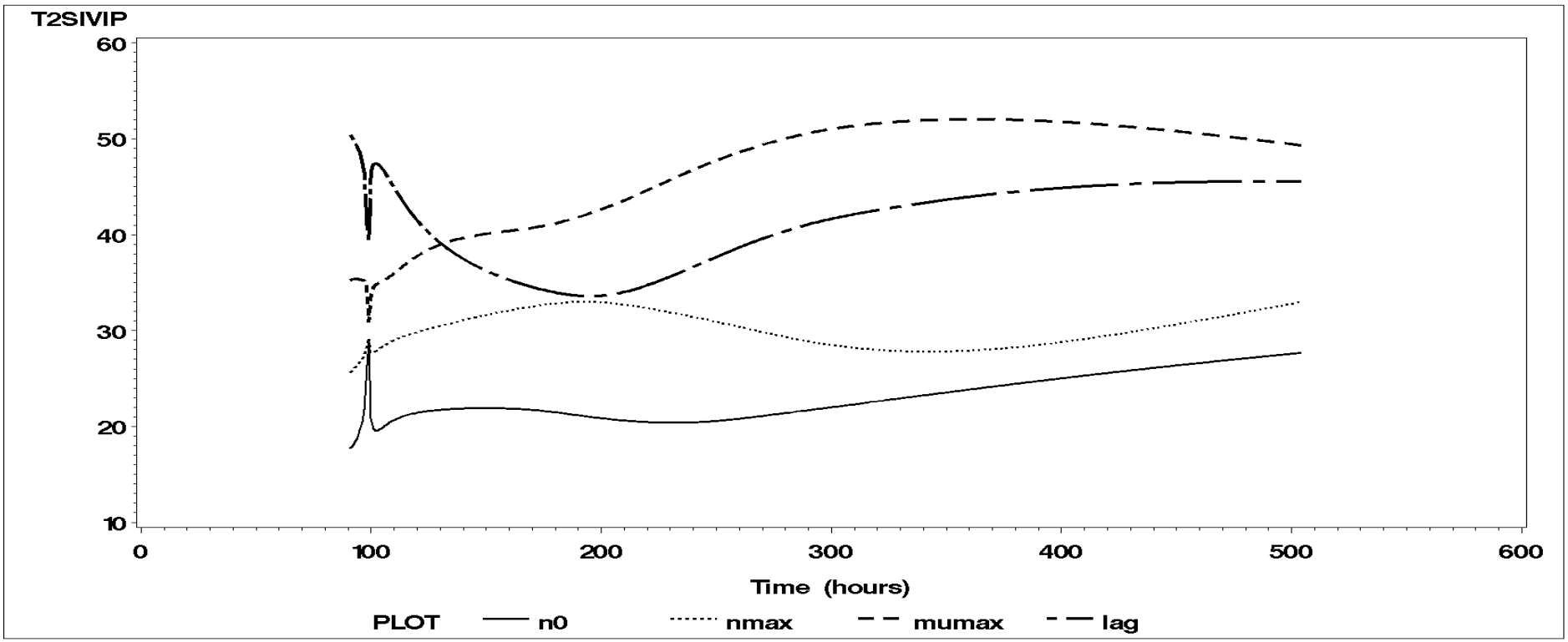
$\Longrightarrow$ **Le premier prochain maximum (validé) est $t^* = 90$, sur la courbe μ_{max}.**



Plan d'expériences -9

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 90$.

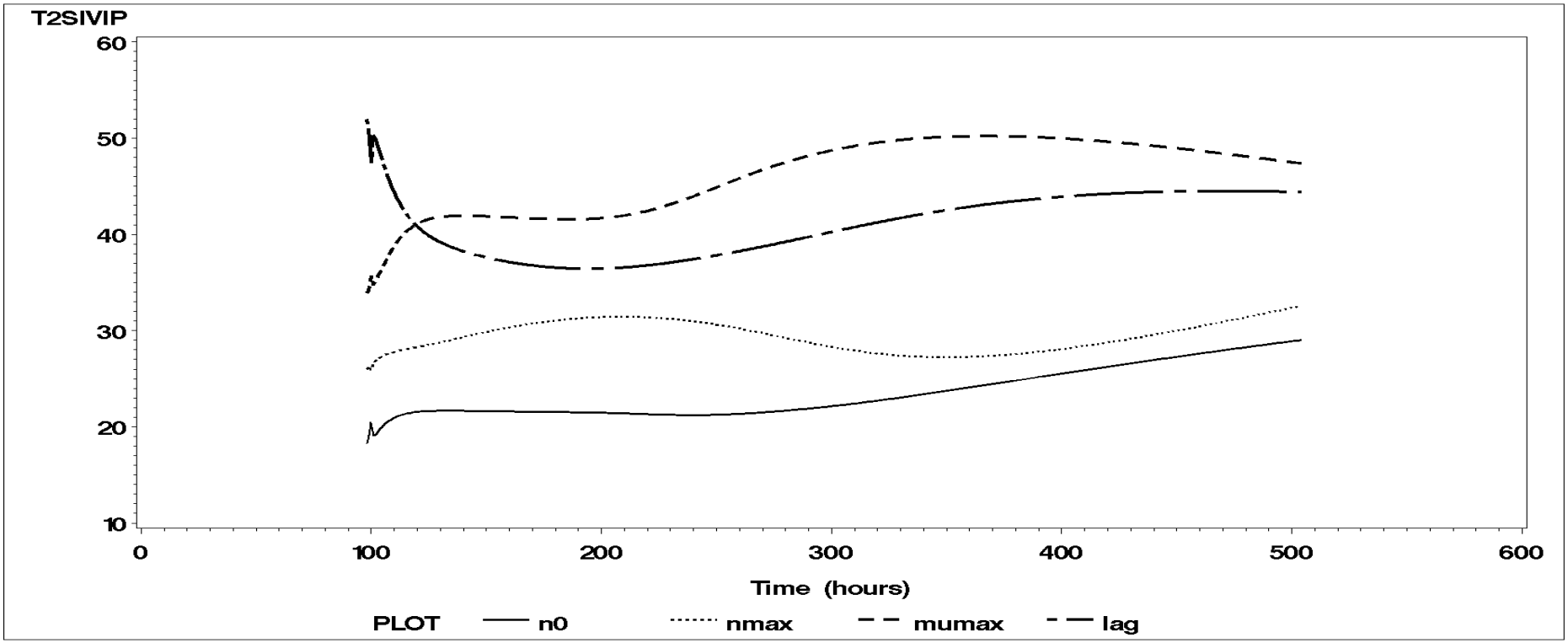
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 97$, sur la courbe μ_{max}.**



Plan d'expériences -10

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 97$.

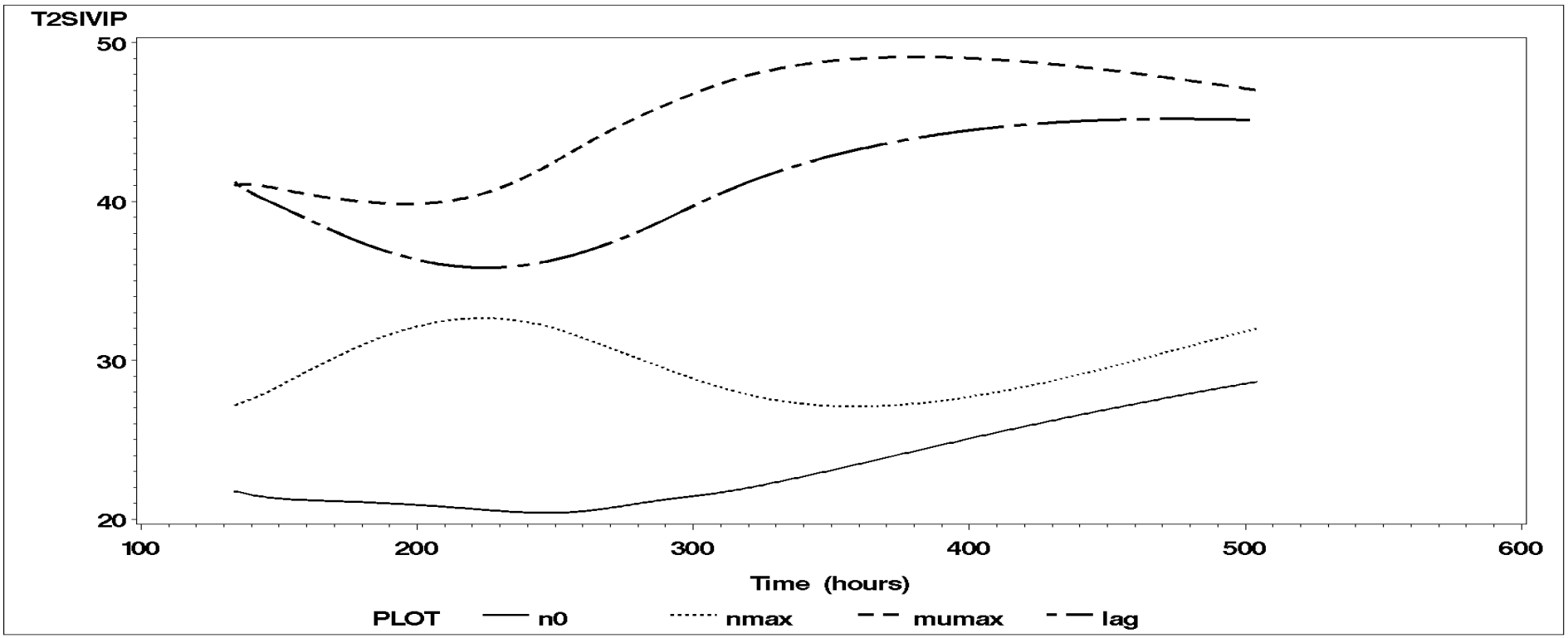
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 133$, sur la courbe μ_{max}.**



Plan d'expériences -11

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 133$.

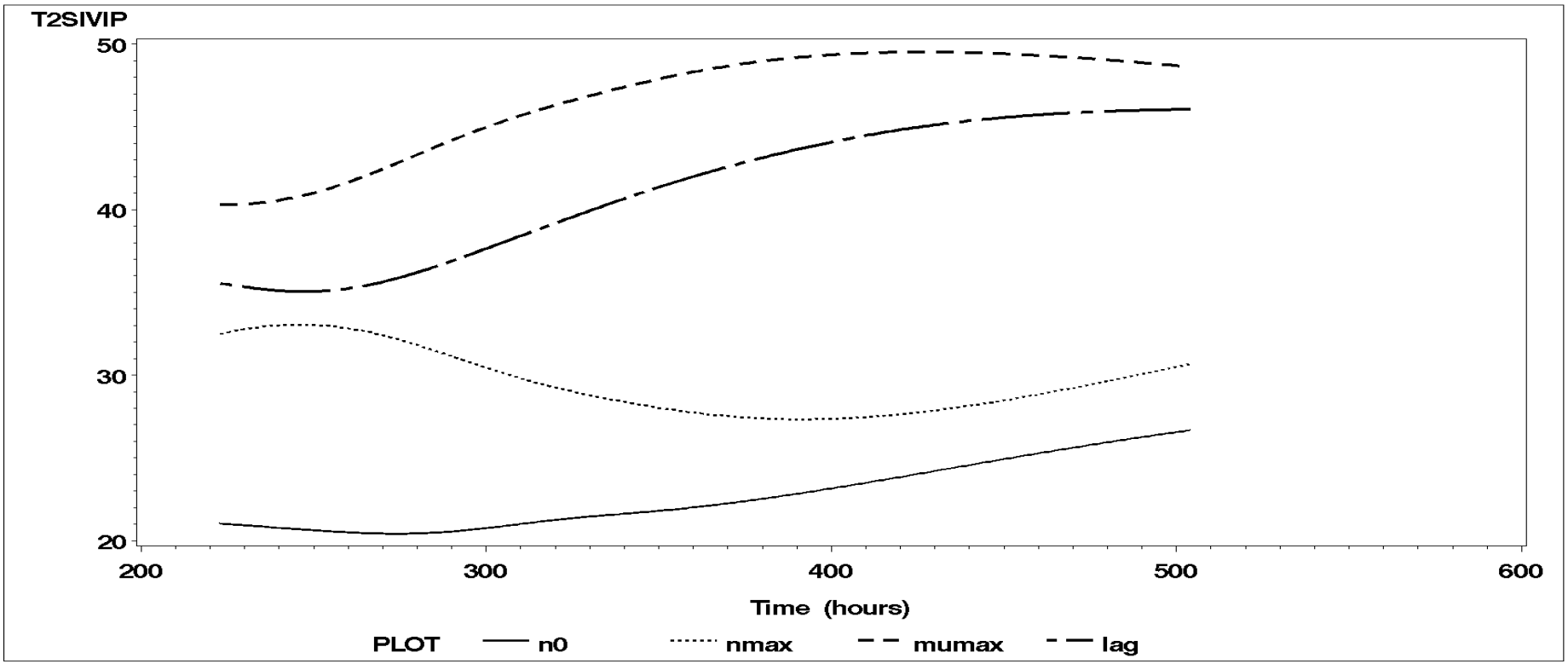
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 222$, sur la courbe $N_{\max}$.**



Plan d'expériences -12

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 222$.

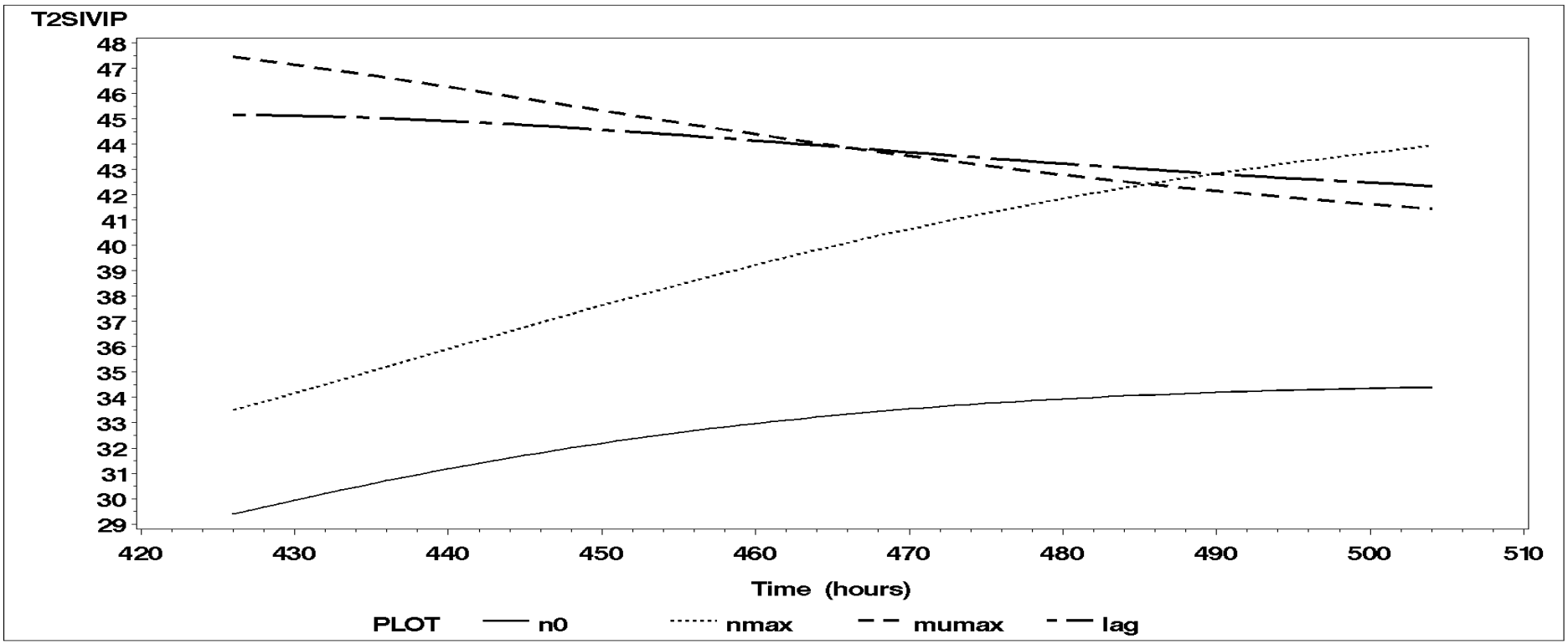
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 425$, sur la courbe μ_{max}.**



Plan d'expériences -13

Les 4 courbes T_2SIVIP des 4 paramètres du modèle de Baranyi-Roberts calculées à partir de $t^* = 425$.

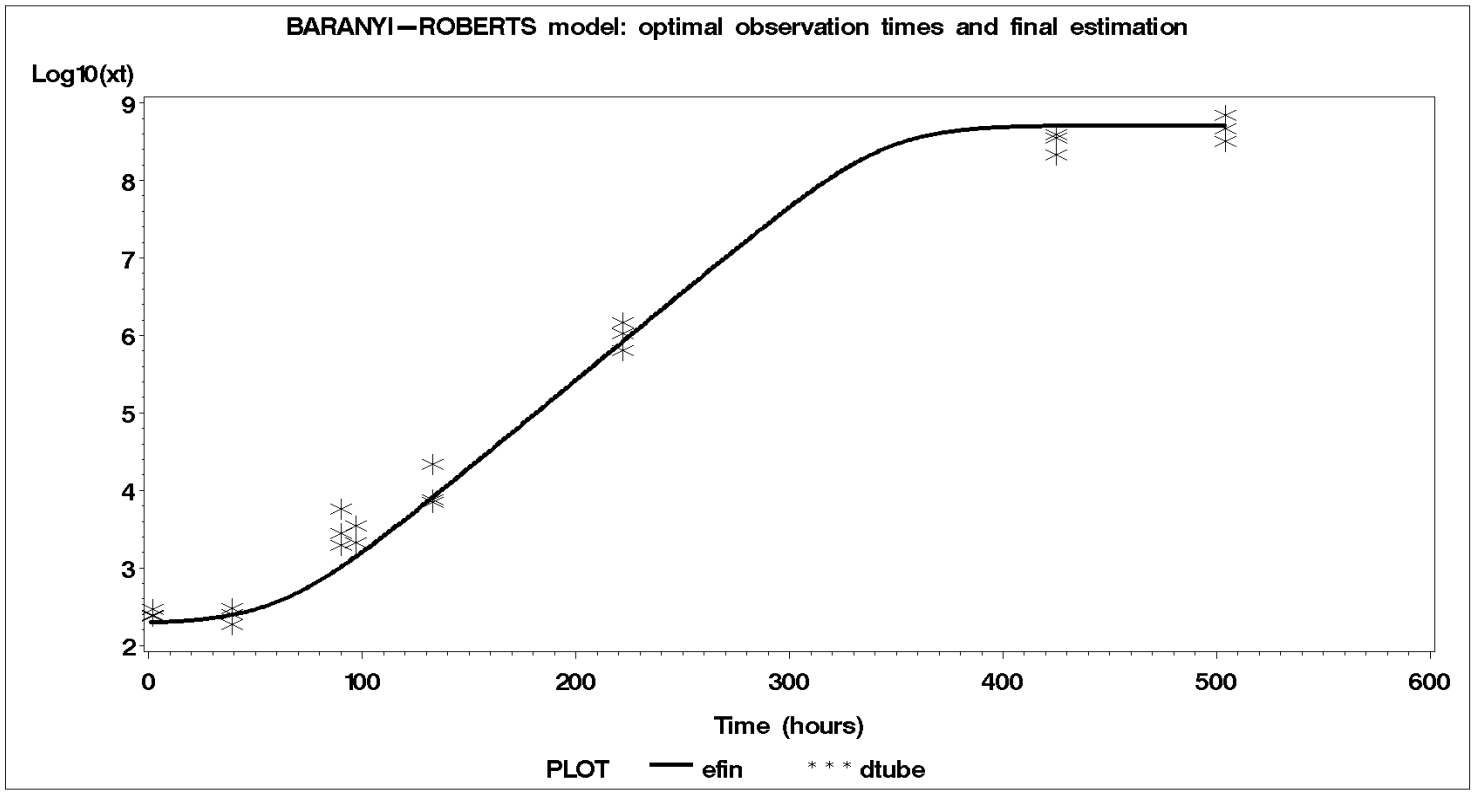
$\Longrightarrow$ **Le premier prochain maximum est $t^* = 504$, sur la courbe $N_{\max}$.**



Plan pour le modèle de Baranyi-Roberts -14

Les temps optimaux :

$$t^* = (2; 39; 90; 97; 133; 222; 425; 504)$$



Plan pour le modèle de Baranyi-Roberts -15

- On observe une accumulation des temps optimaux dans la première partie du processus de filtrage, zone la plus sensible par rapport aux paramètres N_0, λ, μ_{max}.
- Le temps optimal maxi est judicieusement placé à $t_{\max}$ ce qui semble cohérent avec l'estimation du paramètre N_{max}.

Plan pour le modèle de Baranyi-Roberts -16

Estimations "optimales" (avec $n = 10^5$ particules on obtient):

Paramètre	N_0	N_{max}	μ_{max}	λ
Estimation	198	5.6×10^8	0.051	51.6
Borne inf - IC95%	193	5.3×10^8	0.049	48.6
Borne sup - IC95%	204	5.8×10^8	0.053	54.6

$\Longrightarrow$ proches des vraies valeurs :
$N_0^* = 200,\ N_{max}^* = 5 \times 10^8,\ \mu_{max}^* = 0.050,\ \lambda^* = 50.$
$\Longrightarrow$ meilleures que celles obtenues pour des temps équirépartis:

Paramètre	N_0	N_{max}	μ_{max}	λ
Estimation	195	5.29×10^8	0.055	64.0
Borne inf - IC95%	191	5.11×10^8	0.053	59.6
Borne sup - IC95%	199	5.47×10^8	0.056	68.4

$\Longrightarrow$ μ_{max}^* et λ^*ne sont pas dans les IC95%.

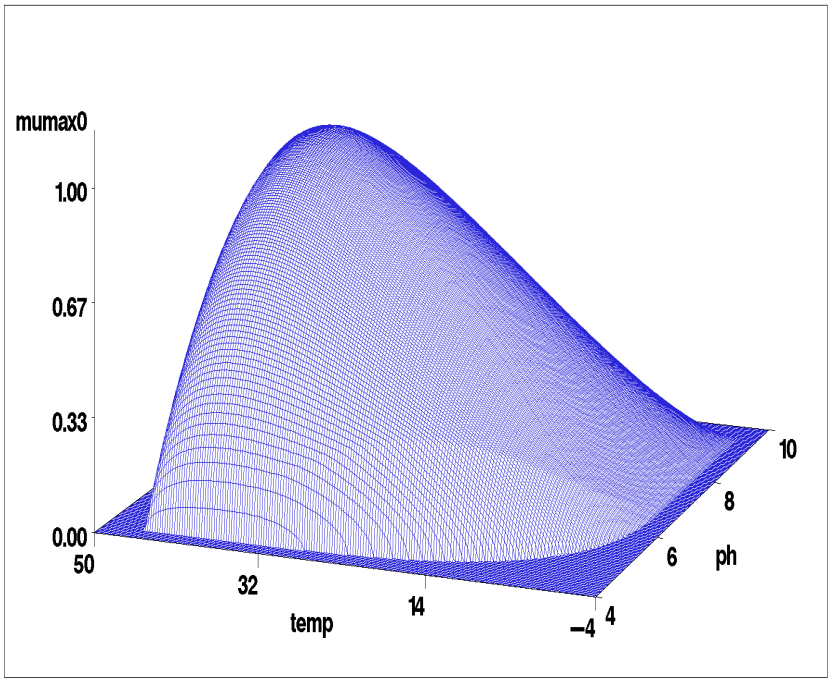
Perspectives sur cette problématique

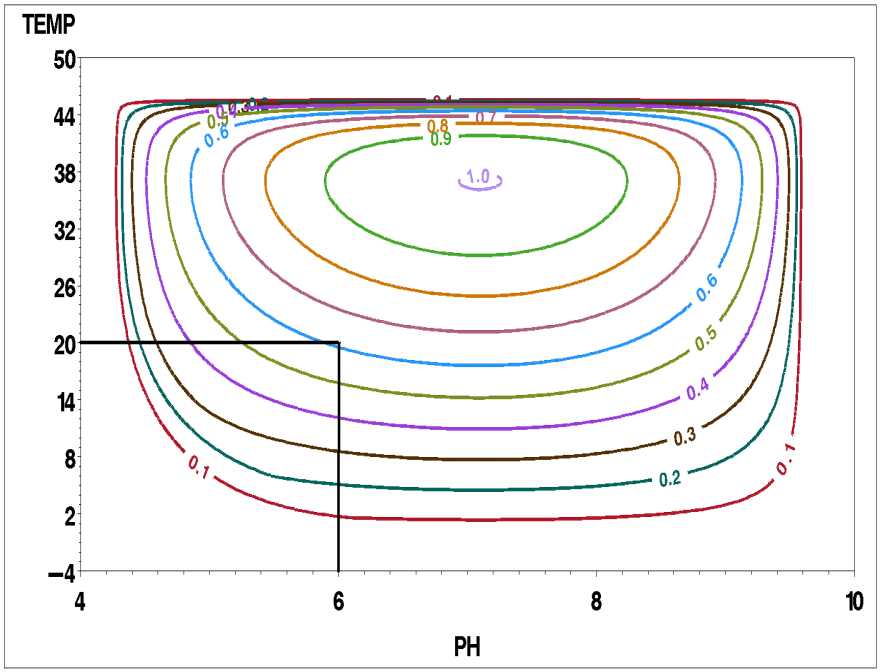
- Au plan statistique : **travail théorique** sur l'argumentation basée sur le lien entre information (au sens de l'entropie de Shanon) et les indices de sensibilité
- Au plan biologique:
  - Prendre en compte **simultanément** les croissances bactériennes de plusieurs types d'espèces en interaction (x_t est un vecteur).
  - **un vrai challenge**: Introduire des profils dynamiques contrôlés ou non pour les facteurs environnementaux qui apparaissent dans les modèles (dits secondaires) du μ_{max} et du λ.

- **Partie II : Plans d'expériences dans un cadre statique**

Un autre challenge -1

Le graphe du modèle secondaire ACL (Augustin & Carlier, 2000) et ses isocontours, pour le paramètre μ_{max} précédent:





Un autre challenge -2

La fonction du modèle ACL en fonction de la température et du pH est:

$$\mu_{\max} = \mu_{opt} \times CM_2(T) \times CM_1(pH) \times \xi(T, pH) + \varepsilon$$

avec:

- Si $T \leq T_{\min}$ alors $CM_2(T) = 0$
- Si $T_{\min} < T < T_{\max}$ alors :

$$CM_2(T) = \frac{(T - T_{\max})(T - T_{\min})^2}{(T_{opt} - T_{\min})\{(T_{opt} - T_{\min})(T - T_{opt}) - C_T\}}$$

où $C_T = (T_{opt} - T_{\max})[T_{opt} + T_{\min} - 2T]$.

Un autre challenge -3

... et :

- Si $pH \leq pH_{\min}$ alors $CM_1(pH) = 0$
- Si $pH_{\min} < pH < pH_{\max}$ alors :

$$CM_1(pH) = \frac{(pH - pH_{\max})(pH - pH_{\min})}{(pH_{opt} - pH_{\min})(pH - pH_{opt}) - (pH_{opt} - pH_{\max})\, C_{pH}}$$

où $C_{pH} = (pH_{\min} - pH)$.

Un autre challenge -4

.... et :

$$\xi(T, pH) = \left\{ \begin{array}{ccc} 1 & \text{si} & \psi \leq 0.5 \\ 2(1-\psi) & \text{si} & 0.5 < \psi < 1 \\ 0 & \text{si} & \psi \geq 1 \end{array} \right\}$$

où

$$\psi = \frac{\left(\frac{T_{opt}-T}{T_{opt}-T_{\min}}\right)^3}{2\left(1-\left(\frac{pH_{opt}-pH}{pH_{opt}-pH_{\min}}\right)^3\right)} + \frac{\left(\frac{pH_{opt}-pH}{pH_{opt}-pH_{\min}}\right)^3}{2\left(1-\left(\frac{T_{opt}-T}{T_{opt}-T_{\min}}\right)^3\right)}$$

et on note

$$\theta = \left\{\mu_{opt}, T_{\min}, T_{\max}, T_{opt}, pH_{\min}, pH_{\max}, pH_{opt}\right\}$$

Un autre challenge -5

Pour ces types de modèles statiques des **critères théoriques et des algorithmes** ont été publiés pour calculer des plans d'expériences:

- basés sur des critères spécifiques du cas non-linéaire,
- semi-bayésiens ou bayésiens.
- pour la modélisation de la probabilité de croissance/non-croissance.

Un autre challenge -6

- Pour le modèle ACL exposé précédemment une version utilisée aujourd'hui présente 10 paramètres et 3 facteurs expérimentaux ($Temp, pH, aw$).
- Le souhait des microbiologistes est une **démarche séquentielle dans la planification des expériences**.

Comment traiter ce problème ?

- $\Longrightarrow$ Passons en revue quelques critères potentiels décrits dans la littérature.

Critère 1

$$\Phi_{DN_F} = \det\left(N_F\left(\theta, \xi_N\right)\right) \quad \text{à} \quad \text{maximiser}$$

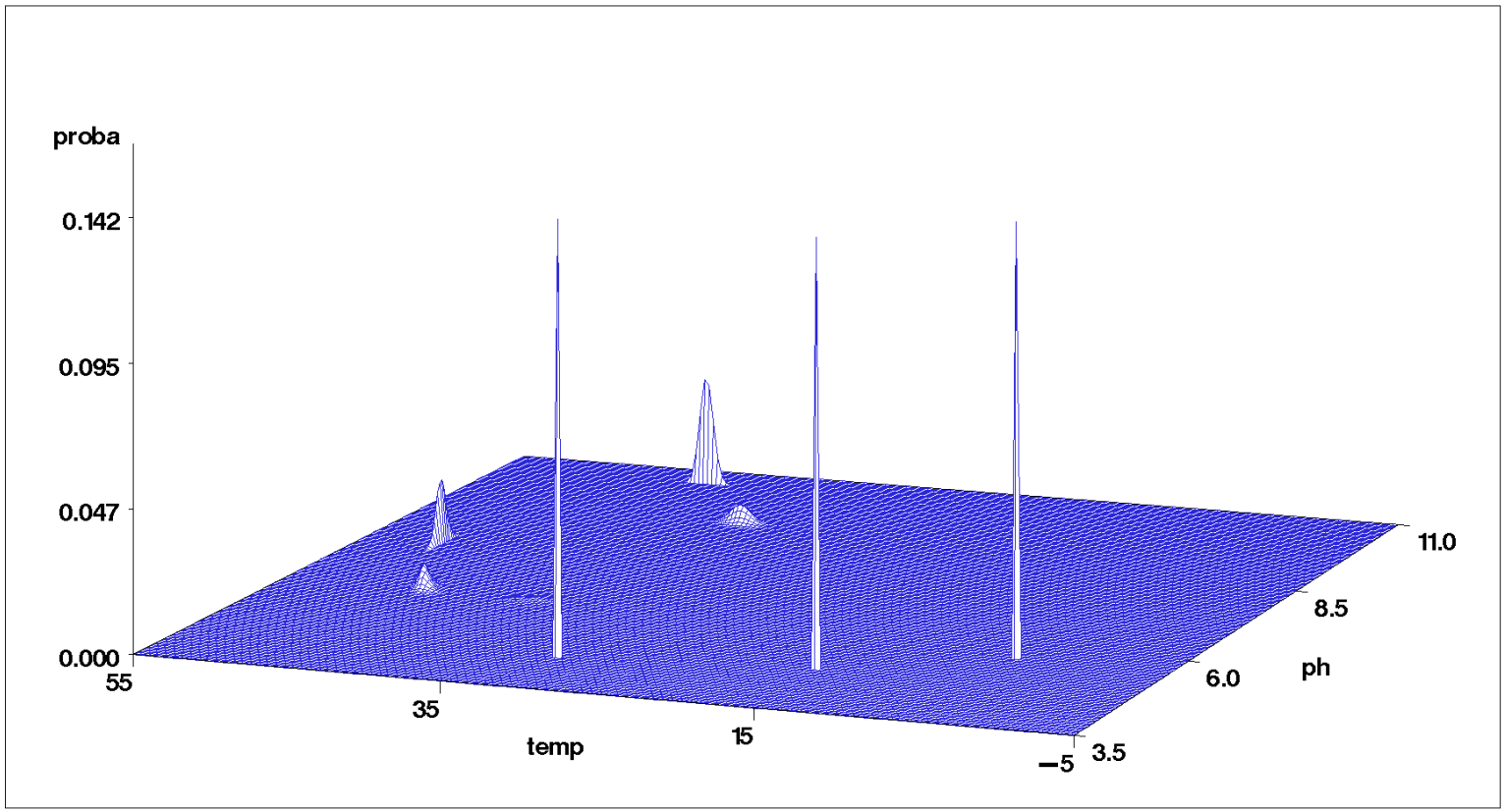
où la matrice d'information de Fisher pour paramètres aléatoires (Sorenson, 1980) s'exprime comme:

$$N_F\left(\theta, \xi_N\right) = E_\theta\left\{M_F\left(\theta, \xi_N\right)\right\} + E_\theta\left\{\frac{\partial \ln\left(p\left(\theta_0, \Sigma_{\theta_0}\right)\right)}{\partial \theta} \frac{\partial \ln\left(p\left(\theta_0, \Sigma_{\theta_0}\right)\right)}{\partial \theta^T}\right\}$$

- pour un plan d'expériences à N expériences ξ_N et $p\left(\theta_0, \Sigma_{\theta_0}\right)$ une loi a priori pour θ,
- avec $M_F\left(\theta, \xi_N\right) =$
  - $X^T X$, matrice d'information de Fisher à $1/\sigma^2$ près pour une variance homoscédastique, X est la matrice de modèle,
  - $X^T \Sigma^{-1} X$ est la matrice d'information de Fisher pour une variance hétéroscédastique.
- Walter & Pronzato, 1997; Pronzato & Walter 1985, dérivent plusieurs critères en espérance dans ce contexte.

Critère 1

Bien sûr une approche "locale" et non séquentielle (Atkinson & Donev, 1992) est facile à obtenir en donnant une valeur θ_0 à θ :
$\Longrightarrow$ le graphe d'une mesure $D(\theta_0)-$optimale à 7 $(= p)$ points de support pour le modèle ACL, pouvant servir de base à un plan d'expériences:



Critère 2

$$\begin{aligned}\Phi_{XB} &= E_y\left\{\mathcal{V}\left[\Re_X(\theta_0, y \mid p(\theta), C_F^0)\right]\right\} \\ &= \int_\Theta \int_\Theta \int_0^{\mathcal{F}_{p,\nu;\alpha}} g(t; p, \nu, \lambda) dt \; dm(\theta) p\left(\theta_0, \Sigma_{\theta_0}\right) d(\theta)\end{aligned}$$

à minimiser,
où:

- $\mathcal{V}(.)$ est le volume d'une région de confiance exacte (Vila & Gauchi, 2007),
- $g(t; p, \nu, \lambda)$ est la densité d'une Fisher monodécentrée (à p et ν ddl), de paramètre de non-centralité λ qui dépend du plan ξ_N.

Critère 3

$$\Phi_{DG} = \det\left[S\left(\xi, \theta^*\right)\right]$$

à minimiser.

- où $S\left(\xi, \theta^*\right)$ est l'erreur quadratique moyenne de $\hat{\theta}_{PWLS}$:

$$S\left(\xi, \theta^*\right) = E_{\theta^*, \xi}\left[\left(\hat{\theta}_{PWLS} - \theta^*\right)\left(\hat{\theta}_{PWLS} - \theta^*\right)^T\right]$$

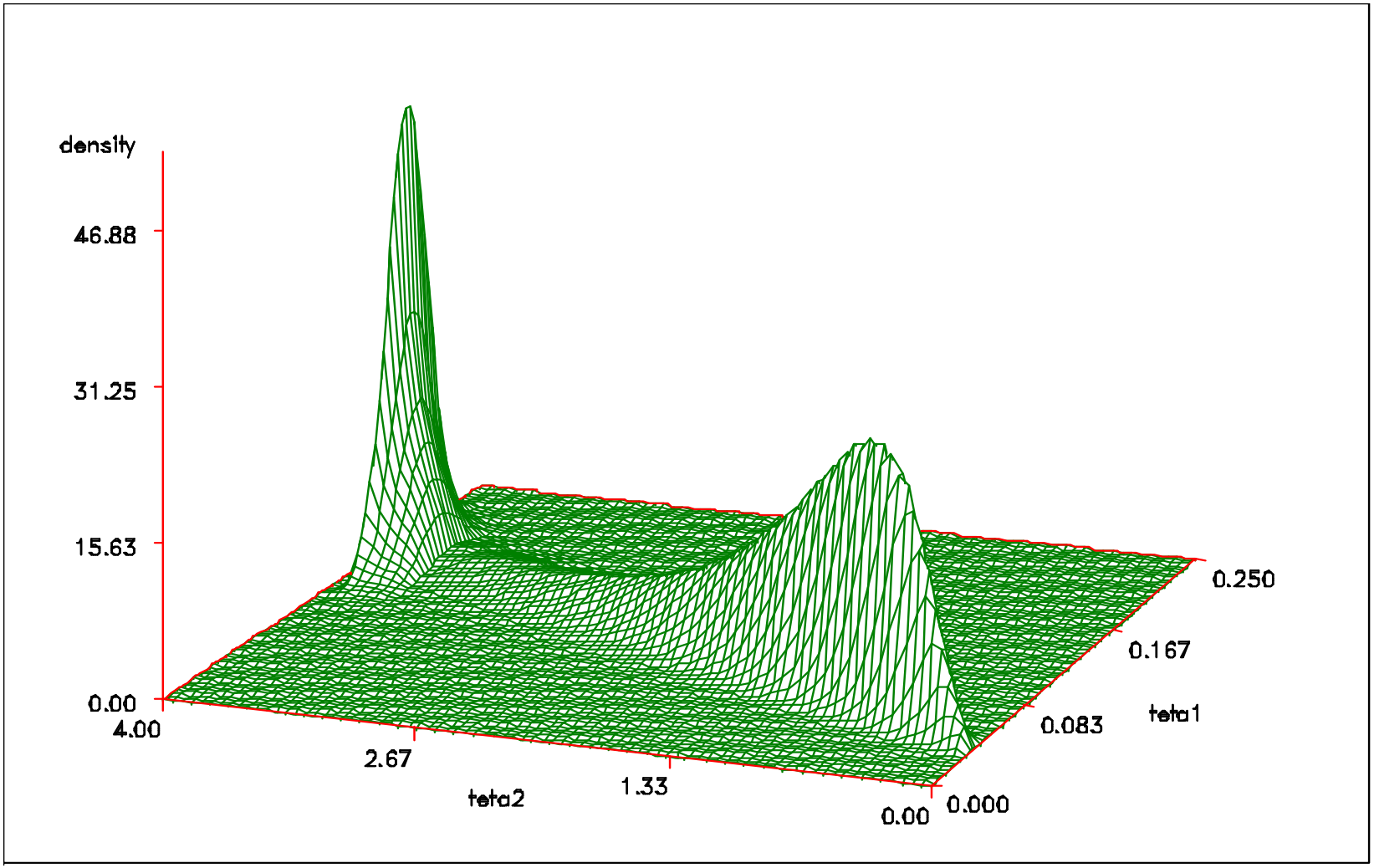
avec $\hat{\theta}_{PWLS}$ l'estimateur des moindres carrés non linéaires pondéré et pénalisé

$$\hat{\theta}_{PWLS} = \arg\min_{\theta \in \Theta}\left\{\left[y - \eta\left(\theta, \xi\right)\right]^T \Sigma^{-1}\left[y - \eta\left(\theta, \xi\right)\right] + 2w^*\left(\theta\right)\right\}$$

- et l'espérance est prise pour $q_\xi\left(\theta \mid \theta^*\right)$ la densité de probabilité exacte (ou presque exacte), à distance finie (en N), de $\hat{\theta}_{PWLS}$ (Pazman, & Pronzato, 1992, Gauchi & Pazman, 2006, Pronzato & Pazman, 2013).

Un exemple d'intégrande

Pour le critère Φ_{DG} (densité exacte de Pazman pour un modèle $p = 2$, et $P = 1$):



Algorithmes

- Les algorithmes d'optimisation stochastique sont les mieux à même de traiter les 3 critères précédents, en évitant la phase d'intégration.
- Un processus stochastique (convergent) de minimisation ($\xi_{N_S}^{k+1}$ et $\xi_{N_S}^{k}$ sont des vecteurs colonne de dimension $(PN_S \times 1)$) s'écrit:

$$\xi_{N_S}^{k+1} = \xi_{N_S}^{k} - a_k \Psi^k|_{\xi_{N_S}^k} \quad , \quad a_k = a_0 / k$$

où θ^k est tiré aléatoirement dans des lois sur Θ:

- $\Psi^k = \frac{\partial \Phi(\theta^k, X)}{\partial X}$ (gradient stochastique, Robbins & Monro, 1951),
- $\Psi^k = W^k = \frac{Z_\Phi(\theta^{k,u}, x^k + c_k) - Z_\Phi(\theta^{k,v}, x^k - c_k)}{2c_k}$ (approximation du gradient stochastique, Kiefer & Wolfowitz, 1952),
- $W_j^k = \sum_{i=1}^{h} \rho_i \frac{1}{N_i} \sum_{s=1}^{N_i} \frac{Z_\Phi^+ - Z_\Phi^-}{2c_k}$ (approximation du gradient stochastique, Fabian, 1967, 1968).

$\Longrightarrow$ **difficultés: choix de a_0 et difficulté d'exploration de Θ dès que** $p \geq 3$ (voir la forme du graphe précédent)

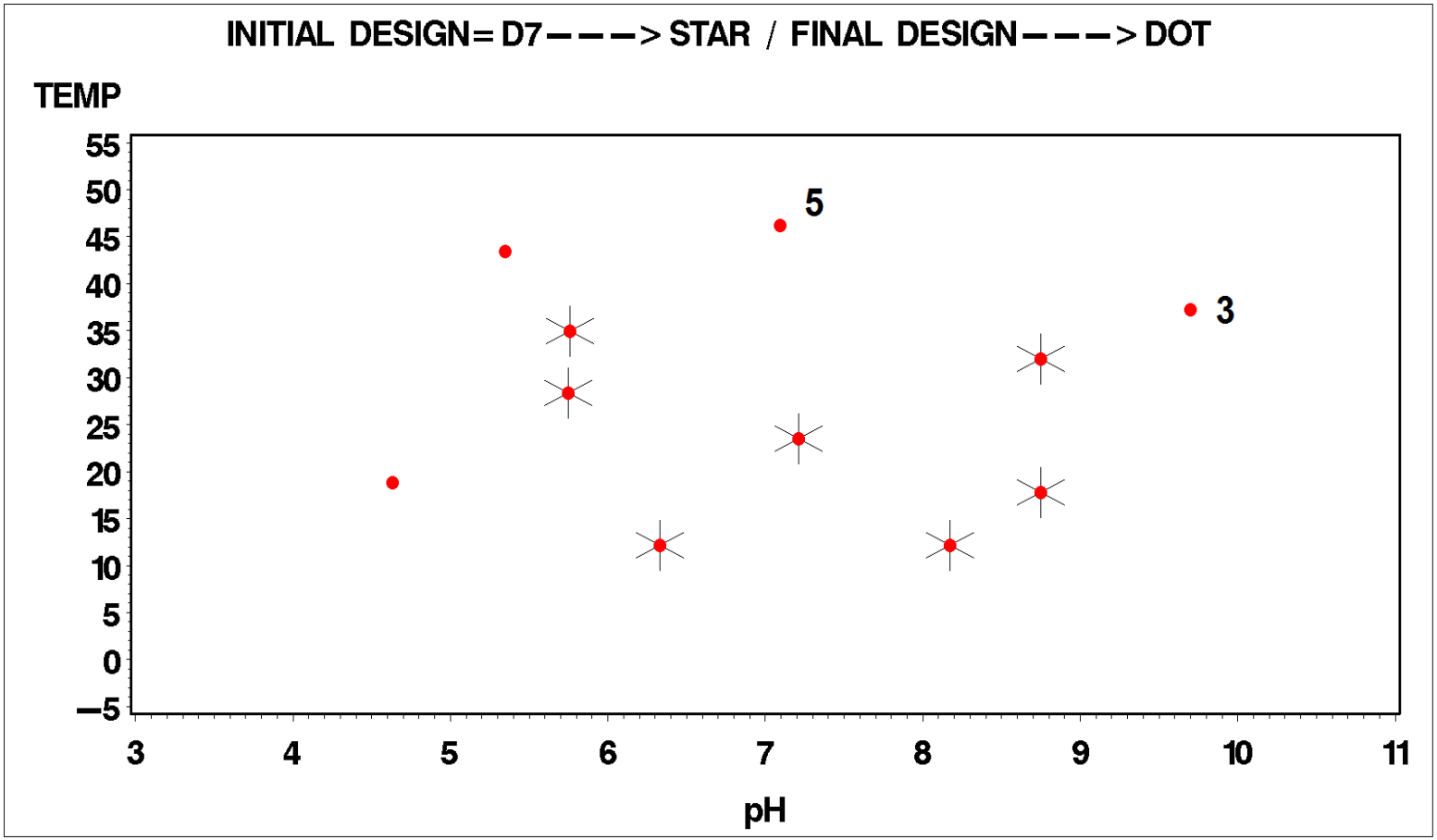
Critère 4

$$\begin{aligned}\Phi_U &= U\left(\xi_N\right) \\ &= \int u\left(\xi_N, \theta, y\right) p_{\xi_N}\left(\theta, y\right) d\theta dy \\ &= \int u\left(\xi_N, \theta, y\right) p\left(\theta\right) p_{\xi_N}\left(y|\theta\right) d\theta dy\end{aligned}$$

à maximiser, où $U\left(\xi_N\right)$ l'utilité espérée.
Algorithmes:
$\Longrightarrow$ approche moderne de résolution : méthodes de **simulation MCMC** (Muller, 1999)**, filtrage particulaire** (Amzal et al.), remplacent les méthodes d'optimisation stochastique.

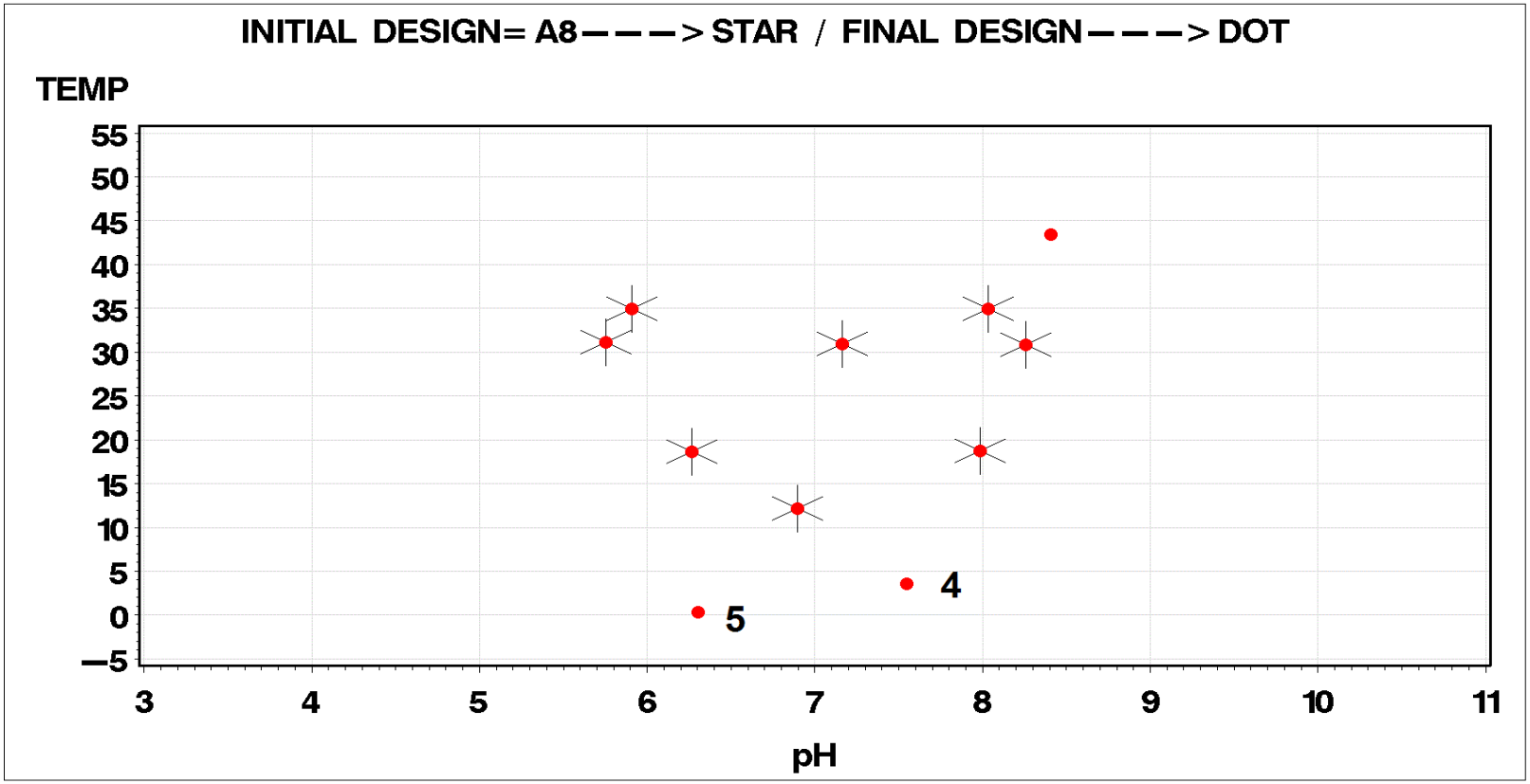
Résultats pour le modèle ACL -1

Une stratégie séquentielle approximative (en cours de publication) est possible sur ce type de modèle secondaire avec un critère de $D(\theta_0)$-$D(\hat{\theta}_k, k = 0, ..., 9)$-optimalité locale:



Résultats pour le modèle ACL -2

.... ou avec un critère de A(θ_0)-A($\hat{\theta}_k, k = 0, ..., 9$)-optimalité locale:



$\Longrightarrow$ Mais une démarche séquentielle plus rigoureuse, basée sur un des 4 critères précédents, qui semblerait plus adaptée, est très lourde à mettre en oeuvre d'un point de vue informatique.

Conclusion

- La grande dimension de Θ (souvent ≥ 5),
- La grande variance de l'erreur (matériel vivant) et son hétéroscédasticité dépendant de la position des points de support du plan d'expériences,
- Le nombre d'expériences à réaliser fortement contraint,

$\Longrightarrow$ La difficulté à construire des plans optimaux pour modèles non linéaires en biologie est aujourd'hui essentiellement de nature **informatique**, et non pas théorique car plusieurs critères bien adaptés existent et sont publiés.
$\Longrightarrow$ Une solution envisageable: la **distribution** et/ou la **parallélisation** des codes informatiques.

Bibliographie -1

- Amzal, B., Bois, FY, Parent, E., Robert, C., 2006, "Bayesian-optimal design via interacting particle systems", JASA, 101, 474, 773-785.
- Atkinson, A.C., Donev, A.N., 1992. "Optimum Experimental Designs". Oxford, Clarendon Press.
- Augustin, JC, Carlier, V., 2000. "Modelling the growth rate of Listeria monocytogenes with a multiplicative type model including interactions between environmental factors", IJFM, 56, 53-70.
- Baranyi, J., Roberts, T.A., 1995. "Mathematics of predictive food microbiology", International Journal of Food Microbiology 26: 199-218.
- Fabian, V., 1967. "Stochastic approximation of minima with improved asymptotic speed", Ann. Math. Statist. 38, 191-200.
- Fabian, V., 1968a. "On asymptotic normality in stochastic approximation", Ann. Math. Statist. 39, 1327-1332.
- Fabian, V., 1968b. "On the choice of design in stochastic approximation methods", Ann. Math. Statist. 39, 457-465.
- FILTREX: http://w3.jouy.inra.fr/unites/miaj/public/logiciels/filtrex/

Bibliographie -2

- Gauchi, J.-P., Pázman, A., 2006. "Designs in nonlinear regression by stochastic minimization of functionals of the mean square error matrix", J. Statist. Planning and Inference, 136, 3, 1135-1152.
- Gauchi, JP., Vila, JP., 2012. "Nonparametric particle filtering approaches for identification and inference in nonlinear state-space dynamic systems", Statistics and Computing, 23:523–533, DOI 10.1007/s11222-012-9327-7.
- Kiefer, J., Wolfowitz, J., 1952. "Ann. Math. Stat.", 23, 462-466.
- Muller, P., 1999. "Simulation-Based Optimal design", Bayesian Statistics, 6, 459-474.
- Pázman, A., Pronzato, L., 1992. "Nonlinear experimental design based on the distribution of estimators", J. Statist. Planning and Inference 33, 385-402.
- Pronzato, L., Walter, E., 1985. "Robust experiment design via stochastic approximation", Math. Biosciences, 75, 103-120.

Bibliographie -3

- Pronzato, L, Pazman, A., 2013. “Design of Experiments in Nonlinear Models", Springer (Lecture Notes in Statistics, Vol. 212), New York, Heidelberg, XV+399 pages.
- Robbins, H., Monro, S., 1951. "A stochastic approximation method", Ann. Math. Stat. 22, 400-407.
- Rossi, V., Vila, J.P. 2005. "Approche non paramétrique du filtrage de système non linéaire à temps discret et à paramètres inconnus", C.R. Acad. Sci. Paris, I, 340, 759-764.
- Rossi, V., Vila, J.P. 2006. "Nonlinear filtering in discrete time: A particle convolution approach", Pub. Inst. Stat. Univ. Paris, L, 3, 71-102.
- Sorenson, H.W., 1980. "Parameter estimation, principles and problems", Marcel Dekker, New York.
- Tenenhaus, M. "Régression PLS - théorie et pratique", Technip, Paris, 1998.

Bibliographie -4

- Vila, J.-P., Gauchi, J.-P., 2007. "Optimal designs based on exact confidence regions for parameter estimation of a nonlinear regression model", J. of Statistical Planning and Inference 137: 9, 2935-2953.
- Walter, E., Pronzato, L., 1997. “Identification of Parametric Models from Experimental Data”, Springer-Verlag, Heidelberg.

Merci de votre attention !