Calcul robuste d'agrégats de données de consommation électrique

Omar Fawzi

2008D019

Septembre 2008

Département Informatique et Réseaux
Groupe IC2 : Interaction, Cognition et Complexité

Calcul robuste d'agrégats de données de consommation électrique

Omar Fawzi
omar.fawzi@ens-lyon.org
Encadrants : Georges Hébrail et Marie-Luce Picard
BILab : TELECOM ParisTech et EDF R&D

30 septembre 2008

1 Contexte

L'installation de compteurs électriques communicants pour relever à distance la consommation électrique des clients permettra d'obtenir des courbes de charge à un pas beaucoup plus fin pour tous les clients. En effet, il est prévu que ces compteurs stockent les consommations électriques demi-heure par demi-heure et envoient leurs courbes de consommation régulièrement. Mais ceci ne va pas sans poser quelques problèmes. Premièrement, le volume des données est important (le nombre de clients s'élève à plus de trente millions), et donc les données doivent être traitées à la volée ; même si ces données seront stockées, l'accès à des vieilles données sera sans doute coûteux. Deuxièmement, dans un tel grand réseau, les pannes et les erreurs ne seront sûrement pas rares, il faut donc pouvoir gérer automatiquement les pannes de compteurs, et corriger les éventuelles erreurs. On s'intéressera ici uniquement au cas des retards qui sont dus soit à des pannes temporaires de compteurs soit à des défaillances dans le réseau de communication. Pour modéliser les retards, on supposera que les données ne sont jamais perdues, mais peuvent arriver avec un certain retard. Néanmoins, on souhaite avoir une estimation presque temps-réel de la consommation à chaque instant (demi heure), raffinant cette estimation au fur et à mesure que les données arrivent.

Idéalement les compteurs envoient leurs données à un pas de temps régulier qui correspond au taux qui leur est demandé. Il est clair qu'en pratique ces délais ne peuvent pas être respectés, en effet, différents phénomènes dans le compteur lui même, ou dans le réseau peuvent introduire un retard ou une avance dans l'envoi des valeurs. De plus, supposer que tous les compteurs sont synchronisés pour envoyer toujours dans les mêmes délais est irréaliste. C'est pour cela que l'on modélise la date d'envoi des mesures d'un compteur par une distribution dont la moyenne est fixée.

Le problème qu'on se pose est de calculer certains agrégats de ces courbes (typiquement les courbes par fournisseur) et non les courbes individuelles des clients, qui sont très difficiles à prédire ou à corriger. Notons que dans ce cadre, on ne contrôle pas directement le plan de sondage, néanmoins, c'est le contrôle de la qualité des compteurs et du réseau qui nous permet de piloter le plan de sondage d'une manière indirecte.

2 Modélisation

2.1 Modèle

2.1.1 Consommations

On notera $\mathcal{C}$ l'ensemble des clients, $C_i(t)$ la consommation électrique du clients i à l'instant t. Pour $A \subset \mathcal{C}$, on notera $C_A(t)$ la consommation totale des clients dans A (A étant par exemple les clients d'un certain fournisseur). Dans la suite, on s'intéressera à estimer $C_A(t)$ pour un certain A, avec $|A| = N$. Pour alléger les notations, on écrira $C(t)$ au lieu de $C_A(t)$.

Remarquons que le temps est discrétisé, et l'unité de temps T (aussi appelé période dans la suite) correspond au pas auquel les données de consommations se présentent. Typiquement cette période sera de l'ordre de la demi-heure ; pour les simulations, la consommation est relevée toutes les 10 minutes. Lorsqu'on parlera d'un instant t, t est un entier qui représente le numéro de la période considérée.

2.1.2 Retards

Chaque compteur envoie régulièrement l'ensemble des consommations depuis le dernier envoi. Si à un instant t, le compteur décide d'envoyer les données stockées, il envoie toutes les consommations pour tous les instants $t' \leq s < t$ où t' désigne le dernier instant où le compteur a rapporté ses valeurs de consommations.

On modélise le temps écoulé entre deux envois successifs par des variables aléatoires indépendantes identiquement distribuées. Ainsi, le compteur i attend après chaque envoi une certaine durée suivant une distribution μ_i avant son prochain envoi. On supposera toutes les variables aléatoires indépendantes pour un même compteur (processus markovien) mais aussi entre les différents compteurs.

Les distributions de retards ne sont pas nécessairement dues à des erreurs mais peuvent être contrôlées. En effet, le fait que les compteurs peuvent recevoir des instructions fait que l'on peut avoir un certain contrôle des délais d'envoi. On voit bien que l'on a intérêt à donner une priorité aux gros consommateurs, puisque on voudrait plus de précision pour les grandes valeurs de consommations, celles-ci influant davantage sur la consommation totale.

2.2 Motivations

Ce modèle assez simple est étudié pour comprendre les phénomènes suivants :

- Sans supposer d'imperfection dans le système, il est possible que la randomisation du protocole utilisé pour récolter les données évite les congestions dans le réseau. La question que l'on se pose alors est : un protocole randomisé peut-il nous permettre d'obtenir un meilleur rapport qualité de l'approximation/bande passante réseau utilisée ? Les protocoles probabilistes sont en effet beaucoup utilisée dans les calculs distribués. Par exemple, les protocoles CSMA/CD, qui gèrent les ressources physiques d'un réseau, résolvent les collisions en attendant une durée aléatoire avant de parler.
- La supposition que les valeurs de tous les compteurs arrivent exactement au moment attendu est irréaliste. La modélisation de ce retard par un certain aléa semble être l'approche la plus naturelle.
- Comment choisir les moyennes des intervalles d'envoi de chaque compteur pour obtenir une bonne précision ?

2.3 Littérature

Le calcul d'agrégats dans les réseaux de capteurs est très étudié dans la littérature. De ce fait, des méthodes robustes qui tolèrent des défaillances de certains éléments du réseaux existent, par exemple dans [5]. Cependant, dans le cadre des réseaux de capteurs, une panne peut affecter l'acheminement de données qui correspondent à d'autres capteurs, notamment dans la cas où l'agrégat est construit en formant un arbre de capteurs. Les solutions proposées pour ce problème reposent souvent sur la construction d'une structure plus complexe qu'un arbre pour assurer que deux capteurs soient connectés par plus d'un chemin.

Dans notre modélisation, il n'y a pas de connexion entre les compteurs, et donc le problème n'est pas dans l'acheminement des données.

D'un autre coté, la sécurité des protocoles d'agrégation par rapport à des capteurs qui ont de fausses données (et même des données adversaires, qui visent à fausser l'agrégat) fait l'objet d'études concernant la "resilient agregation" [6, 2].

3 Méthodes d'estimation

On peut voir le problème comme un problème de sondage. Considérons la consommation à un instant t_0 donné. Dans toute cette partie, on cherche à estimer la consommation en t_0. On omettra parfois d'indiquer la dépendance en t_0 pour alléger les notations. Les valeurs de consommation concernant t_0 arrivent à différents instants $t > t_0$. Ainsi, pour tout $t > t_0$, on a un échantillon $S_t \subset \mathcal{C}$ à l'instant t des compteurs ayant envoyé une valeur pour t_0. Nous avons donc une suite croissante d'échantillons. On notera $n_t = |S_t|$. Cet échantillon est aléatoire puisque les dates d'envois le sont.

3.1 Sondage simple

3.1.1 Estimateur

On peut donner une estimation du total par une formule du type

$$\hat{C}_t(t_0) = \frac{N}{n_t} \sum_{i \in S_t} C_i.$$

Pour affiner cette estimation on peut effectuer une post-stratification de la population en fonction de leur consommation et définir un estimateur en conséquence. En effet, si on suppose que la population est divisée en différentes classes $\mathcal{C}_1, \ldots, \mathcal{C}_r$, de taille $N_1, \ldots, N_r$ alors on peut estimer le total

$$\hat{C}_t(t_0) = \sum_{h=1}^{r} \frac{N_h}{n_{h,t}} \sum_{i \in S_t \cap \mathcal{C}_h} C_i.$$

On peut donc aussi voir cet estimateur comme un estimateur avec des poids où les poids sont $w_i = \frac{N_h}{n_h}$ si $i \in \mathcal{C}_h$. On voit donc que le poids d'un individu est d'autant plus grand que le nombre de valeurs reçues de sa classe est faible.

3.1.2 Précision

Pour analyser notre estimateur, il faut mettre un modèle probabiliste en place. Pour faciliter les calculs, on supposera que l'on a affaire à un sondage avec un nombre d'individus échantillonné fixé. Ceci n'est pas vrai dans notre modèle de retard mais le calcul est plus simple. En effet on suppose que la taille de l'échantillon étant suffisamment grande, la taille de l'échantillon est assez proche de sa moyenne, on va alors considérer $n_t \approx \mathbb{E}(n_t) = \sum_{i \in \mathcal{C}} \mathbb{P}(i \in S_t)$. Ceci revient à étudier un autre estimateur $\hat{C}'_t = \frac{N}{\mathbb{E}(n_t)} \sum_{i \in S_t} C_i$ qui est plus simple, et qui est avec grande probabilité très proche de notre estimateur original. Donc pour tous les calculs, on se basera sur $\hat{C}'_t$ sans le préciser, et on confondra n_t et $\mathbb{E}(n_t)$.

Tout d'abord commençons par regarder le biais de notre estimateur, pour ceci, on écrit

$$\hat{C}_t(t_0) = \frac{N}{n_t} \sum_{i \in \mathcal{C}} \mathbb{I}_{i \in S_t} C_i$$

où $\mathbb{I}_A$ représente la fonction indicatrice de l'évènement A. Il en découle

$$\begin{aligned}
\mathbb{E}(\hat{C}_t - C) &= \frac{N}{n_t} \sum_{i \in \mathcal{C}} \mathbb{E}(\mathbb{I}_{i \in S_t}) C_i - C \\
&= \frac{N}{n_t} \sum_{i \in \mathcal{C}} \mathbb{P}(i \in S_t) C_i - C \\
&= \sum_{i \in \mathcal{C}} \Big(p_i^t \frac{N}{n_t} - 1 \Big) C_i
\end{aligned}$$

où on note $p_i^t = \mathbb{P}(i \in S_t)$.

On a donc un biais nul lorsque les éléments sont tous échantillonnés avec la même probabilité, i.e. lorsque les p_i^t sont tous égaux. Pour quantifier ensuite la précision de l'estimateur, on calcule la variance de notre estimateur. Notre estimateur s'écrivant comme la somme de variables aléatoires indépendantes, la variance est simplement la somme des variances, on a donc

$$\begin{aligned} Var(\hat{C}_t) &= \sum_{i \in \mathcal{C}} \frac{N^2}{(n_t)^2} C_i^2 Var(\mathbb{I}_{i \in S_t}) \\ &= \frac{N^2}{(n_t)^2} \sum_{i \in \mathcal{C}} C_i^2 p_i^t (1 - p_i^t) \end{aligned}$$

On voit donc que la variance est d'autant plus grande que les p_i^t sont proches de $\frac{1}{2}$ à taille d'échantillon $n_t = \sum_i p_i^t$ fixée. Pour minimiser cette variance on a donc intérêt à affecter une grande probabilité d'inclusion aux compteurs ayant une grosse consommation. Remarquons qu'au vu de l'expression de la variance, il semblerait que l'on peut tout aussi bien prendre p_i très petit pour les gros consommateurs, mais rappelons nous que ceci fait que le biais de l'estimateur devient plus important, donc on concentre l'estimateur mais sur des mauvaises valeurs. On vérifie que lorsque t est assez grand, on aura $p_i^t = 1$ pour tout i et donc une variance nulle.

On se pose maintenant la question de l'estimateur stratifié. L'estimateur stratifié est simplement une somme d'estimateurs simples pour lesquelles on suppose une certaine homogénéité dans la population. On voit directement l'intérêt de la stratification déjà sur le biais de l'estimateur, en effet si les individus d'une classe ont des probabilités d'inclusions semblables, alors comme on l'a vu le biais est petit. En ce qui concerne la variance de l'estimateur, celle-ci diminue si les échantillons dans les classes qui ont des grandes consommations sont plus grands, et donc elle diminue si les grands consommateurs ont des grandes probabilités d'inclusion.

3.1.3 Estimation de la précision

Il est clair que l'on ne peut pas calculer exactement la variance de notre estimateur pour donner un intervalle de confiance, puisqu'on n'a pas toutes les valeurs des consommations. Mais on peut tout de même en donner une estimation. On cherche donc à estimer la quantité $\frac{N^2}{(n_t)^2} \sum_{i \in \mathcal{C}} C_i^2 p_i^t (1 - p_i^t)$. Pour ceci, on peut faire comme notre estimation de la consommation totale et donc estimer $\sum_{i \in \mathcal{C}} C_i^2 p_i^t (1 - p_i^t)$ par $\frac{N}{n_t} \sum_{i \in S_t} C_i^2 p_i^t (1 - p_i^t)$. Maintenant quelle est la qualité de cet estimateur de la variance ? Si on fait les calculs, on trouve que la variance ici est un ordre de grandeur plus petit que la variance que l'on cherche à estimer, et donc on choisit de négliger la précision de cet estimateur de la variance, on considère donc que notre estimateur de la variance

$$\hat{V} = \frac{N^3}{n_t^3} \sum_{i \in S_t} C_i^2 p_i^t (1 - p_i^t)$$

est exact. Pour vérifier la cohérence de notre formule, on peut toujours regarder le cas simple où $p_i^t = \frac{n_t}{N}$, et on retrouve bien l'estimateur de la variance pour un sondage simple. Avec ce calcul, on peut donc définir un intervalle de confiance à 95%, on a alors

$$IC = \left[\hat{C}_t - 2\sqrt{\hat{V}}, \hat{C}_t + 2\sqrt{\hat{V}}\right].$$

Le problème est que notre estimateur est biaisé, donc cette formule n'est plus valable normalement. Cependant, si on fait des classes et que les consommations sont similaires et donc que le biais est petit, on peut quand même supposer que la formule reste correcte.

3.1.4 Coût et informations nécessaires

Cette méthode ne nécessite aucune information ni sur les consommations ni sur les retards, pour l'estimateur stratifié, il suffit d'avoir une classification des compteurs qui correspondent à des plages de consommations typiques.

A chaque pas de temps t, l'algorithme reçoit un certain nombre de valeurs qui correspondent à différents intervalles de temps. Il va alors revoir son estimation pour tous les instants passés qui sont concernés par l'échantillon et mettre à jour leur estimation. Par exemple supposons que nous ayons reçu au total $n_t(t_0)$ valeurs de compteurs concernant l'instant t_0, alors on revoit notre estimateur

$$\hat{C}_{t+1}(t_0) = \frac{n_t(t_0)}{n_{t+1}(t_0)} \hat{C}_t(t_0) + \frac{N}{n_{t+1}(t_0)} \sum_{i \in S_{t+1} - S_t} C_i.$$

Ainsi, on voit que la mise à jour est très simple. Chaque nouvelle valeur reçue est donc traitée une fois, le coût de calcul à chaque instant est donc linéaire en le nombre de valeurs reçues à cet instant, ce qui est optimal.

Par contre pour donner une estimation de la précision sous le modèle probabiliste mis en place, on a besoin des probabilités d'inclusions, et donc de la distribution des retards pour chaque compteur. De plus le calcul de la variance ne s'écrit pas aussi facilement de manière incrémentale comme la somme vu la présence des termes p_i^t qui dépendent de t dans la somme.

3.2 Méthode avec probabilité d'inclusion

3.2.1 Estimateur

Notre échantillon est en fait issu d'un sondage à probabilités inégales. Dans ce cas, il est naturel de regarder l'estimateur linéaire sans biais (dit de Horvitz Thompson). Si on connait la distribution des retards, on peut calculer les probabilités d'inclusion de chaque compteur dans l'échantillon S_t. La distribution de cet échantillon dépend de la dernière

date d'envoi pour chaque compteur. En effet, la probabilité d'inclusion $\mathbb{P}(i \in S_t)$ est donnée par la probabilité qu'il y ait un envoi entre t_0 et t pour le compteur i sachant la dernière date d'envoi pour le compteur i.

Si on note $l_i^{t_0}$ le dernier instant pour lequel nous avons reçu une valeur pour le compteur i avant t_0, alors on a

$$\mathbb{P}(i \in S_t) = \mathbb{P}(l_i^{t_0} + R_i \leq t | R_i > t_0 - l_i^{t_0})$$

où $R_i \sim \mu_i$ représente le délai d'envoi suivant l'envoi de l'instant $l_i^{t_0}$.

Notre estimateur s'écrit donc

$$\hat{C}_t(t_0) = \sum_{i \in S_t} w_i^t C_i(t_0)$$

où les w_i^t sont des poids associés aux compteurs, qui vont être choisis de sorte que l'estimateur $\hat{C}_t(t_0)$ soit sans biais. Il faut prendre :

$$w_i^t = \frac{1}{\mathbb{P}(l_i^{t_0} + R_i \leq t | R_i > t_0 - l_i^{t_0})}$$

Remarquons que notre solution regarde l'échantillon d'ensemble ne faisant pas de différence entre les éléments qui sont arrivés à des instants différents entre t_0 et t. A chaque instant $t > t_0$ on refait une toute nouvelle estimation à partir l'échantillon S_t. On pourrait peut être imaginer une estimation incrémentale qui corrige l'estimateur en fonction des nouvelles valeurs qui sont arrivées à l'instant t.

3.2.2 Précision de l'estimation

Pour analyser cet estimateur, on calcule sa variance. On rappelle les notations $p_i^t = \mathbb{P}(l_i^{t_0} + R_i \leq t | R_i > t_0 - l_i^{t_0})$. On a alors (la dépendance en t_0 est omise pour alléger les notations), en utilisant le fait que les évènements $(i \in S_t)_{i \in \mathcal{C}}$ sont deux à deux indépendants,

$$\begin{aligned} Var(\hat{C}_t) &= \sum_{i \in \mathcal{C}} \frac{C_i^2}{(p_i^t)^2} Var(\mathbb{I}_{i \in S_t}) \\ &= \sum_{i \in \mathcal{C}} C_i^2 (\frac{1}{p_i^t} - 1) \end{aligned}$$

Rappelons que dans notre cas, l'échantillon n'est pas de taille fixe, mais il est le résultat des "décisions" indépendantes de chaque compteur. Ainsi, contrairement au cas où la taille de l'échantillon est fixée, on n'obtient pas un estimateur de variance nulle lorsque les p_i^t sont proportionnels aux C_i. Cependant si on suppose fixé le nombre moyen d'éléments dans l'échantillon, l'expression de la variance est quand même minimisée lorsque les p_i^t sont proportionnels aux C_i. En effet, en notant $\mathbb{E}(n_t) = \sum_{i \in \mathcal{C}} p_i^t$ le nombre moyen d'éléments dans l'échantillon on a

$$\begin{aligned}\sum_{i\in\mathcal{C}} C_i^2(\frac{1}{p_i^t}-1) &= \Big(\sum_{i\in\mathcal{C}} C_i^2\frac{1}{p_i^t}\Big)\Big(\frac{\sum_{i\in\mathcal{C}} p_i^t}{\mathbb{E}(n_t)}\Big) - \sum_{i\in\mathcal{C}} C_i^2 \\ &\geq \frac{1}{\mathbb{E}(n_t)}\Big(\sum_{i\in\mathcal{C}} C_i\sqrt{\frac{1}{p_i^t}}\sqrt{p_i^t}\Big)^2 - \sum_{i\in\mathcal{C}} C_i^2 \\ &= \frac{1}{\mathbb{E}(n_t)}\Big(\sum_{i\in\mathcal{C}} C_i\Big)^2 - \sum_{i\in\mathcal{C}} C_i^2\end{aligned}$$

en utilisant l'inégalité de Cauchy-Schwarz. On a donc intérêt (comme nous l'avions vu intuitivement) à faire en sorte que les plus gros consommateurs aient des délais plus faibles et donc une plus grande probabilité d'inclusion dans l'échantillon. Mais même si on arrive à bien choisir les probabilités p_i^t, on ne peut pas espérer un estimateur parfait, en effet, on ne peut pas obtenir d'estimateur ayant une variance plus petite que $\frac{1}{n_t}(\sum_{i\in\mathcal{C}} C_i)^2 - \sum_{i\in\mathcal{C}} C_i^2$. Remarquons quand même que choisir les p_i^t proportionnels aux C_i n'est pas toujours possible, en effet si n_t est grand, $n_t\frac{C_i}{\sum_{j\in\mathcal{C}} C_j}$ peut être plus grand que 1.

3.2.3 Estimation de la précision

Pour calculer un intervalle de confiance, il nous faut estimer la variance de notre estimateur. L'estimateur le plus naturel dans ce cas est

$$\hat{V} = \sum_{i\in S_t} \frac{1}{p_i^t} C_i^2(\frac{1}{p_i^t}-1).$$

Maintenant, puisque notre estimateur $\hat{C}_t$ est sans biais, on a un intervalle de confiance à 95% de

$$IC = \Big[\hat{C}_t - 2\sqrt{\hat{V}}, \hat{C}_t + 2\sqrt{\hat{V}}\Big].$$

3.2.4 Coût et informations nécessaires

Pour cette méthode, à chaque instant, il faut recalculer les probabilités d'inclusion pour tous les instants du passé pour lesquels on a reçu des valeurs. Si on note Δ les instants du passé pour lesquels un nouvel échantillon a été reçu en t, alors le nombre de pas de calcul que l'on effectue est proportionnel à $\sum_{s\in\Delta} n_t(s)$, vu qu'il faut reprendre en considération toutes les valeurs qui concernent l'instant s et non seulement les valeurs qui sont arrivées en t. Ce nombre peut être beaucoup plus grand que le nombre de valeurs reçues à l'instant t. En effet, ce nombre peut très bien aller jusqu'à $N \times P$ où P est la taille typique de la période concernée par les échantillons.

3.3 Estimation par la régression

Une autre manière de voir le problème est la suivante. On a un modèle général de prévision de consommation et les échantillons qu'on reçoit nous permettent de calculer des paramètres du modèle. Donc on essaye d'effectuer un sondage qui prend en compte une variable corrélé (qui est ici la consommation dans le passé qui est connue pour chaque client). On écrit alors un modèle simple de relation entre la consommation en t_0 et la consommation en $t_0 - d$,

$$C_i(t_0) = a + bC_i(t_0 - d) + U_i. \tag{1}$$

On effectue une régression linéaire en choisissant a et b pour minimiser la somme des distances au carré, c'est à dire pour minimiser $\sum_i U_i^2$.

Ceci donne pour b l'expression suivante

$$b^* = \frac{\sum_{i\in\mathcal{C}}(C_i(t_0) - \overline{C(t_0)})(C_i(t_0 - d) - \overline{C(t_0 - d)})}{\sum_{i\in\mathcal{C}}(C_i(t_0 - d) - \overline{C(t_0 - d)})^2}$$

où $\overline{C(t_0)} = \frac{1}{N}\sum_{i\in\mathcal{C}} C_i(t_0)$ désigne la moyenne des $C_i(t_0)$, et

$$a^* = \overline{C(t_0)} - b^*\overline{C(t_0 - d)}.$$

Jusqu'à présent, on a simplement réécrit $C_i(t)$ différemment. Vu que nous n'avons pas toutes les valeurs, on ne peut pas calculer b^*, mais on peut tout de même l'estimer $\hat{b}$ en prenant la somme uniquement sur les éléments de l'échantillon. On a alors

$$\hat{b}_t = \frac{\sum_{i\in S_t}(C_i(t_0) - \overline{C_{S_t}(t_0)})(C_i(t_0 - d) - \overline{C(t_0 - d)})}{\sum_{i\in\mathcal{C}}(C_i(t_0 - d) - \overline{C(t_0 - d)})^2}$$

où $\overline{C_{S_t}(t_0)} = \frac{1}{n_t}\sum_{i\in S_t} C_i(t_0)$ désigne la moyenne des $C_i(t_0)$ sur l'échantillon S_t. Et de même pour $\hat{a}$,

$$\hat{a}_t = \overline{C_S}(t_0) - \hat{b}_t\overline{C_S}(t_0 - d).$$

Lorsqu'on somme notre égalité 1, on obtient

$$C(t_0) = Na + bC(t_0 - d) + \sum_{i\in\mathcal{C}} U_i.$$

On va négliger le terme $\sum_{i\in\mathcal{C}} U_i$. Donc si on résume, on effectue une régression linéaire entre les vecteurs $(C_i(t_0))_{i\in S_t}$ et $(C_i(t_0 - d))_{i\in S_t}$, et on prend ce modèle pour estimer la somme $\sum_i C_i(t_0)$. L'échantillon est utilisé pour déterminer le modèle en estimant b^*. Ce qui donne l'estimateur pour le total

$$\hat{C}_t(t_0) = N\hat{a} + \hat{b}C(t_0 - d)$$

On peut voir cet estimateur aussi comme un estimateur avec un système de poids dépendant de l'échantillon et des valeurs de la variable bien connus. On a alors

$$w_i^t = \frac{1}{n_t} + (\overline{C(t_0-d)} - \overline{C_S(t_0-d)})\frac{C_i(t_0-d) - \overline{C_S(t_0-d)}}{\sum_{i\in S}(C_i(t_0-d) - \overline{C_S(t_0-d)})^2}$$

et

$$\hat{C}_t(t_0) = \sum_{i\in S_t} w_i^t C_i(t_0)$$

Remarquons aussi que cette méthode ne repose pas sur la validité du modèle, même si les données sont très loin de vérifier la relation affine, l'erreur sera plus grande, certes, mais pas pire qu'une estimation simple sans régression.

3.3.1 Précision

Ici, on fait la supposition que les valeurs reçues viennent d'un sondage simple avec probabilités d'inclusion égales. Ceci n'est pas vrai dans notre modèle en général puisque les distributions de retards peuvent être très différentes, mais l'analyse est beaucoup plus difficile dans la cas de probabilités inégales, et prendre les probabilités d'inclusion en compte dans l'estimation augmente considérablement le coût en calcul de la méthode comme on le verra lorsqu'on analysera les coûts.

L'estimateur $\hat{C}(t_0)$ est légèrement biaisé, et la variance

$$Var(\hat{C}(t_0)) \simeq N^2 \frac{1-\frac{n}{N}}{n}\frac{1}{N-1}\sum_i U_i^2$$

où $\frac{1}{N-1}\sum_i U_i^2$ est la dispersion vraie dans la population de la variable U_i. Donc en quelque sorte, on voit qu'on a remplacé la dispersion de C_i par celle de U_i qui est censée être beaucoup plus petite si le modèle colle bien. En réalité, on a

$$Var(\hat{C}_{regr}(t_0)) \approx (1-\rho^2)Var(\hat{C}_{sond}(t_0))$$

où ρ représente le coefficient de corrélation entre $(C_i(t_0))_{i\in\mathcal{C}}$ et $(C_i(t_0-d))_{i\in\mathcal{C}}$. Pour plus de détail concernant la précision de cet estimateur, voir la partie III.4 de [1] qui concerne l'estimateur par la régression.

3.3.2 Estimation de la précision

On peut également estimer la précision à l'aide de l'échantillon, il suffit de remplacer la dispersion vraie par la dispersion dans l'échantillon. On a donc

$$\hat{V}(\hat{C}_{regr}(t_0)) = N^2 \frac{1-\frac{n}{N}}{n}\frac{1}{n-1}\sum_{i\in S_t}\Big(C_i(t_0) - \hat{a} - \hat{b}C_i(t_0-d)\Big)$$

et donc on peut donner un intervalle de confiance aussi.

3.3.3 Coût et informations nécessaires

Si on applique directement les formules données, on recalcule à chaque instant t les poids pour les estimateurs de la totalité des échantillons de tous les instants passés qui sont concernés par le nouvel échantillon. En effet les poids w_i^t changent également pour les compteurs qui sont déjà arrivés.

Mais il s'avère que l'on n'est pas obligé de re-calculer explicitement ces poids, et que l'on peut faire les calculs de manière incrémentale. On ne re-visite donc pas chaque compteur pour revoir son poids mais on a simplement besoin de garder quelques agrégats qui représentent la précédente estimation.

On définit les variables suivantes, et on garde pour chaque instant t_0 des variables courantes $\alpha_t, \beta_t, \gamma_t$ et δ_t qui évoluent dans le temps au fur et à mesure que l'échantillon grandit.

- $\alpha_t = \sum_{i \in S_t} C_i(t_0)$
- $\beta_t = \sum_{i \in S_t} C_i(t_0 - d)$
- $\gamma_t = \sum_{i \in S_t} C_i(t_0 - d)^2$
- $\delta_t = \sum_{i \in S_t} C_i(t_0 - d) C_i(t_0)$

et on a les relations de récurrences

- $\alpha_{t+1} = \alpha_t + \sum_{i \in S_{t+1} - S_t} C_i(t_0)$
- $\beta_{t+1} = \beta_t + \sum_{i \in S_{t+1} - S_t} C_i(t_0 - d)$
- $\gamma_{t+1} = \gamma_t + \sum_{i \in S_{t+1} - S_t} C_i(t_0 - d)^2$
- $\delta_{t+1} = \delta_t + \sum_{i \in S_{t+1} - S_t} C_i(t_0 - d) C_i(t_0)$

qui nous permettent de mettre à jour efficacement toutes ces quantités. Maintenant pour mettre à jour notre estimateur, en notant $d_t = \gamma_t - 2\beta_t \frac{\alpha_t}{n_t} + \frac{\beta_t^2}{n_t} = \sum_{i \in S_t} (C_i(t_0 - d) - \overline{C_{S_t}(t_0 - d)})^2$

$$\hat{C}_t(t_0) = \frac{N}{n_t}\alpha_t + N(\overline{C(t_0 - d)} - \frac{1}{n_t}\beta_t)\frac{1}{d_t}(\delta_t - \frac{\beta_t}{n_t}\alpha_t).$$

On peut donc bien faire le calcul de manière incrémentale et on n'effectue qu'un nombre d'opérations linéaire en le nombre de valeurs reçues.

Mais cette méthode nécessite d'avoir accès aux données détaillées du passé, et que l'on ait une garantie que dans un délai d, toutes les données seront disponibles.

3.3.4 Variantes et extensions de la méthode

Plusieurs extensions sont possibles pour cette méthode d'estimation :

- Il est possible d'écrire une relation linéaire avec plus de variables, par exemple

$$C_i(t_0) = a_0 + a_1 C_i(t_0 - d_1) + a_2 C_i(t_0 - d_2) + \cdots + a_p C_i(t_0 - d_p) + U_i.$$

  On espère que prendre en compte plus d'éléments dans le passé nous donnera une précision plus grande dans la prévision, ce qui va être certainement le cas. Remarquons aussi que les données qu'on utilise ne doivent pas nécessairement être des consommations passées mais peuvent correspondre à n'importe quelle variable qui dépend

du compteur. Un candidat potentiel est la température, si on dispose d'une valeur pour la température par zone géographique (donnée par une source extérieure par exemple) il est possible de prendre en compte cette variable.
Le problème est que plus on augmente le nombre de variables prises en compte, plus on augmente le temps de calcul, en effet avec k variables, on est amené à multiplier des matrices $p \times N$ ce qui est coûteux en général. Il faut donc trouver un compromis entre temps de calcul et précision. De plus, ceci va grandement compliquer l'expression incrémentale qui permet à cette méthode de mettre à jour efficacement son estimation à l'arrivée de nouveaux échantillons.

- Intégrer les probabilités d'inclusion est également possible, mais il ne sera plus possible dans ce cas de mettre à jour les estimations de manière incrémentale. Donc cette possibilité ne semble pas très efficace.
- On a supposé que l'on fait un régression par rapport aux données du temps $t_0 - d$. Ceci non seulement nécessite le fait que toutes les valeurs du temps $t_0 - d$ soient arrivées, mais aussi que il faut les retenir en détail. Donc à ce problème on peut proposer deux solutions :

  1. Soit faire la régression par rapport à un jour type (ou une semaine, un mois type) fixé à l'avance. Ces données types peuvent être mises à jour manuellement de temps en temps pour prendre en compte des phénomènes de plus longue durée. Cette méthode a le mérite de ne pas faire de supposition sur les retards maximums. Par exemple, si les retards peuvent atteindre 1 mois, il faut prendre d au moins 1 mois.
  2. Soit faire un calage par rapport à une donnée approximative, c'est à dire par exemple de construire un résumé de petite taille qui contient une approximation des consommations individuelles pour l'instant $t_0 - d$. Typiquement, on stockerait le vecteur $(C_i(t_0 - d))_{i \in \mathcal{C}}$ dans un sketch count-min (voir [3] pour une description). Ceci nous permettrait de calculer les valeurs $C_i(t_0 - d)$ à $\epsilon C(t_0 - d)$ près. Le problème est que ces erreurs vont s'ajouter, et donc cette approche ne semble pas très prometteuse.

Récapitulons les diverses caractéristiques de chaque méthode :

4 Simulations

4.1 Comparaison

Pour comparer nos méthodes, il faut fixer les paramètres de chaque méthode. On va donc tester les méthodes suivantes :

1. La méthode par sondage simple sans partition et avec une partition contenant 5 classes. Les données à notre disposition représentent déjà une classe assez restreinte de clients, il semble donc qu'à l'échelle du système grandeur réelle, une classe ne sera pas plus fine que les données à notre disposition. Néanmoins, pour comparer

	Sondage simple	**Probabilités inégales**	**Régression**
Coût calcul par valeur	constant	$\Theta(N)$	constant
Informations	-/une partition	distribution retards	consommations individuelles
Condition de précision	clients identiques	consommations proportionnelles aux probabilités d'inclusion	forte corrélation entre consomation t et $t-d$

TAB. 1 – Comparaison des méthodes

la méthode par sondage simple aux autres méthodes, il est légitime d'introduire une utilisation de la consommation passée, et donc on peut se permettre une petite classification faite sur la base des consommations d'un certain jour fixé dans le passé. Les classes définies par des intervalles de consommations choisis de sorte qu'il n'y ait pas de classes ayant peu de clients.

2. La méthode avec des probabilités inégales est appliquée directement comme décrit dans la section correspondante.
3. Concernant l'estimateur par la régression, pour que cette méthode soit réalisable à plus grande échelle, on choisit de simuler le cas avec une régression par rapport à une demi-semaine type dans le passé, et cet instant du passé est fixé et ne glisse pas avec l'instant t_0 que l'on cherche à estimer.

4.2 Simulations : données

Les données à notre disposition concernent 1000 clients ayant un même tarif pour lesquels on a les mesures de consommations à un pas de 10 minutes pendant une année. Seules les données de consommation sont disponibles, les retards doivent donc être générés pour simuler les méthodes d'estimations. Ceux-ci sont générés suivant des lois normales ou log-normales avec une moyenne et une variance fixées pour chaque compteurs. La loi log-normale modélise le fait que des très gros retards sont probables. Les retards des compteurs sont indépendants. Deux familles de simulations sont effectuées :

- les compteurs sont identiques, c'est à dire qu'ils ont tous les mêmes moyennes et variances
- les compteurs sont différents, et les caractéristiques dépendent des consommations typiques du client.

Les tests exécutés simulent uniquement pendant une durée d'une semaine les 1000 clients pour lesquels nous avons des mesures.

La stratification choisie donne lieu à 5 strates avec des effectifs semblables.

4.2.1 Compteurs identiques

Pour les compteurs identiques on effectue deux séries de simulations.

1. Moyenne : 12 heures, écart type : 3 heures.
2. Moyenne : 12 heures, écart type : 6 heures.

4.2.2 Compteurs différents, retards corrélés aux consommations

On peut fixer des moyennes et variances qui sont fonctions de la classe du client, on peut pour cela utiliser les strates définies pour l'estimateur stratifié. On effectue la simulation avec les valeurs suivantes

- Classe 1 : moyenne : 24 h, écart type : 12 h
- Classe 2 : moyenne : 12 h, écart type : 4 h
- Classe 3 : moyenne : 6 h, écart type : 2 h
- Classe 4 : moyenne : 2 h, écart type : 40 minutes
- Classe 5 : moyenne : 30 minutes, écart type : 10 minutes

4.3 Résultats

Les résultats de ces simulations sont présentés dans le dossier `resultats`.

De cette série de simulations il ressort que l'estimateur par la régression est le plus adapté, pour la précision de l'estimation et son efficacité dans les différents cas de figures.

On peut aussi conclure qu'utiliser les probabilités d'inclusion n'est pas une bonne idée : le temps de calcul augmente considérablement et la précision n'est pas bonne. En effet, sauf lorsque les retards sont liés aux consommations, c'est le plus mauvais estimateur. Ceci vient sans doute du fait que sur les données, la variance de cet estimateur est très grande.

Il serait intéressant d'évaluer le comportement de chacune des méthodes en présence d'anomalies. Des anomalies peuvent être des données non reçues ou des valeurs inexactes de consommation. Pour des valeurs non reçues, on voudrait que notre estimateur soit très précis lorsque le nombre de valeurs manquantes est très petit. On a pu observer que les estimateurs par sondage et par la régression sont tous les deux assez précis lorsqu'il manque moins de 1% des données. Pour les valeurs aberrantes, on voudrait que l'estimateur ne soit pas trop influencé par ces valeurs. Ces simulations n'ont pas été effectuées mais il semble que l'estimateur de la regression serait le plus adapté dans ce cadre également.

5 Perspectives

Cette étude visait à regarder ce qui se passe en présence de retards dans l'acheminement des données. Quelques perspectives pour poursuivre ce travail :

- Suite de l'analyse
    1. Comment repérer automatiquement des données aberantes et les rectifier ?

2. Est ce que la prise en compte d'autres variables permettrait d'améliorer la méthode par régression ?
3. On a vu que le coefficient de corrélation entre les données à estimer et les données passées jouent un role important. Est-il possible de choisir de façon dynamique la date du passé (le d de la méthode par régression) pour optimiser la précision de l'estimateur ?
4. La gestion des retards moyen permet aussi de ne pas surcharger le système. On pourrait donc définir un problème d'optimisation qui maximiserait la précision de notre estimateur tout en maintenant une certaine contrainte sur le nombre de compteurs dont on reçoit la valeur à un instant donnée.
5. Est ce que si on programme les compteurs à envoyer leur valeur de cette manière est une bonne idée pour éviter de surcharger le réseau ?

– Questions plus générales

1. Pour ce modèle, est ce que combiner des méthodes de prévision avec un sondage est réalisable de façon générale, par exemple la prévision donnant une distribution a priori sur la consommation (approche bayesienne) ?
2. Les distributions de retard ont été choisies comme des lois normales ou log-normales de façon un peu arbitraire. Peut-on définir un modèle plus réaliste des retards ?
3. Concernant la quantité de communication dans le réseau, si on suppose que l'on peut faire des requêtes sur les compteurs, comment établir un protocole qui assure une certaine précision tout en ne surchargeant pas le réseaux (voir [4]) ? Comment des retards et des omissions influencent-elles ce protocole ?
4. On peut également se poser des questions de sécurité en présence de compteur "malicieux". Des compteurs qui sont contrôlés par un adversaire qui cherche à fausser les agrégats, dans quelle mesure est-il possible de calculer des agrégats dans ce cadre ? (Des éléments de réponse sont présentés dans [6]).

Références

[1] Pascal Ardilly. *Les techniques de sondages*. Editions TECHNIP, 2006.

[2] Levente Buttyán, Peter Schaffer, and István Vajda. *Resilient Aggregation : Statistical Approaches*, chapter Chapter 10. in N.P.Mahalik (ed.) : Sensor Networks and Configuration, Springer, 2006.

[3] G. Cormode and S. Muthukrishnan. An improved data stream summary : The count-min sketch and its applications. In *Proceedings of Latin American Theoretical Informatics (LATIN)*, pages 29–38, 2004.

[4] Graham Cormode and Minos Garofalakis. Sketching streams through the net : Distributed approximate query tracking. *ACM Transactions on Database Systems (TODS)*, 2008.

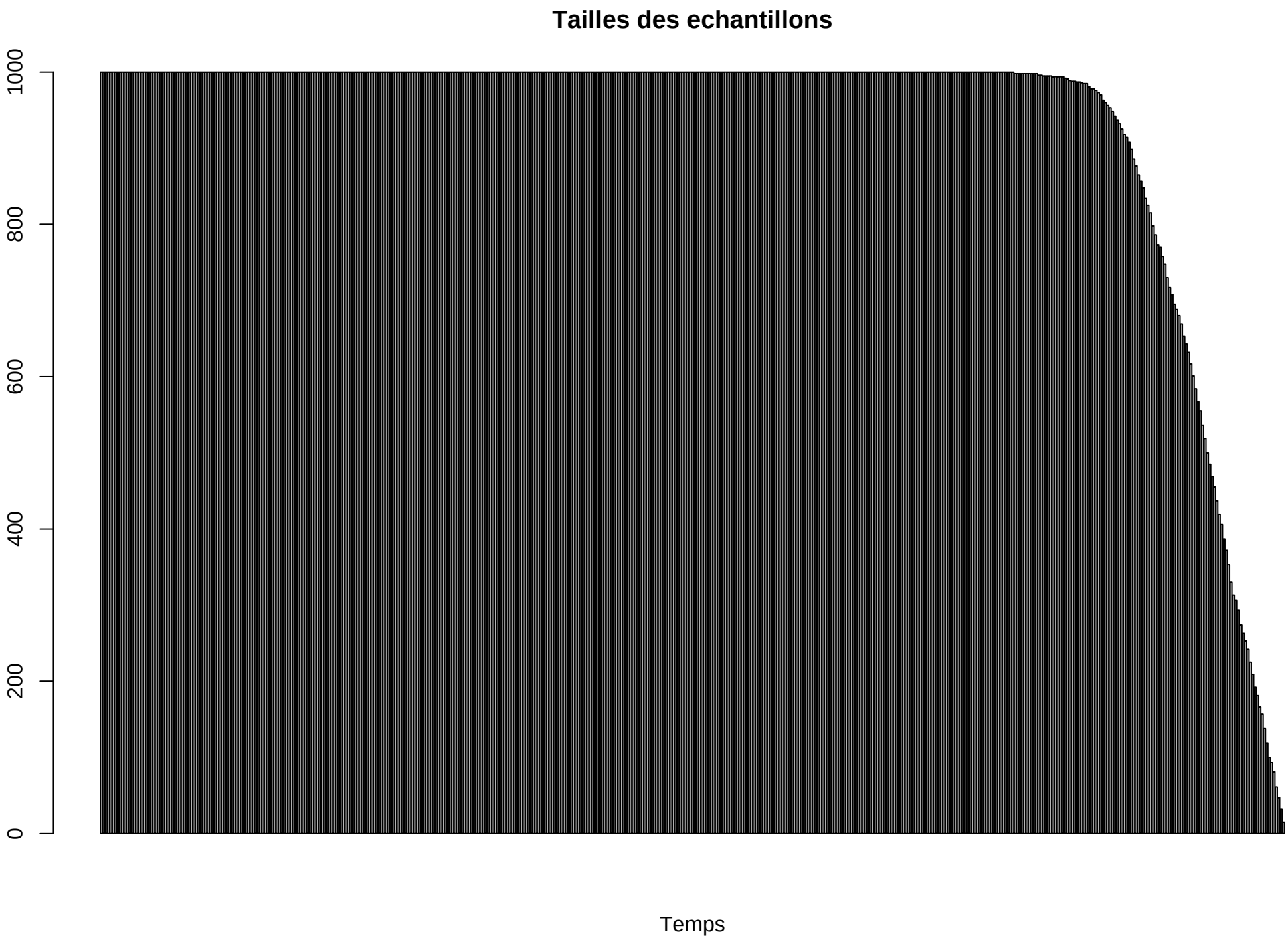
[5] Suman Nath, Phillip B. Gibbons, Srinivasan Seshan, and Zachary Anderson. Synopsis diffusion for robust aggregation in sensor networks. *ACM Trans. Sen. Netw.*, 4(2) :1–40, 2008.

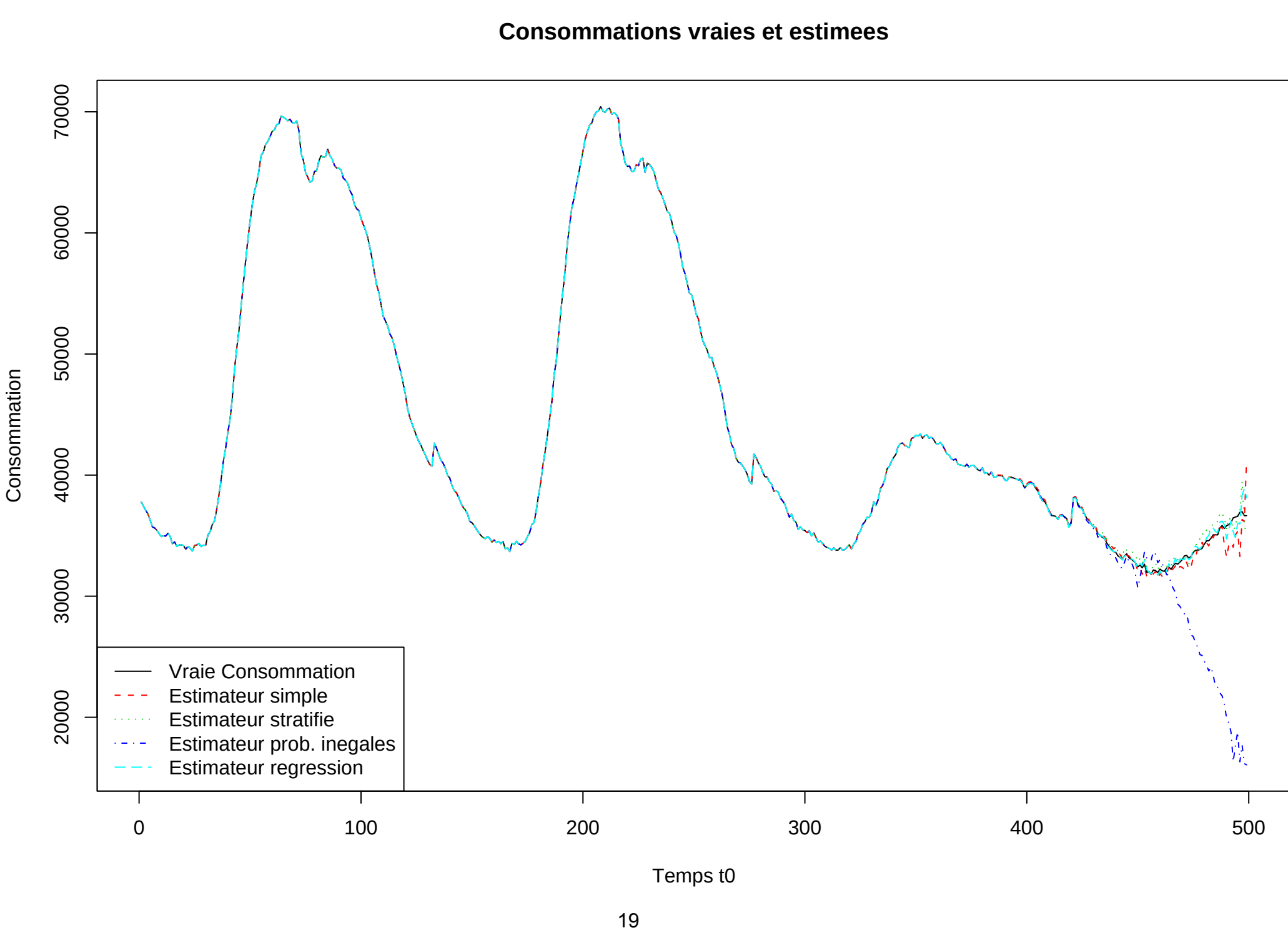
[6] David Wagner. Resilient aggregation in sensor networks. In *SASN*, pages 78–87, 2004.

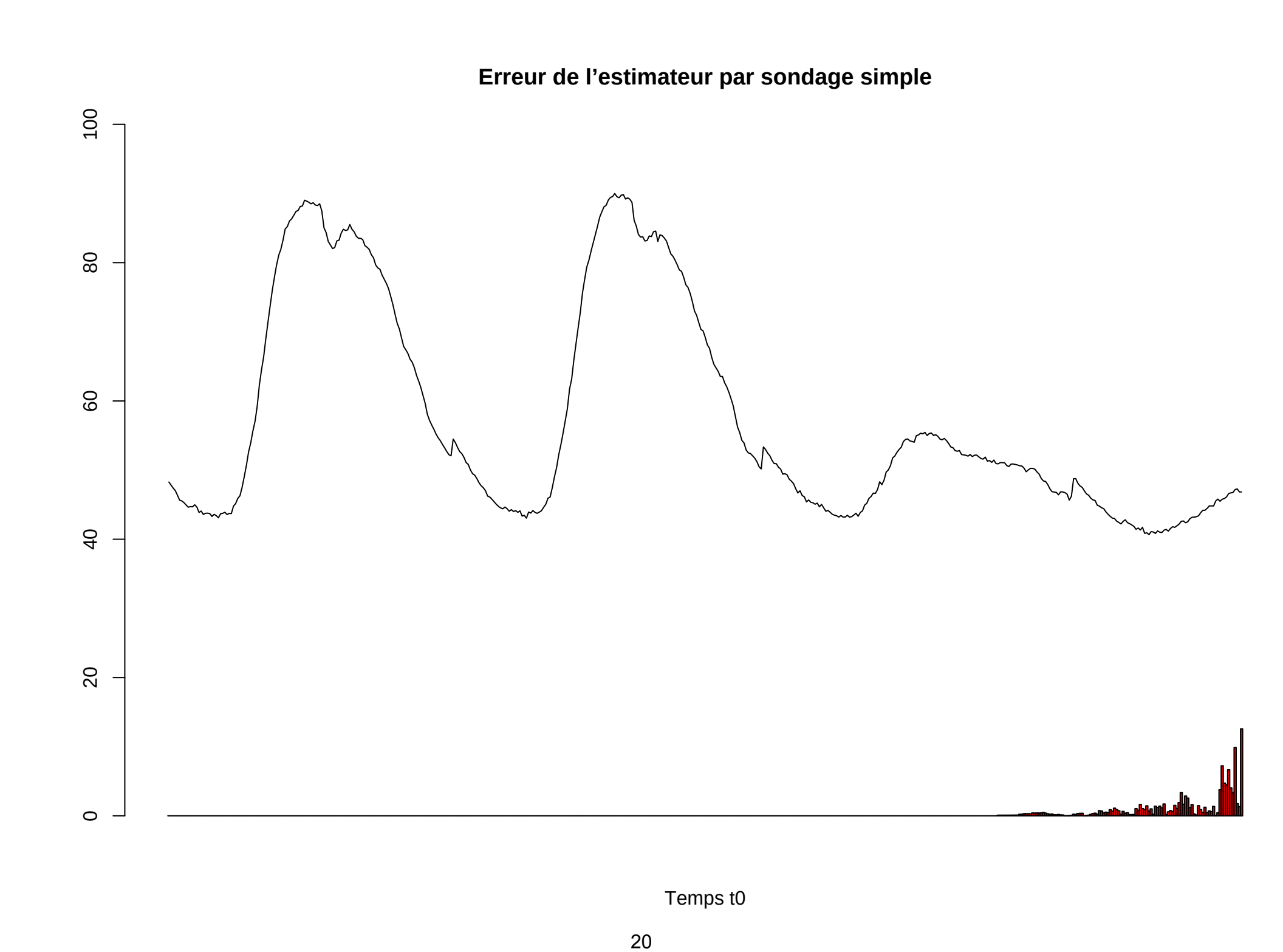
RESULTATS DES SIMULATIONS

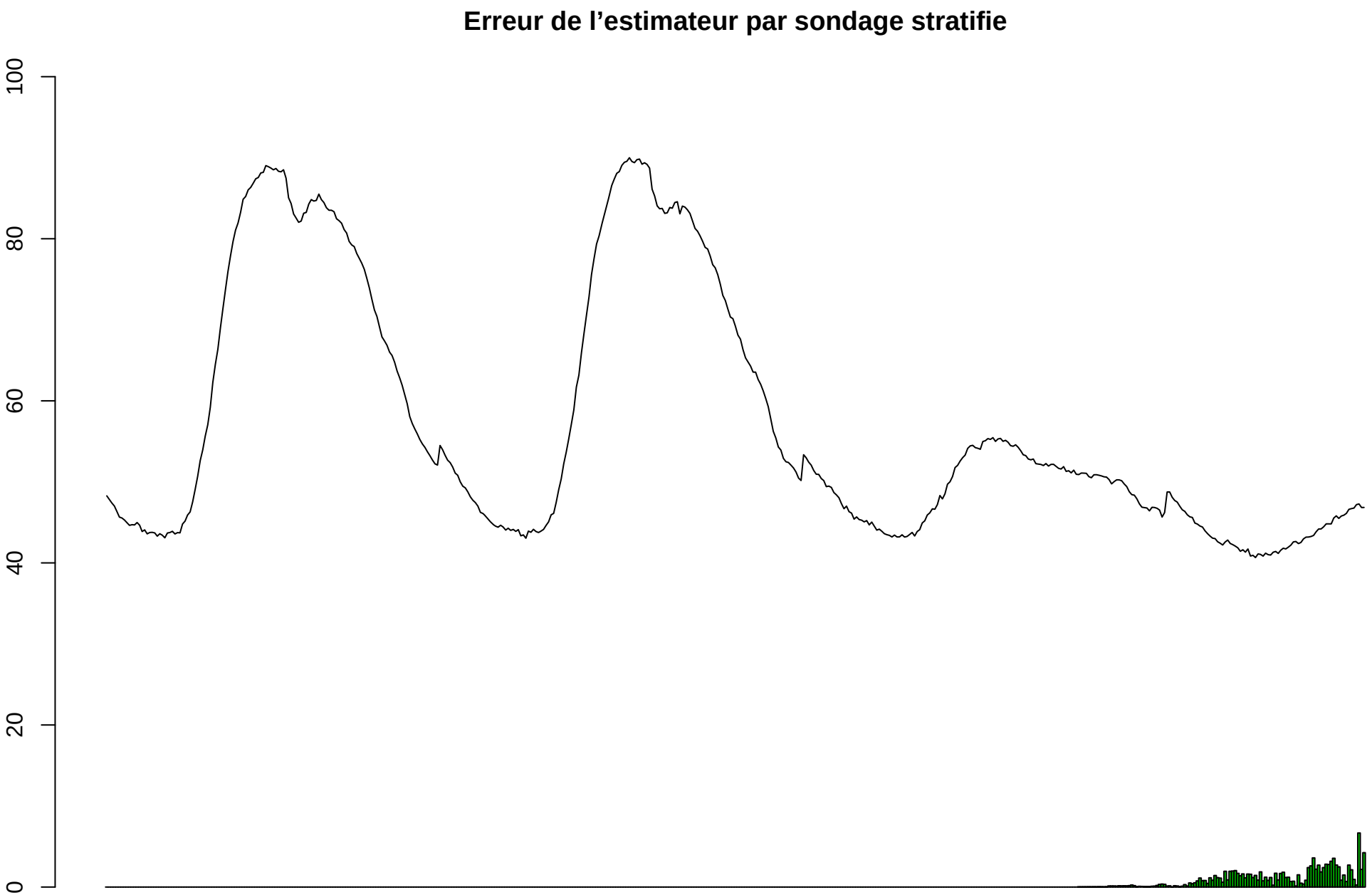
Loi Log-normale pour les retards

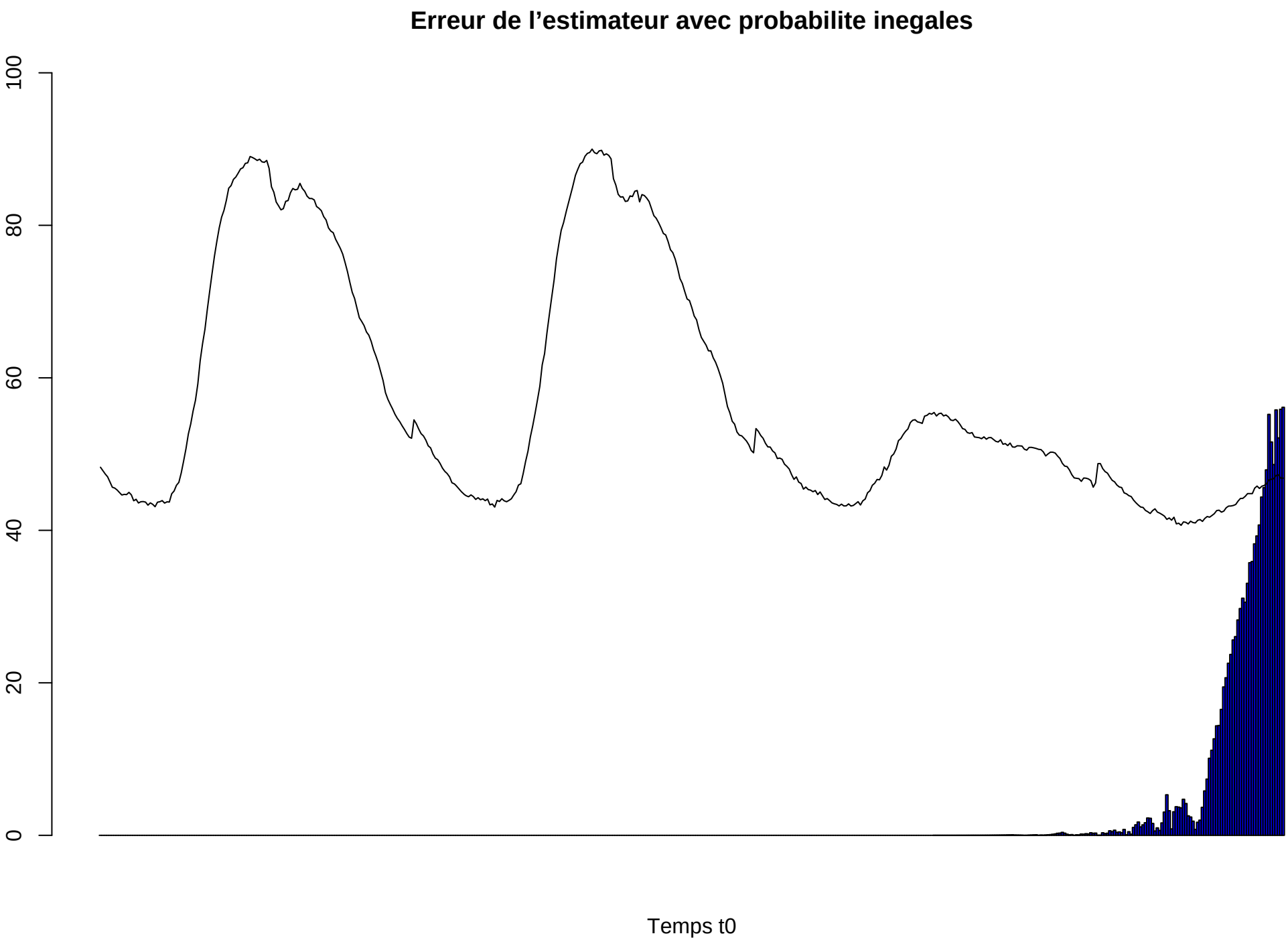
Tous les compteurs suivent la même loi : moyenne 12h, écart-type 6h

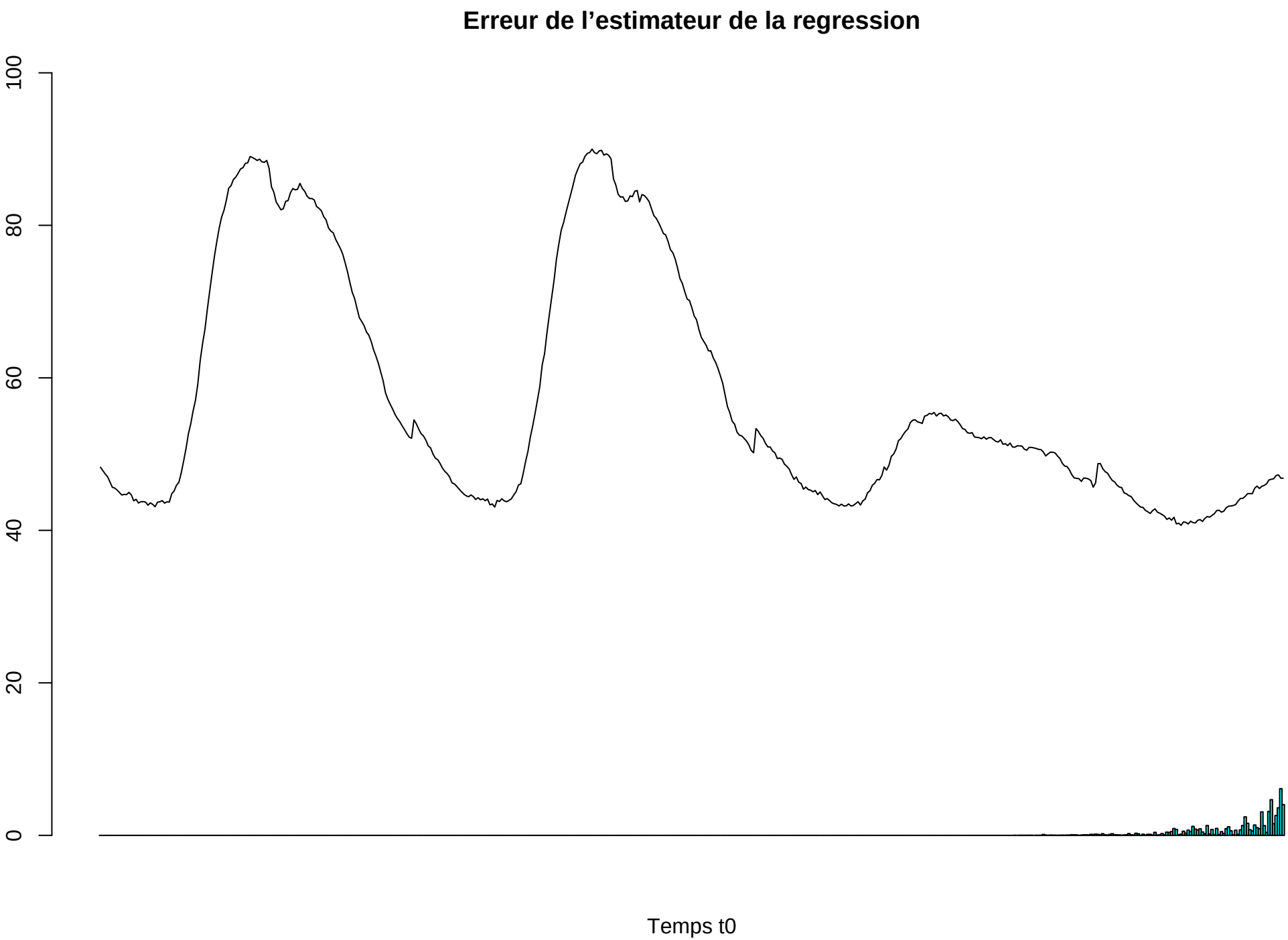

Tailles des echantillons
0
200
400
600
800
1000
Temps

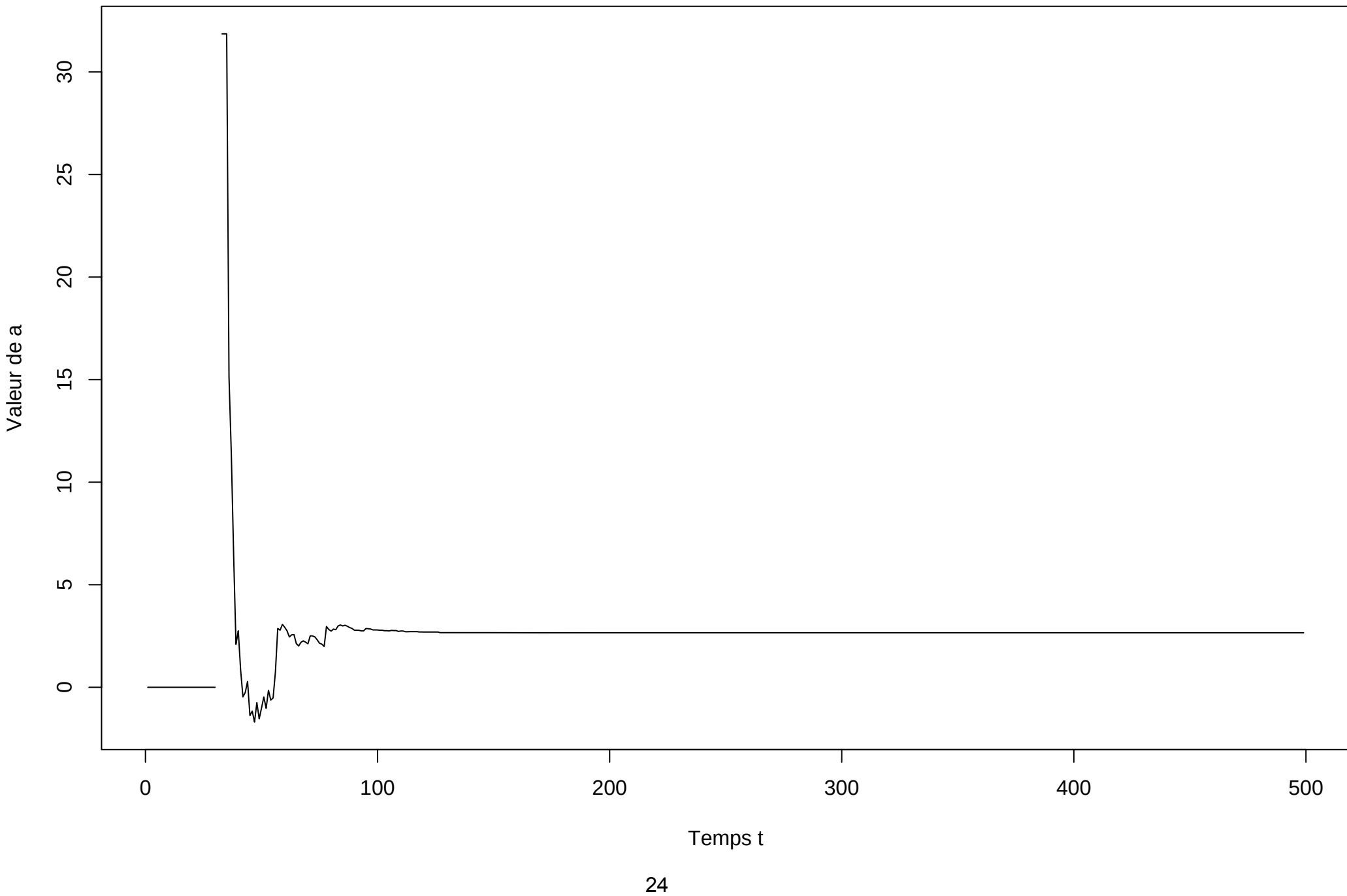

Consommations vraies et estimees
Consommation
70000
60000
50000
40000
30000
20000
0
100
200
300
400
500
Temps t0
Vraie Consommation
Estimateur simple
Estimateur stratifie
Estimateur prob. inegales
Estimateur regression

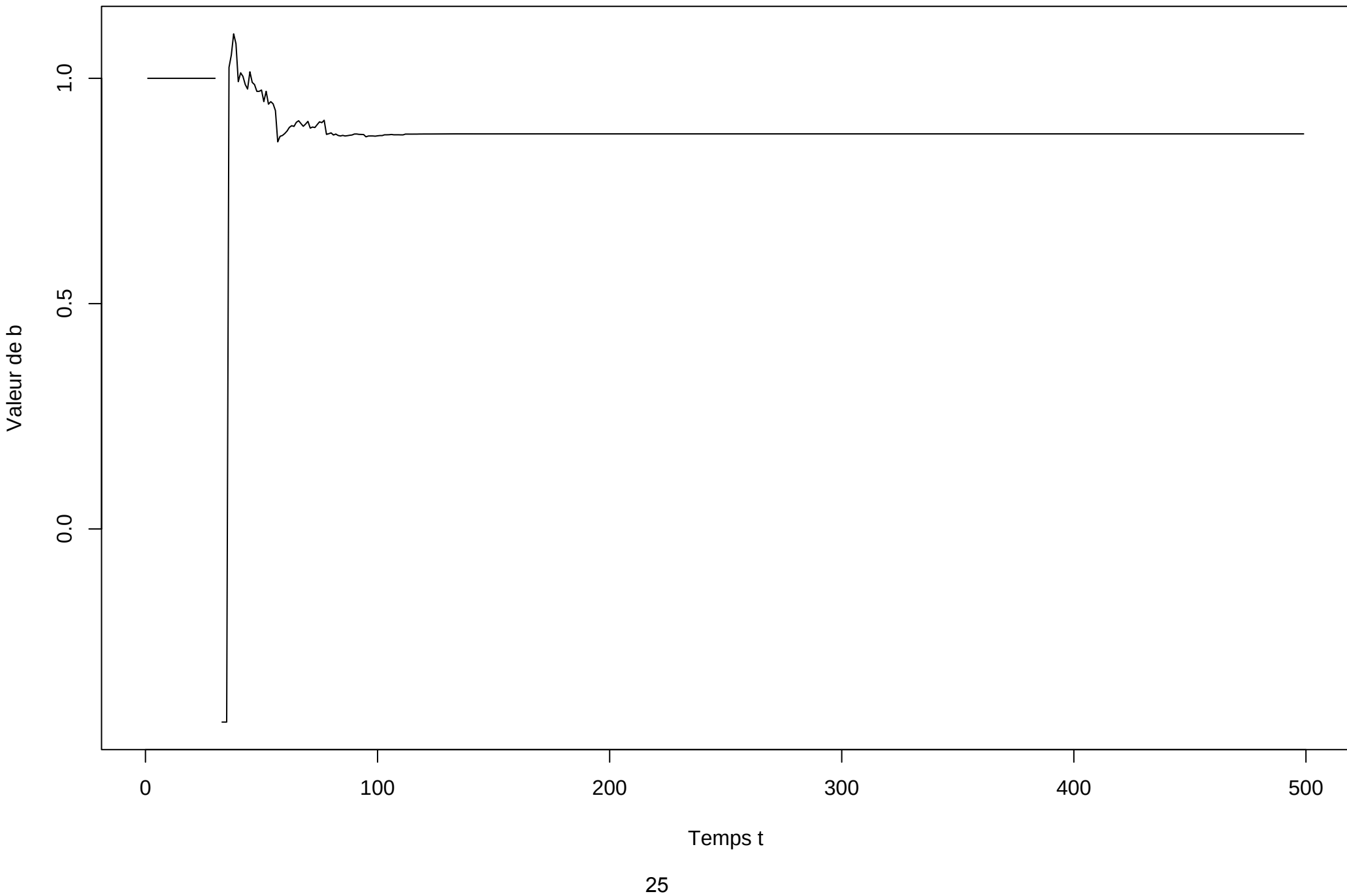

Erreur de l’estimateur par sondage simple
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur par sondage stratifie
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur avec probabilite inegales
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur de la regression
100
80
60
40
20
0
Temps t0

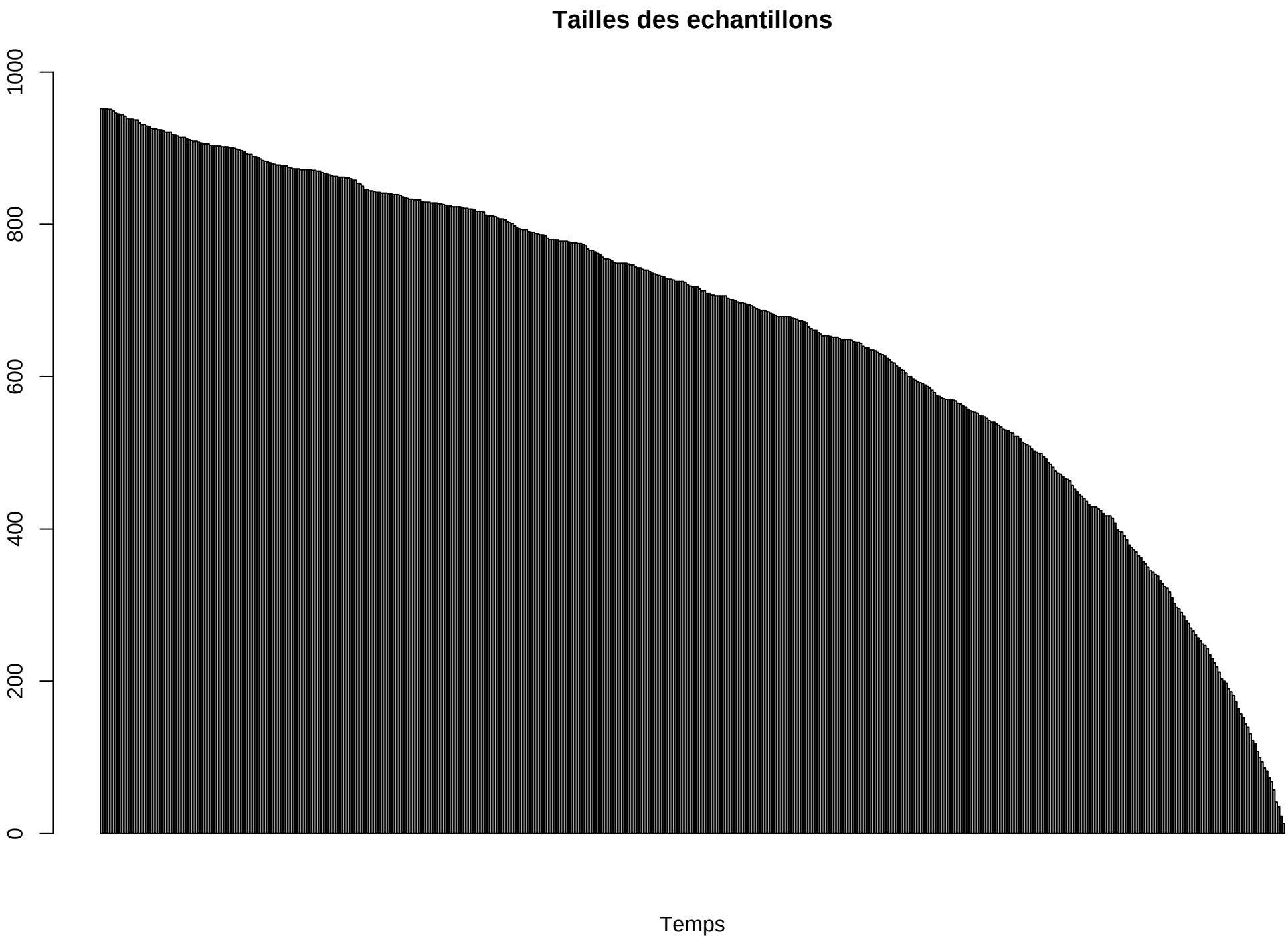

Valeur de a
0
5
10
15
20
25
30
0
100
200
300
400
500
Temps t

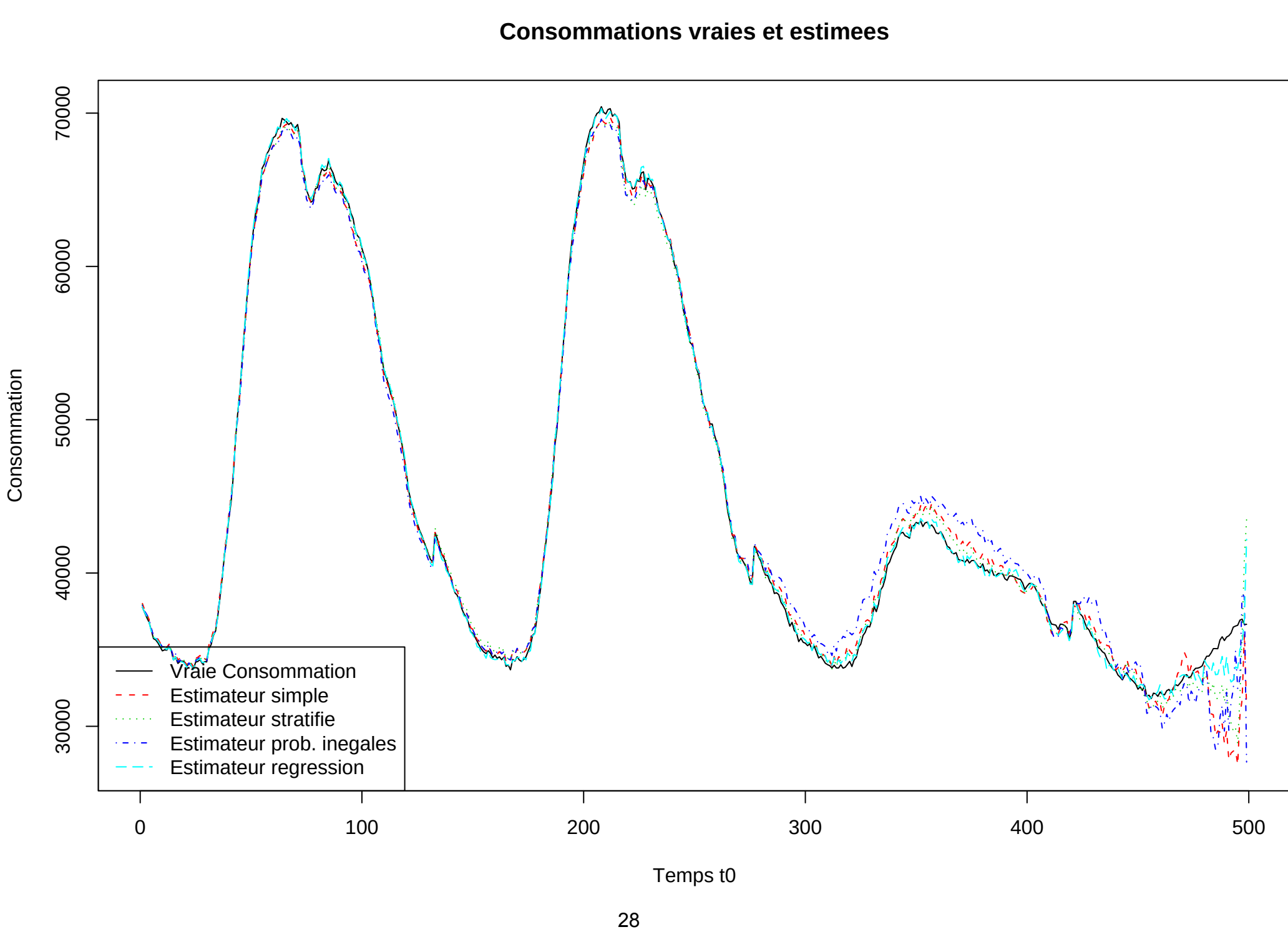

1.0
0.5
0.0
Valeur de b
0
100
200
300
400
500
Temps t

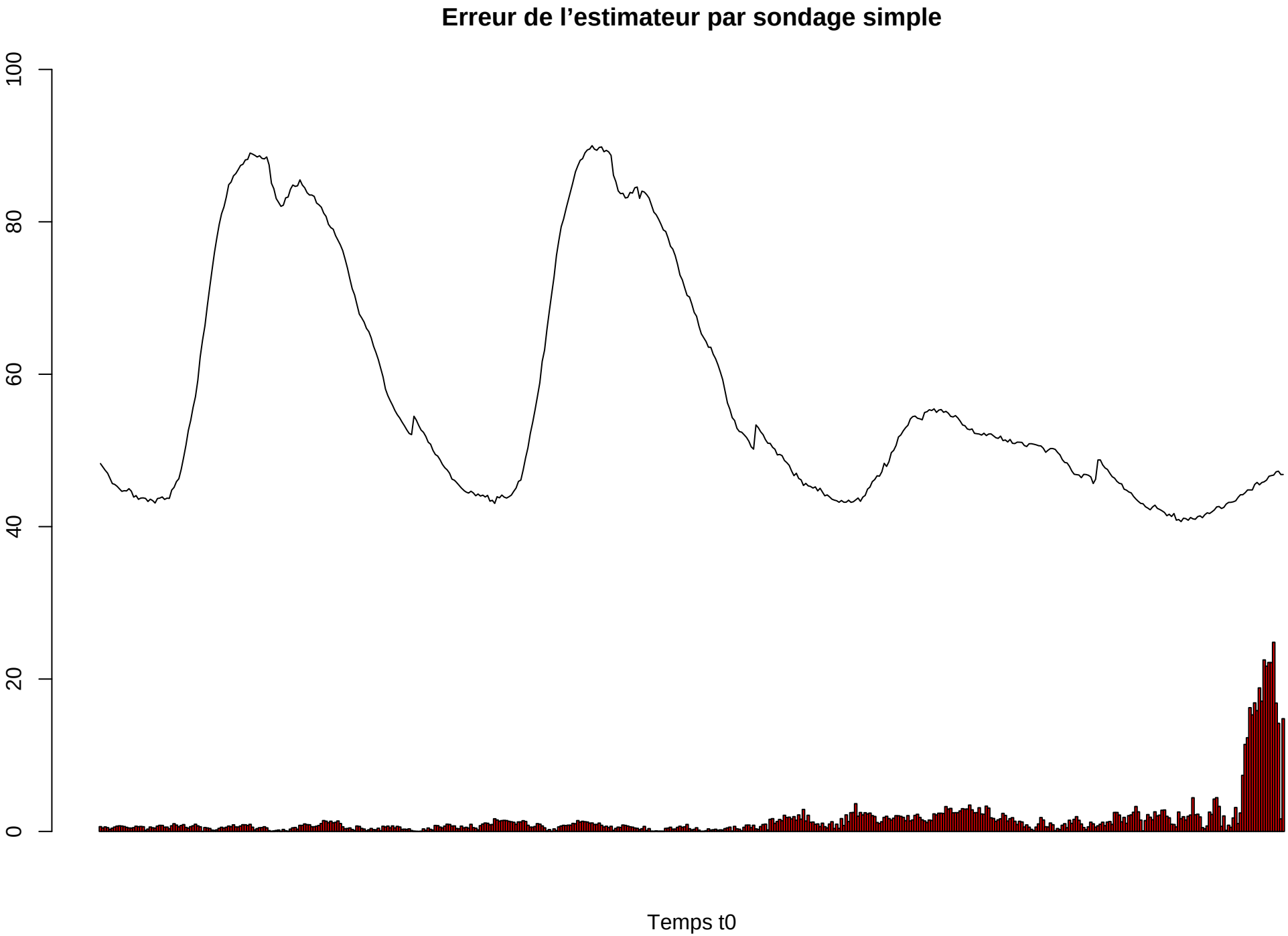
RESULTATS DES SIMULATIONS

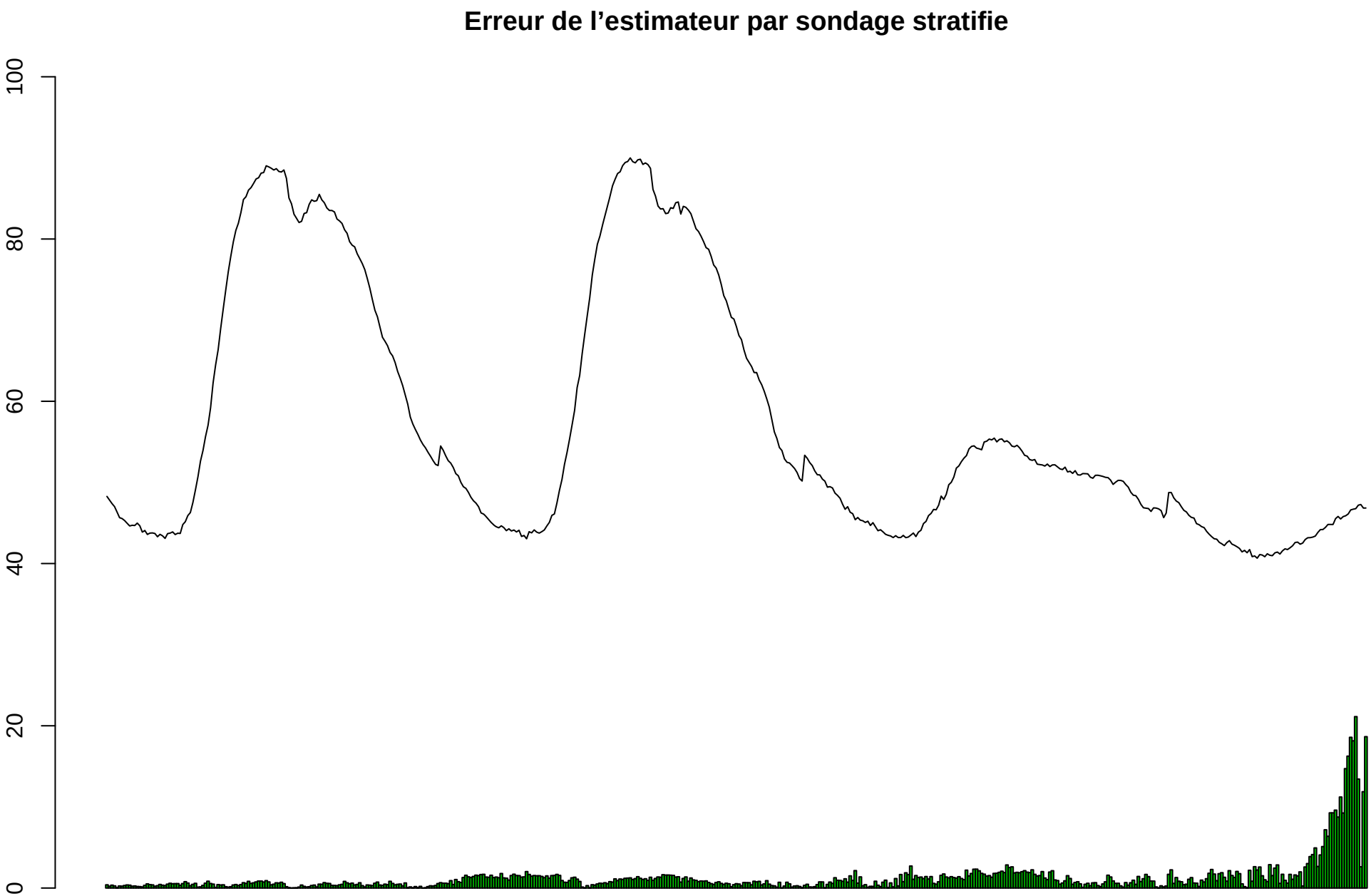
Loi Log-normale pour les retards

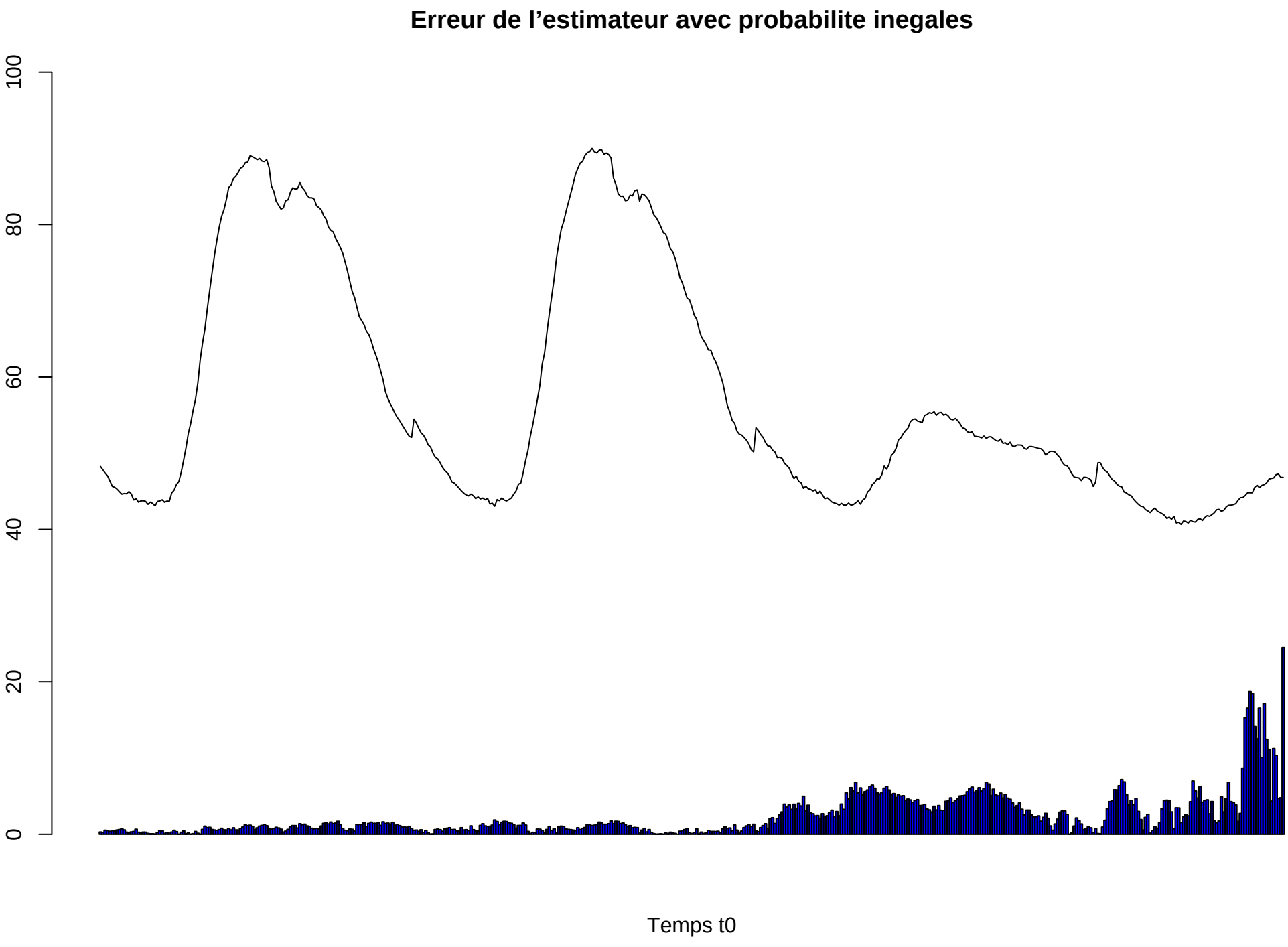
Tous les compteurs suivent la même loi : moyenne 12h, écart-type 24h

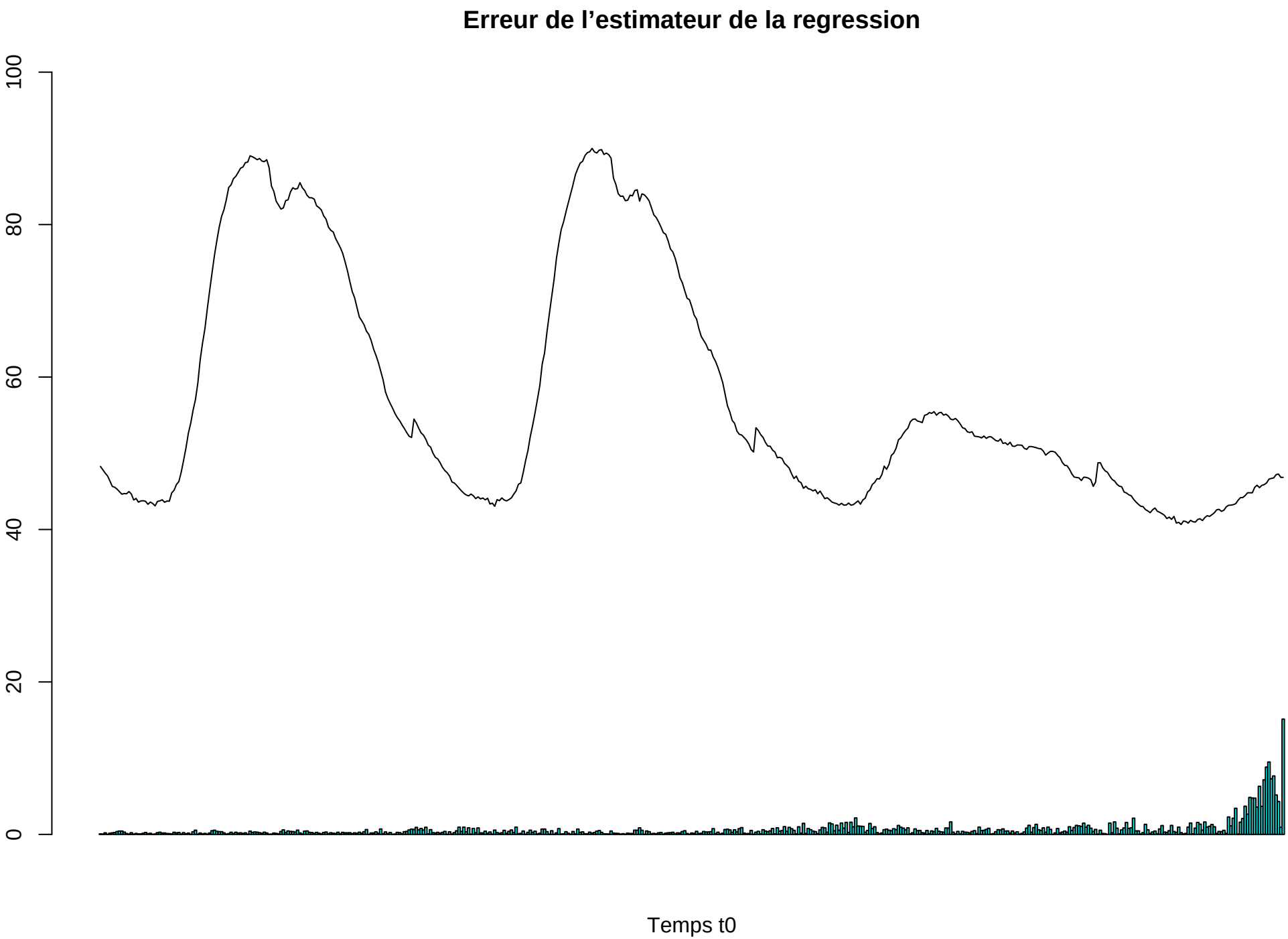

Tailles des echantillons
1000
800
600
400
200
0
Temps

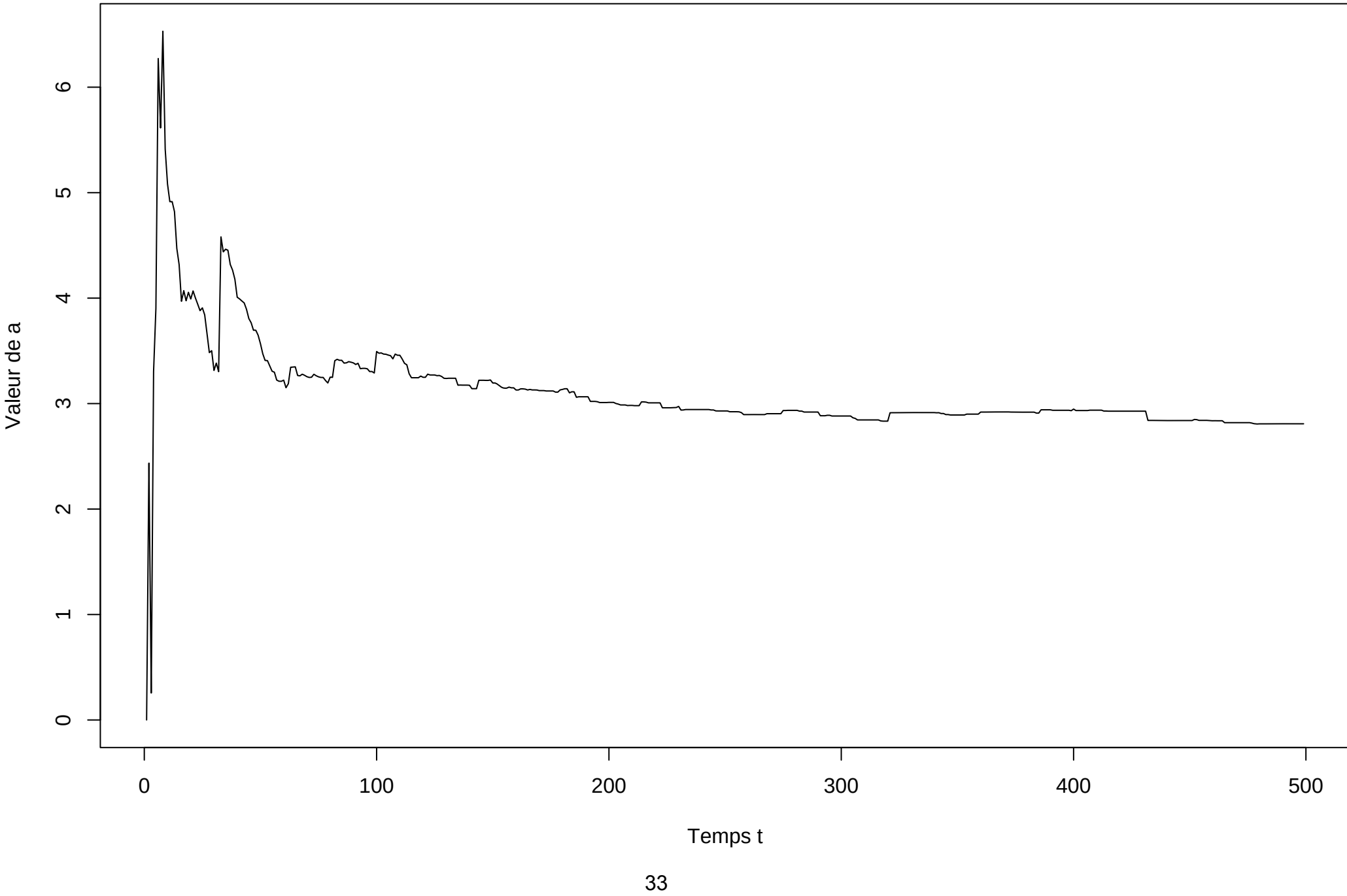

Consommations vraies et estimees
Consommation
70000
60000
50000
40000
30000
0
100
200
300
400
500
Temps t0
Vraie Consommation
Estimateur simple
Estimateur stratifie
Estimateur prob. inegales
Estimateur regression

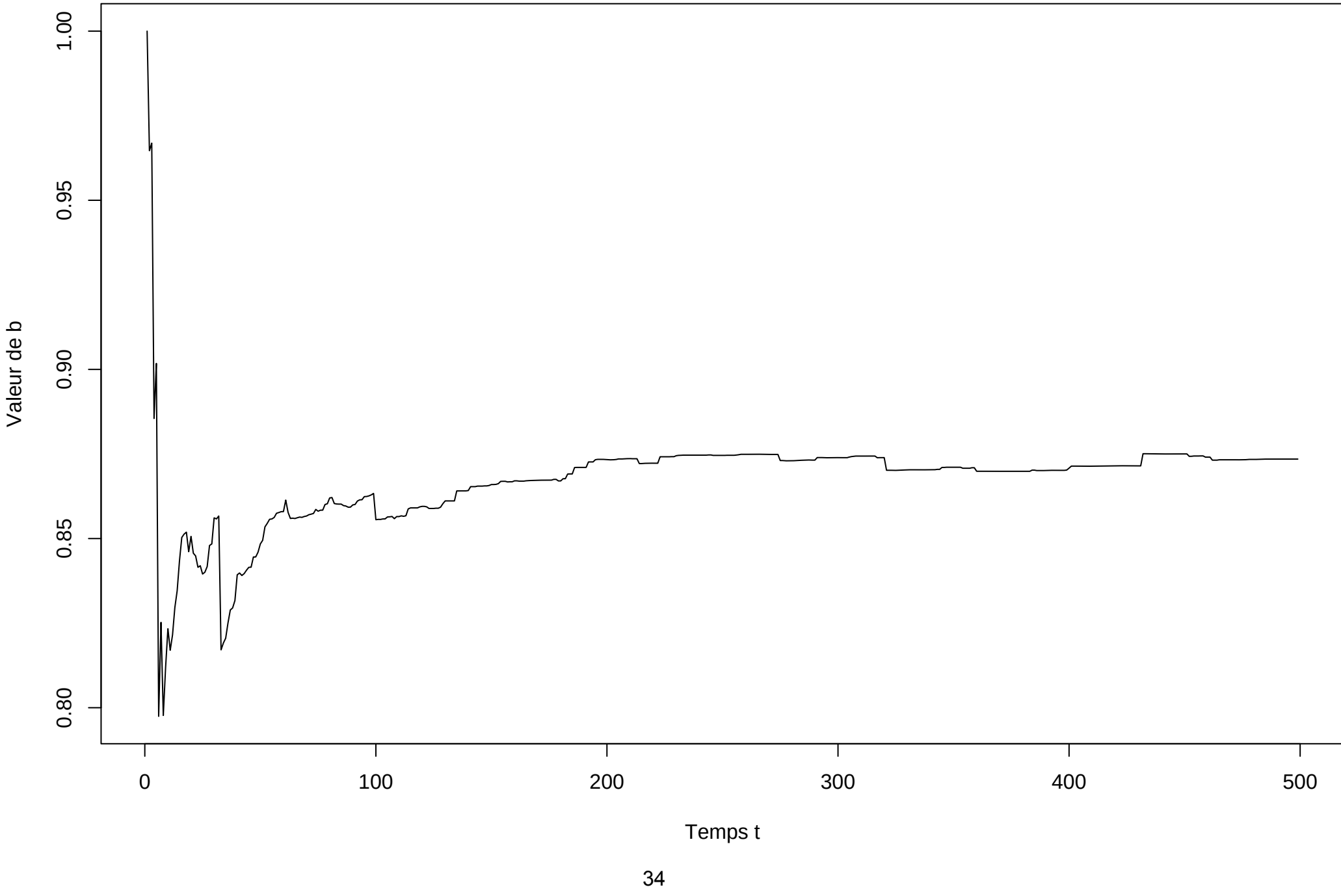

Erreur de l’estimateur par sondage simple
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur par sondage stratifie
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur avec probabilite inegales
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur de la regression
100
80
60
40
20
0
Temps t0


Valeur de a
0
1
2
3
4
5
6
0
100
200
300
400
500
Temps t

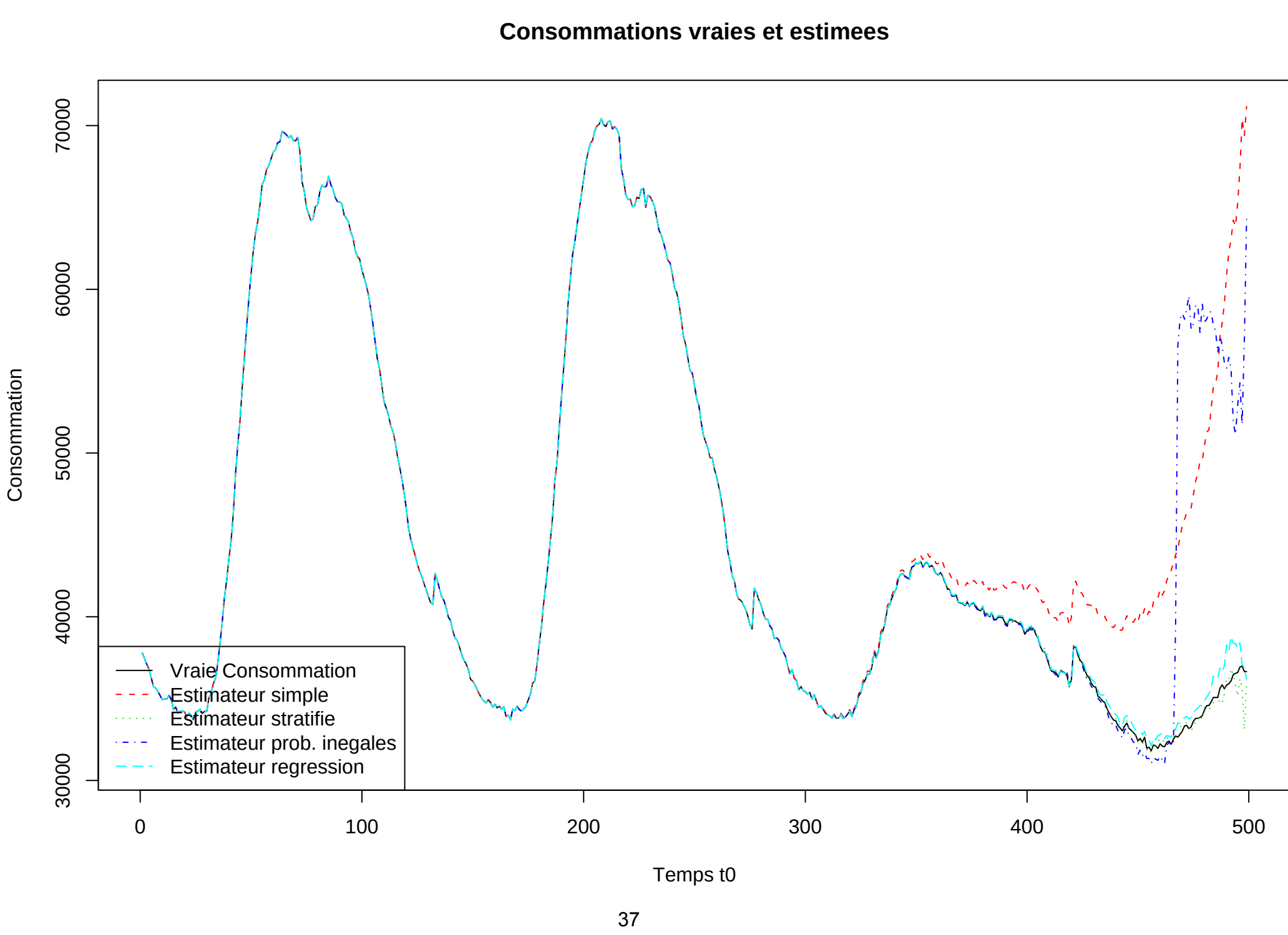

1.00
0.95
0.90
0.85
0.80
Valeur de b
0
100
200
300
400
500
Temps t

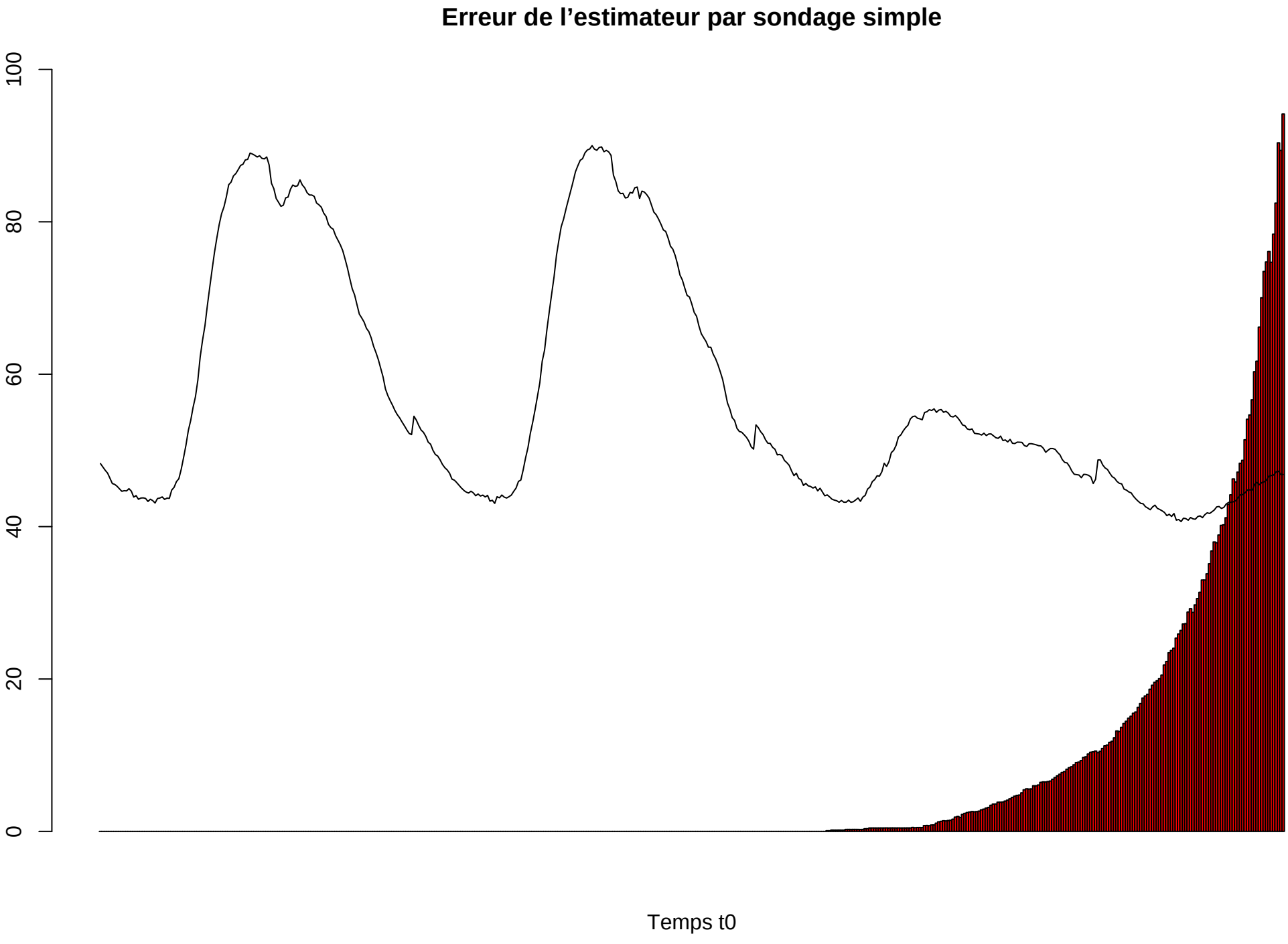
RESULTATS DES SIMULATIONS

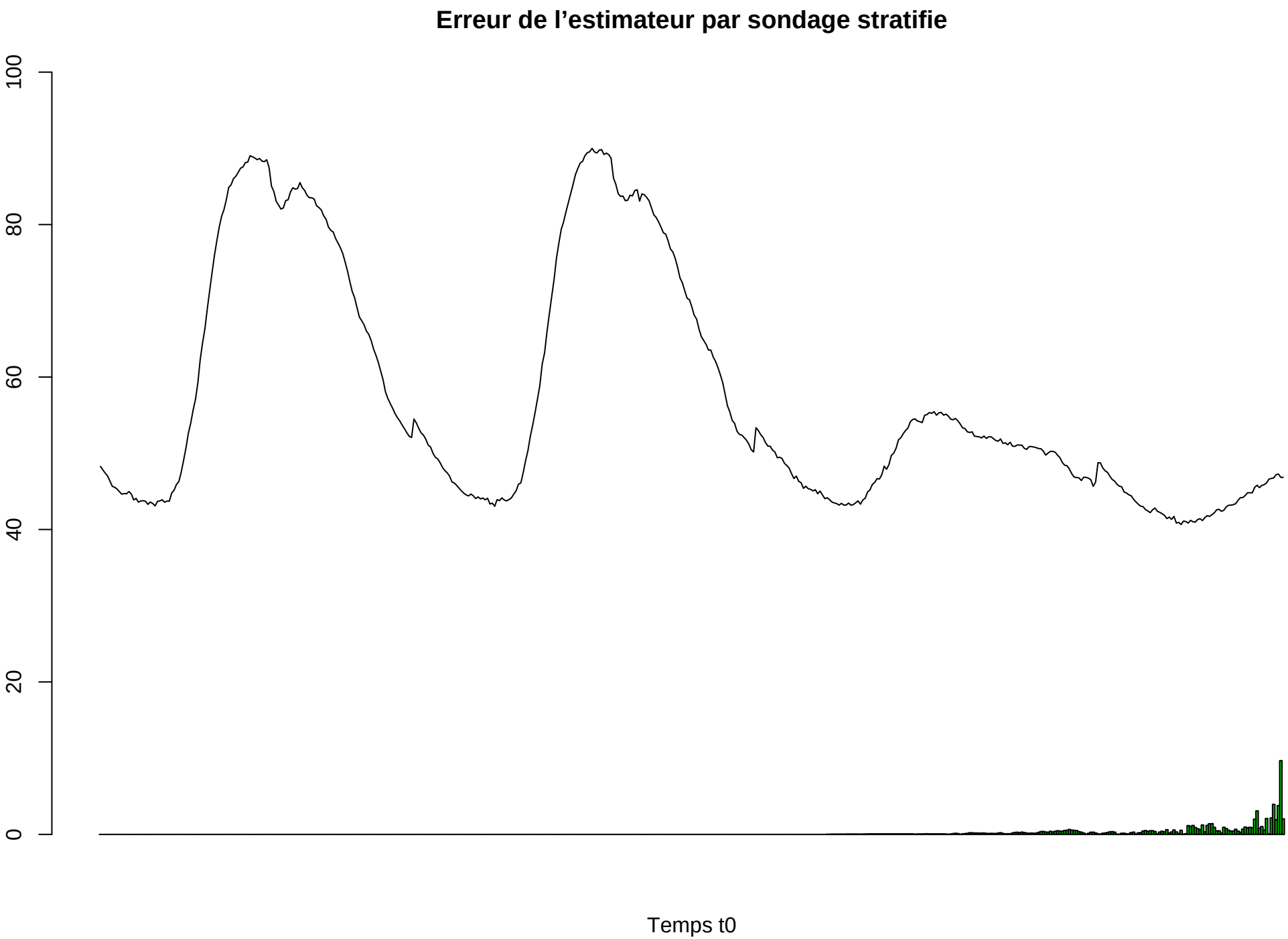
Loi Log-normale pour les retards

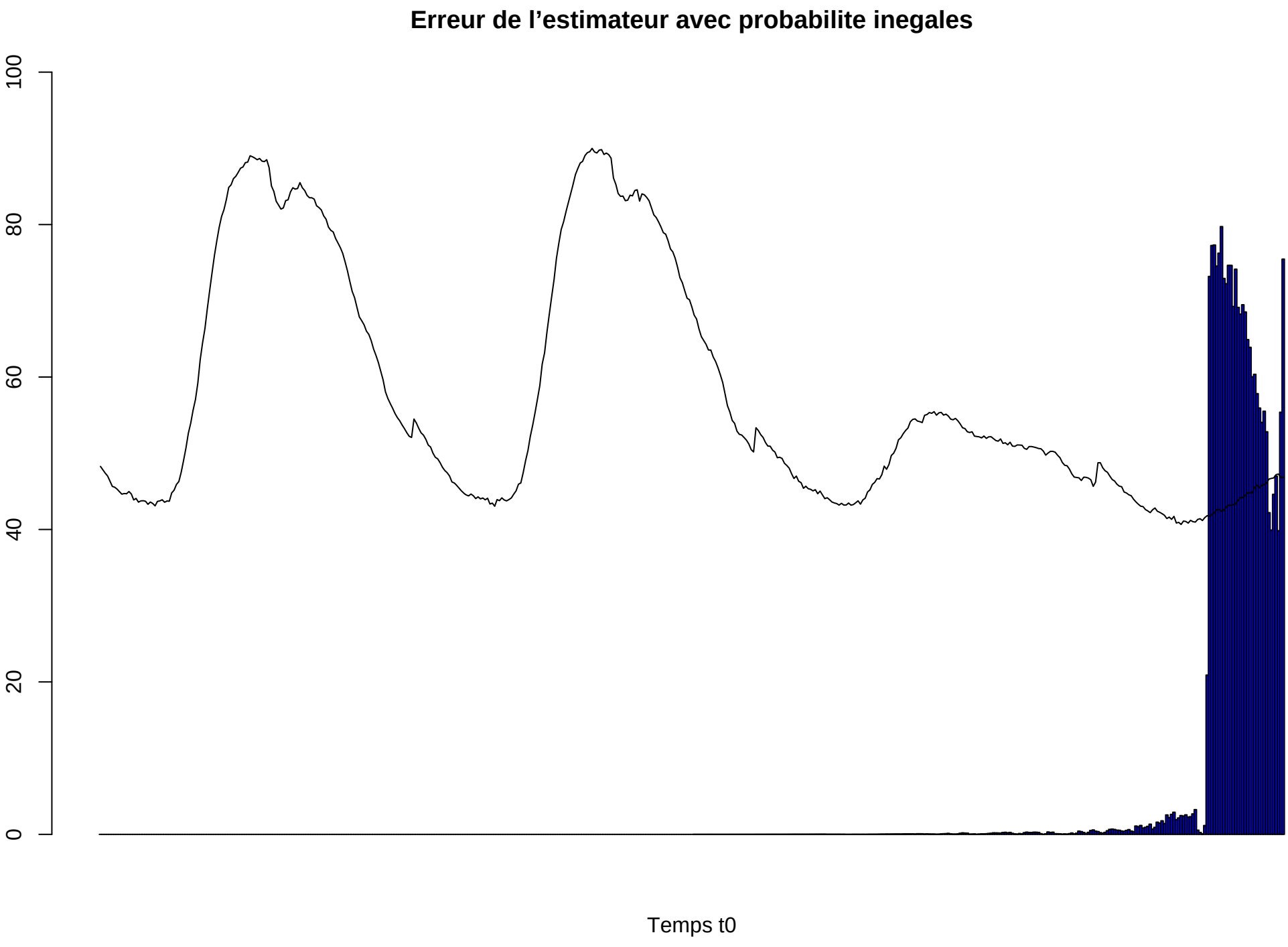
Retards différenciés suivant 5 classes de compteurs

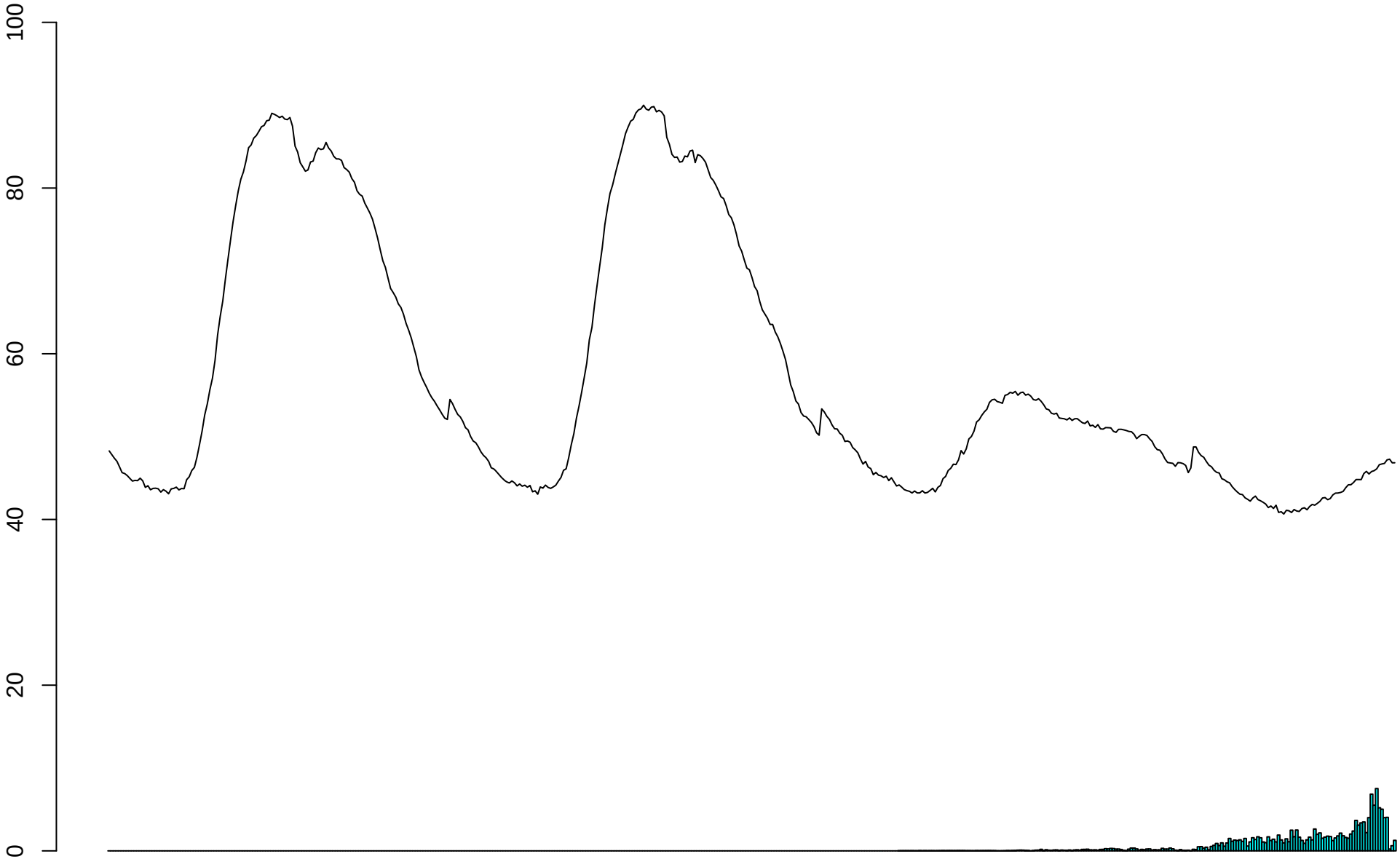

Tailles des echantillons
0
200
400
600
800
1000
Temps

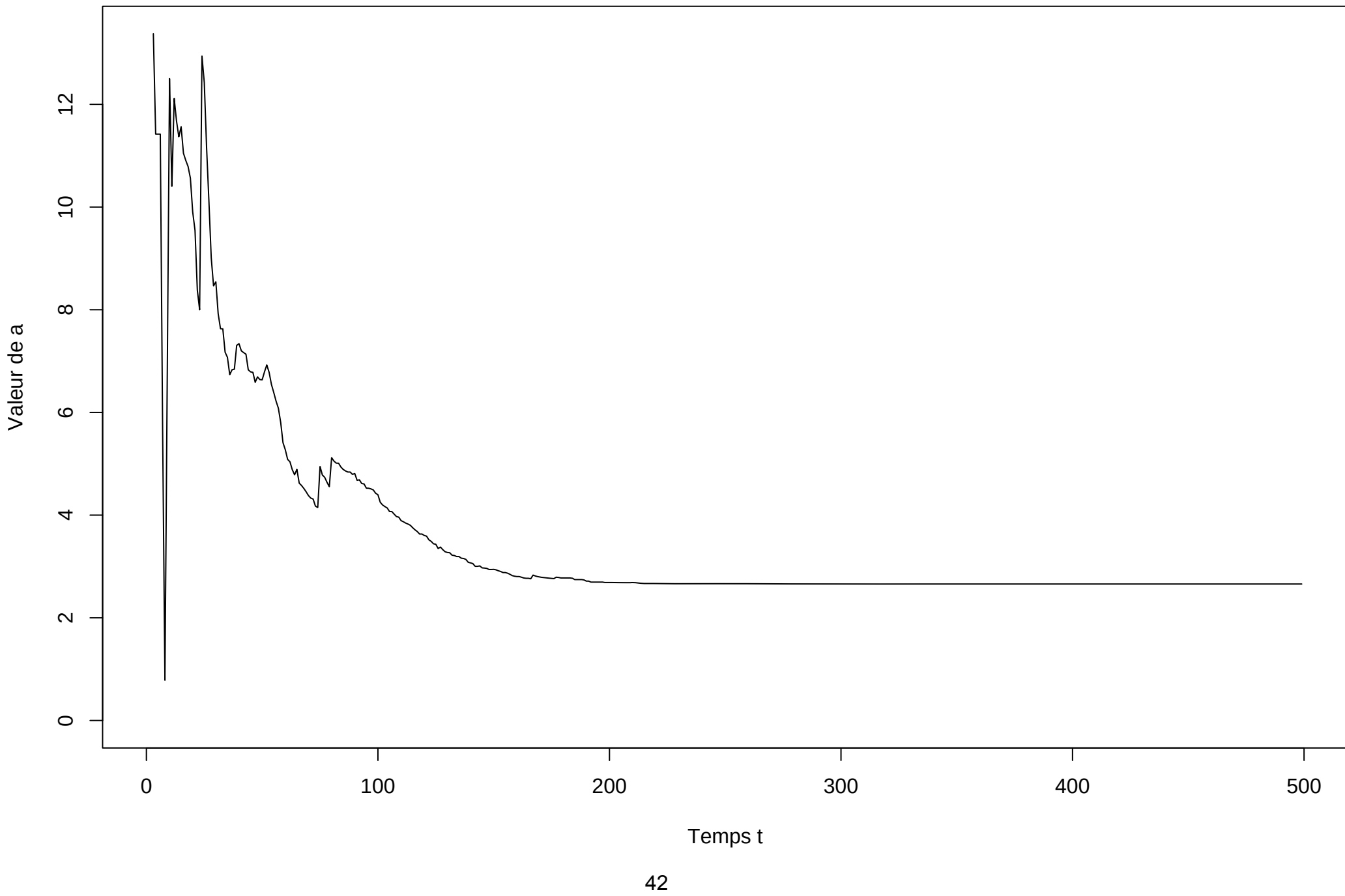

Consommations vraies et estimees
Consommation
30000
40000
50000
60000
70000
0
100
200
300
400
500
Temps t0
Vraie Consommation
Estimateur simple
Estimateur stratifie
Estimateur prob. inegales
Estimateur regression

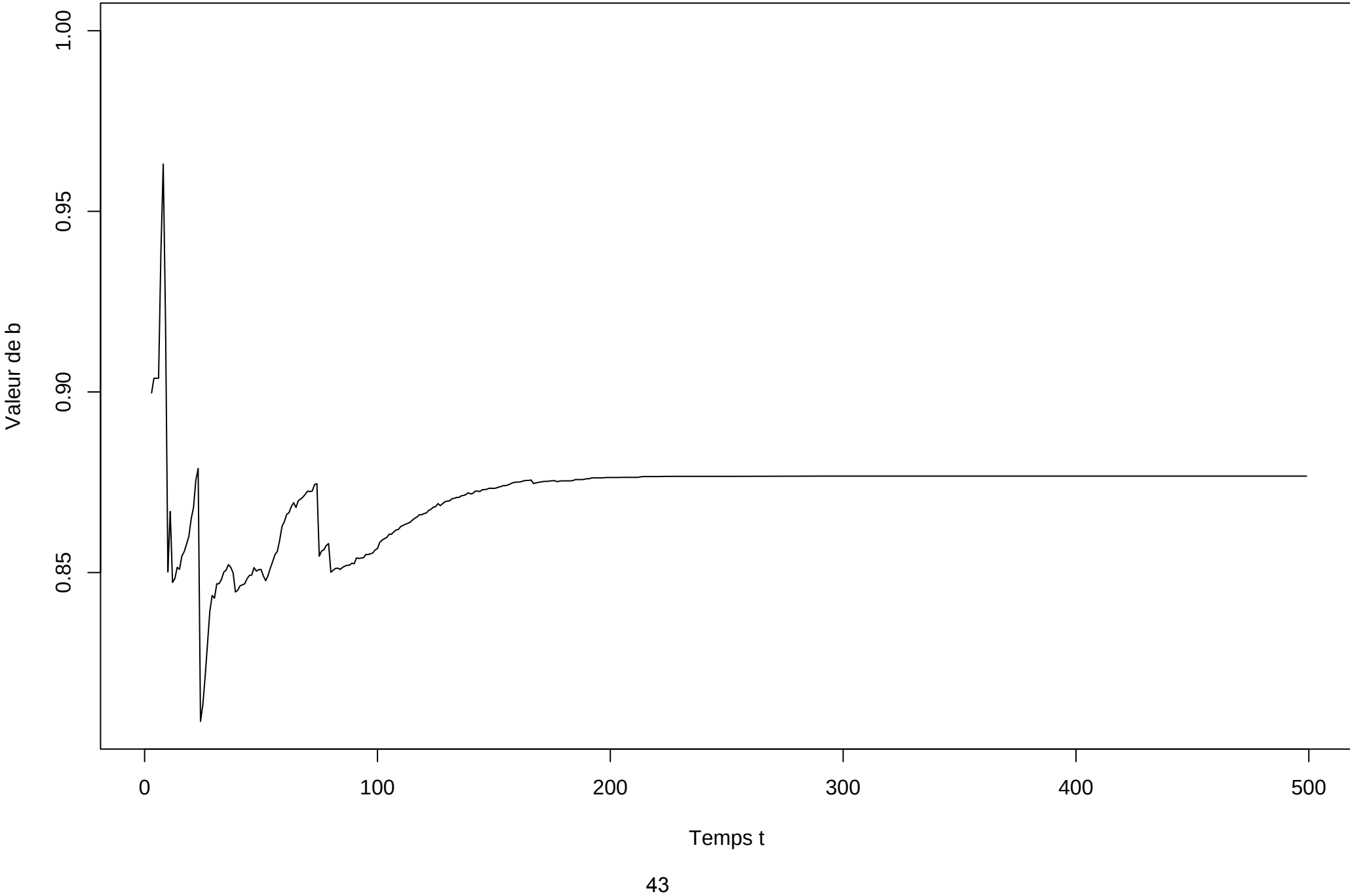

Erreur de l’estimateur par sondage simple
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur par sondage stratifie
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur avec probabilite inegales
0
20
40
60
80
100
Temps t0


Erreur de l’estimateur de la regression
100
80
60
40
20
0
Temps t0

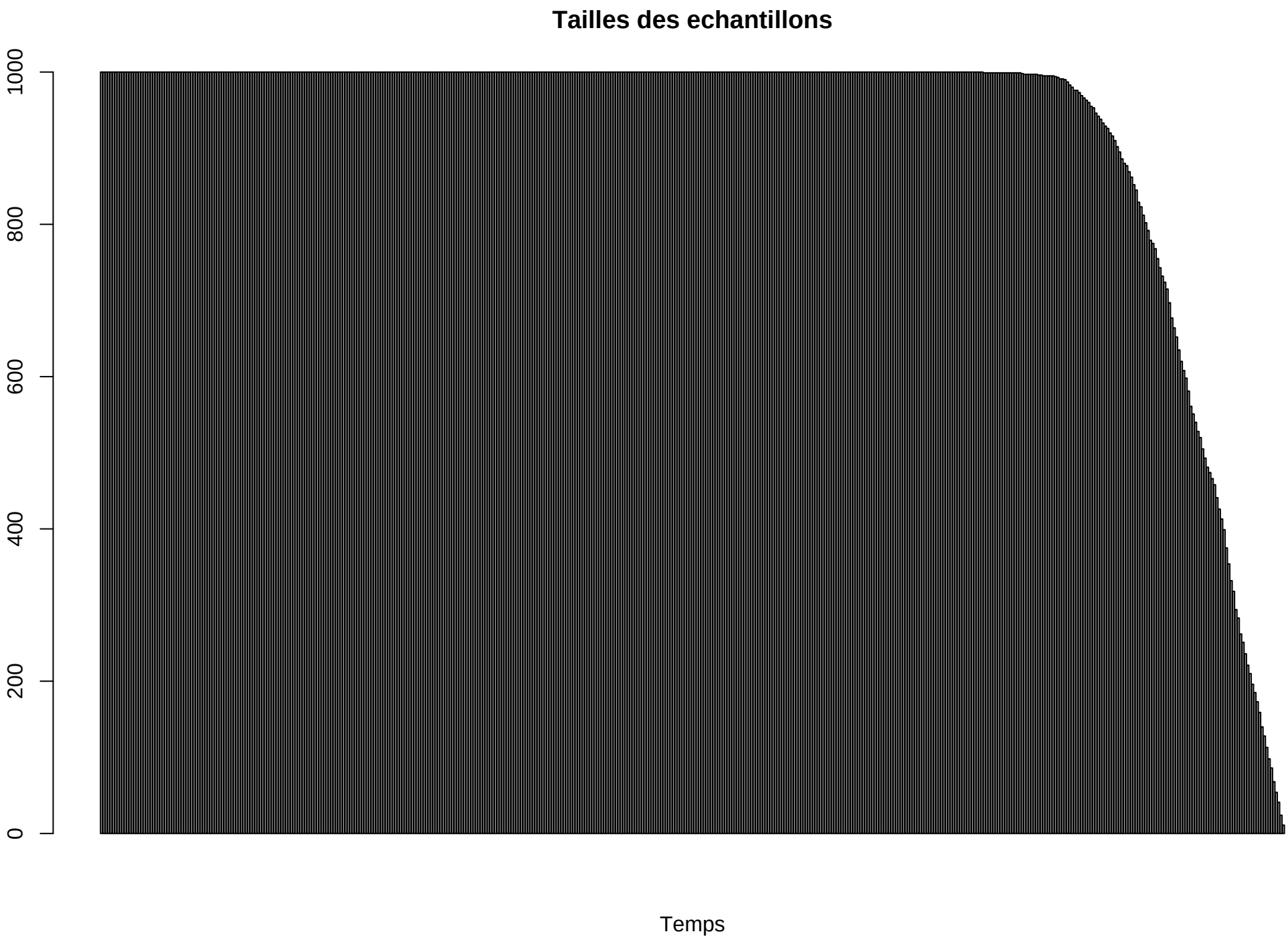

Valeur de a
0
2
4
6
8
10
12
0
100
200
300
400
500
Temps t

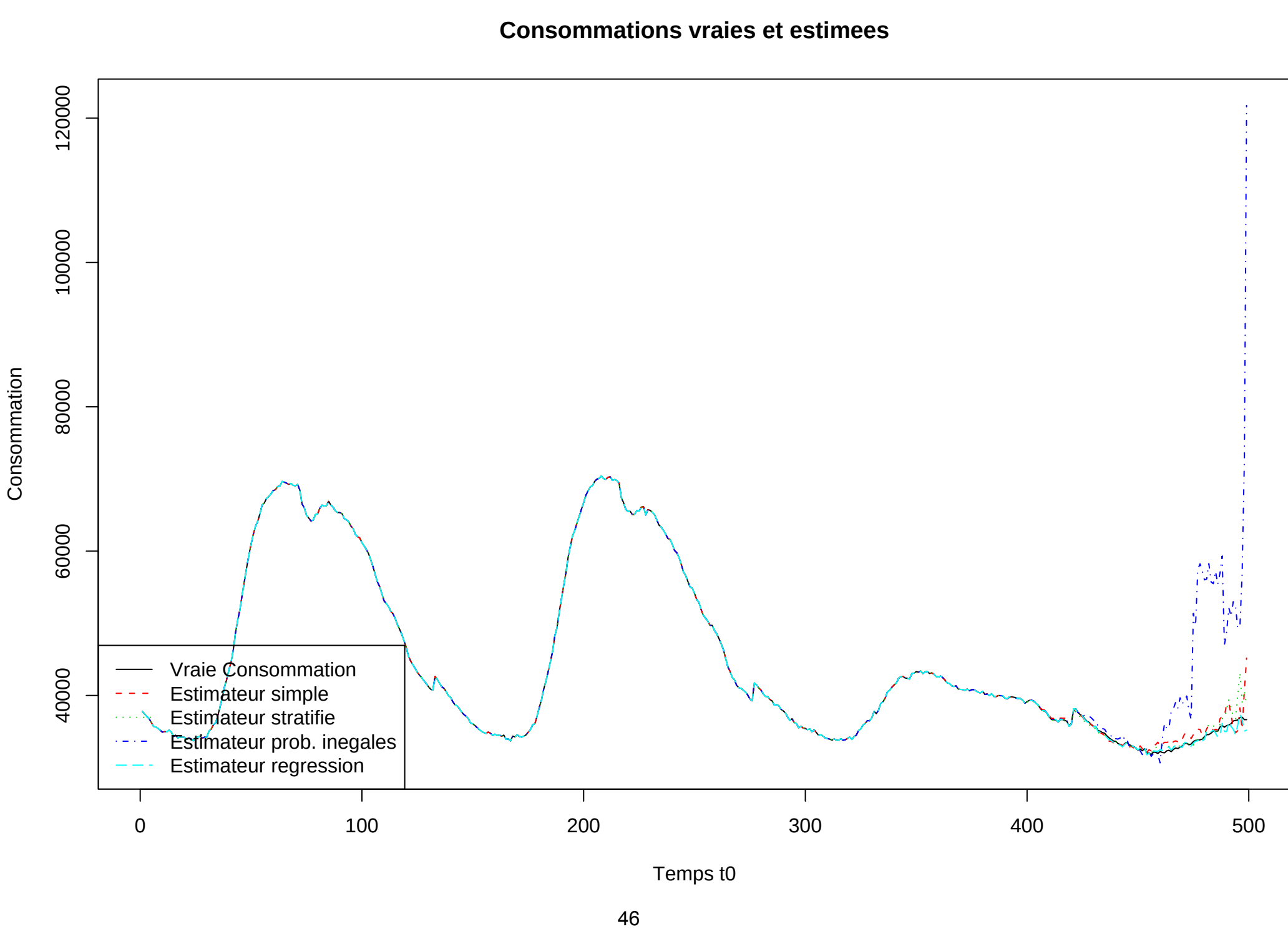

1.00
0.95
0.90
0.85
Valeur de b
0
100
200
300
400
500
Temps t

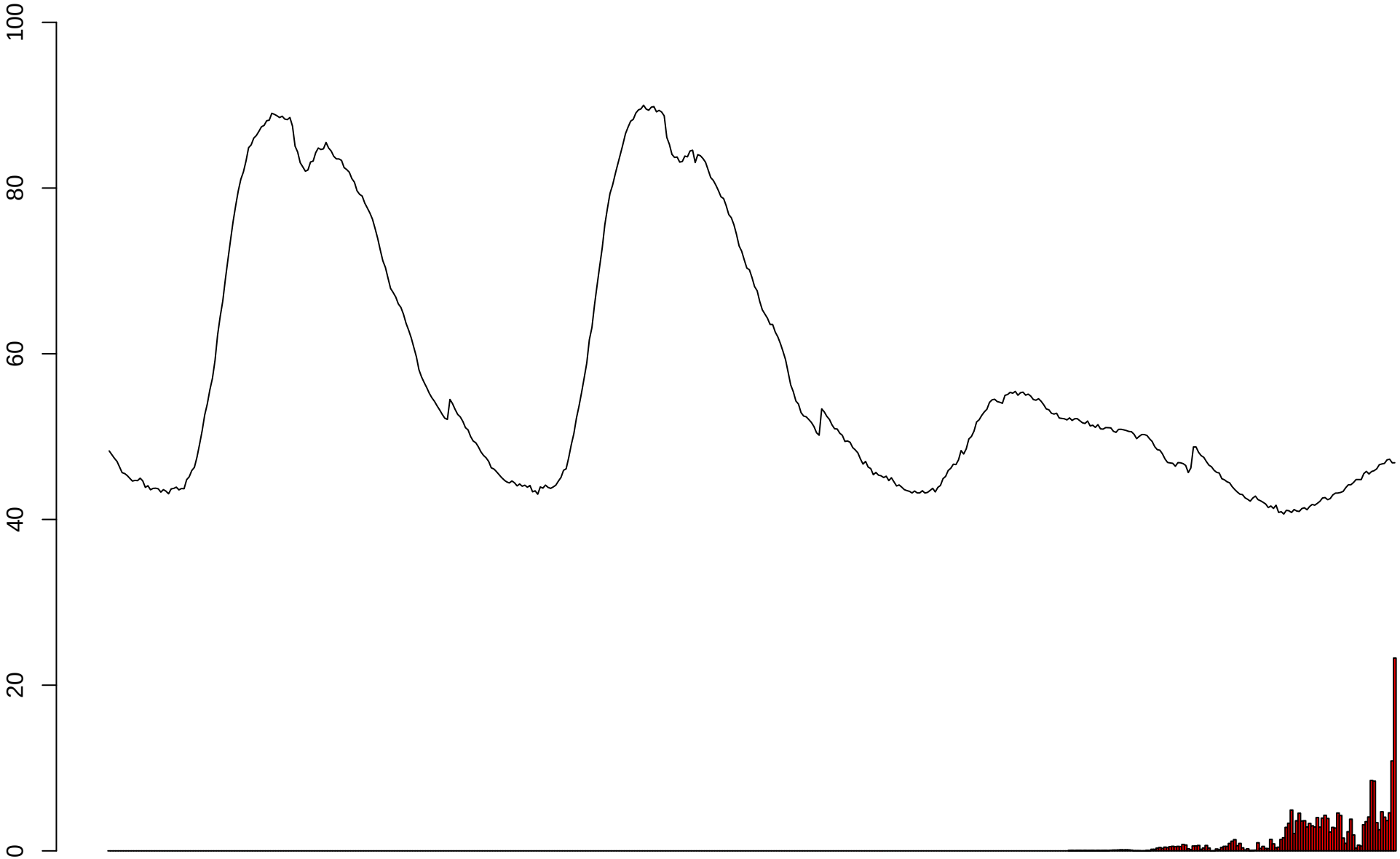
RESULTATS DES SIMULATIONS

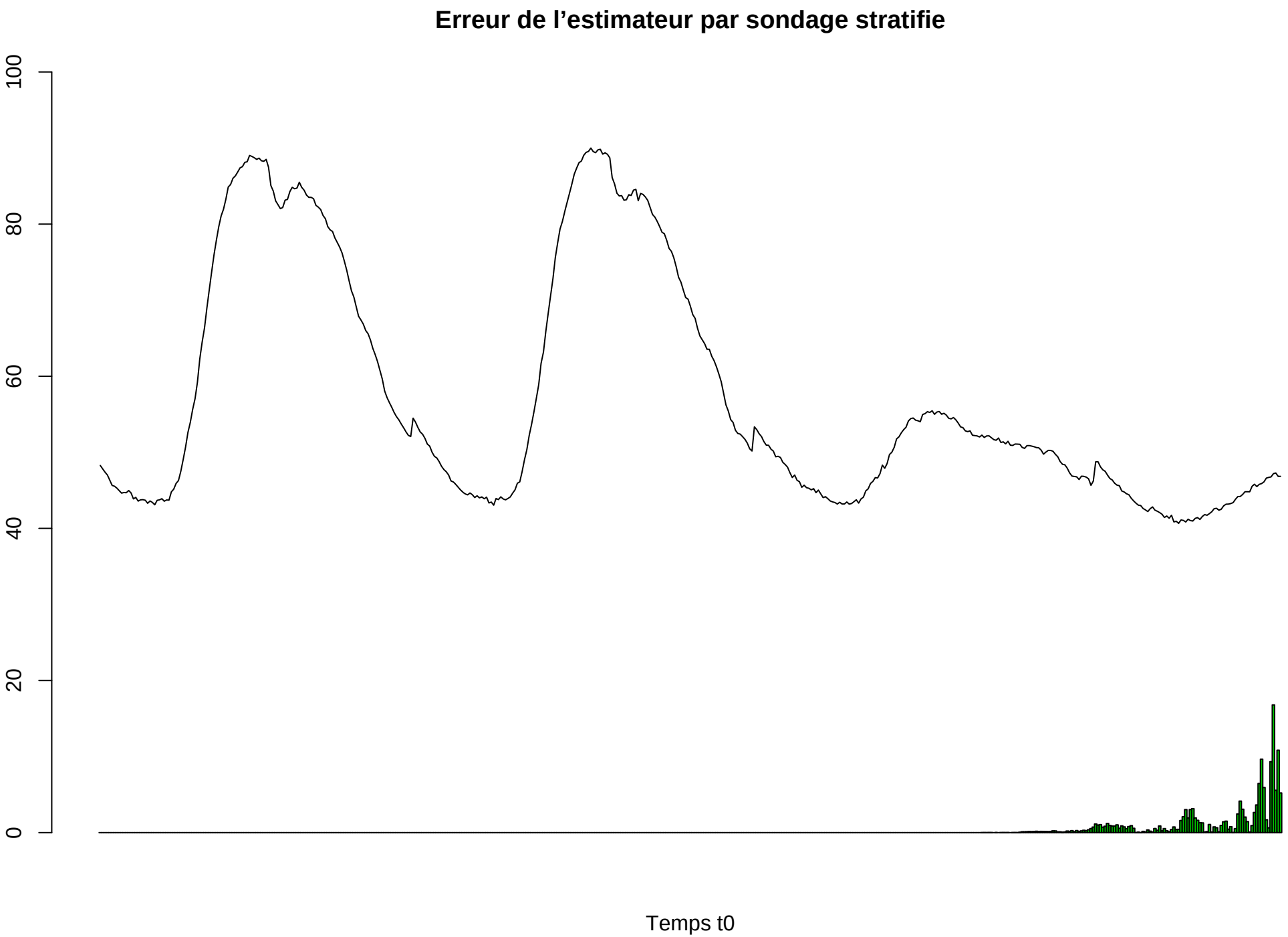
Loi Normale pour les retards

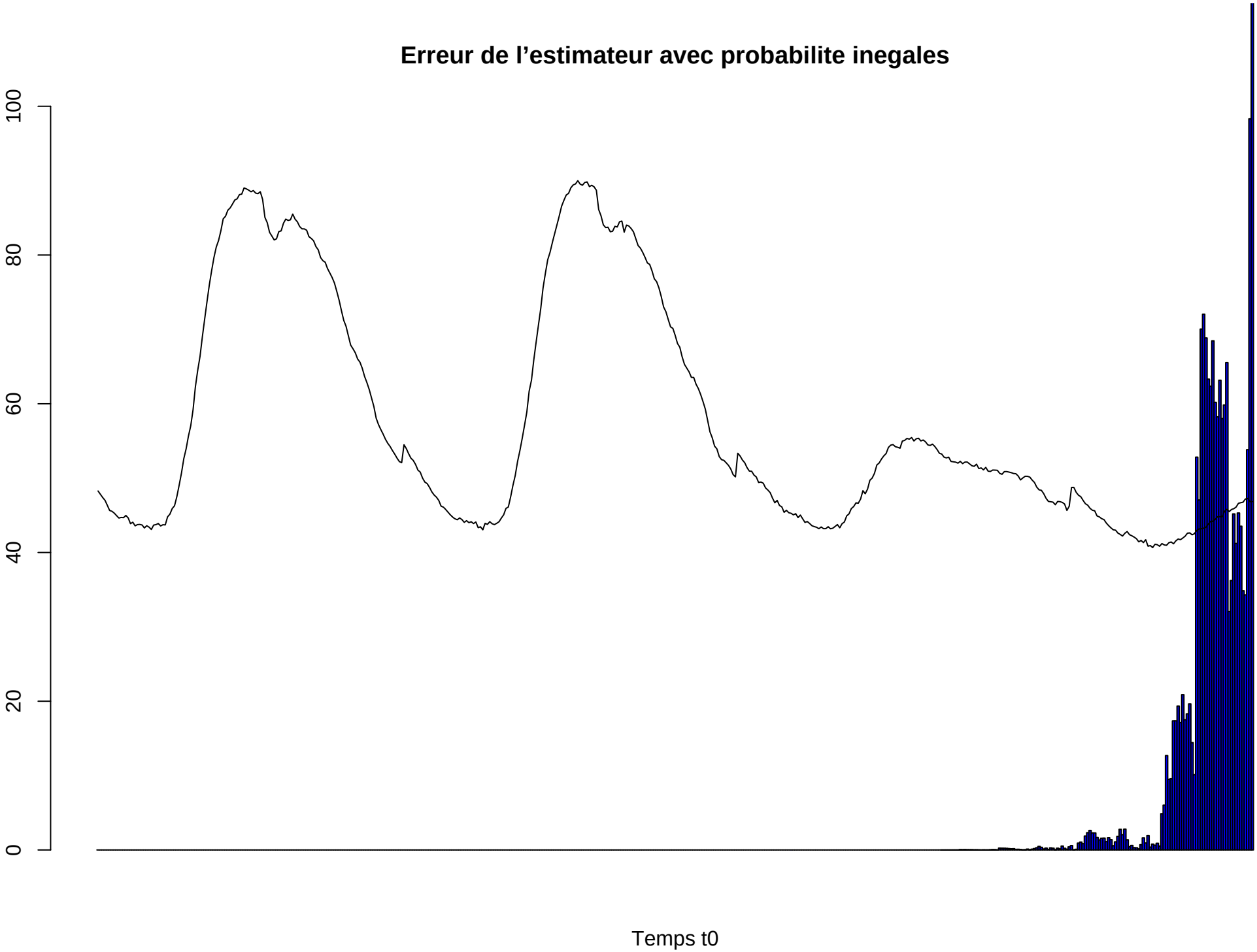
Tous les compteurs suivent la même loi : moyenne 12h, écart-type 6h

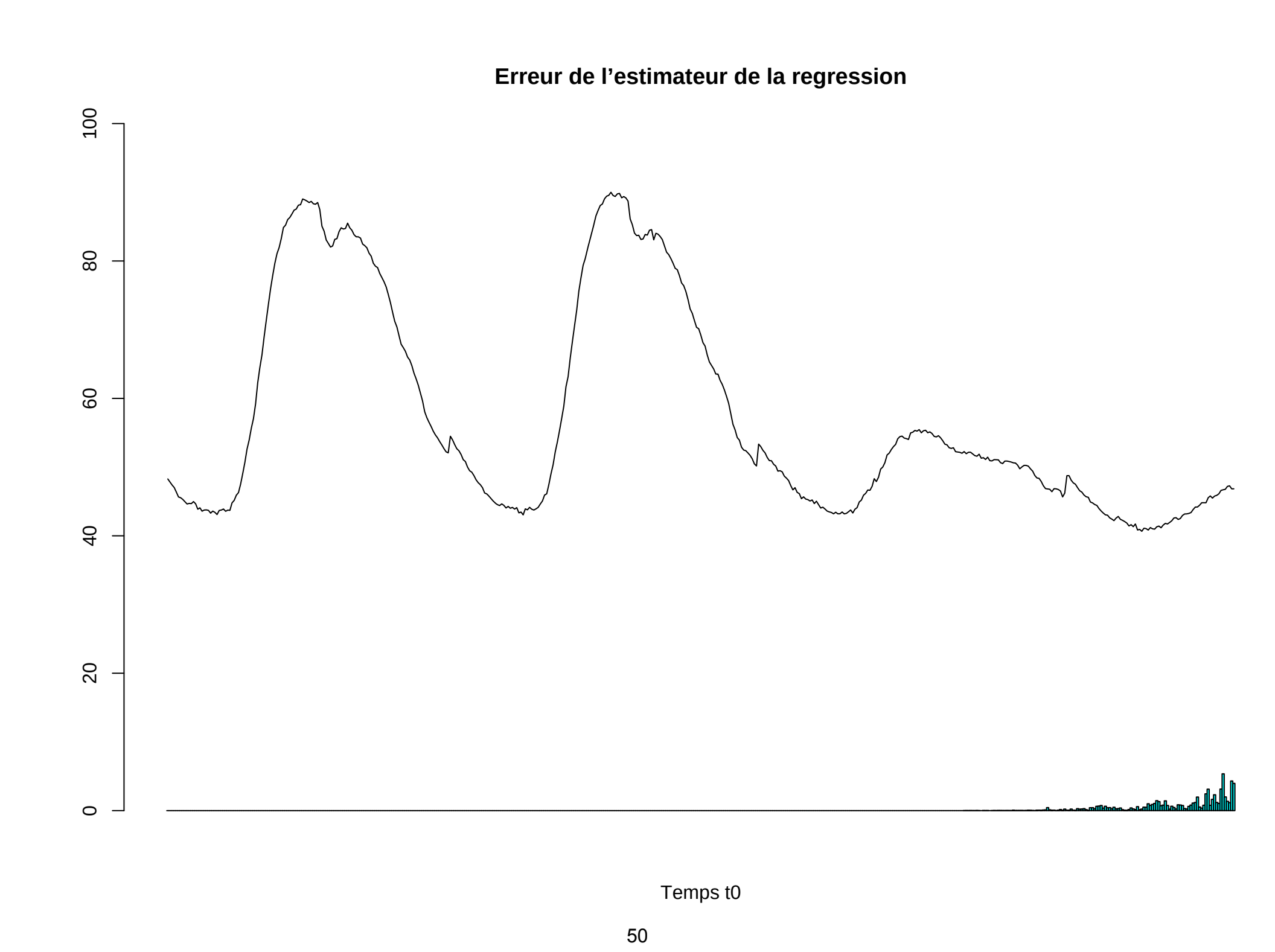

Tailles des echantillons
1000
800
600
400
200
0
Temps

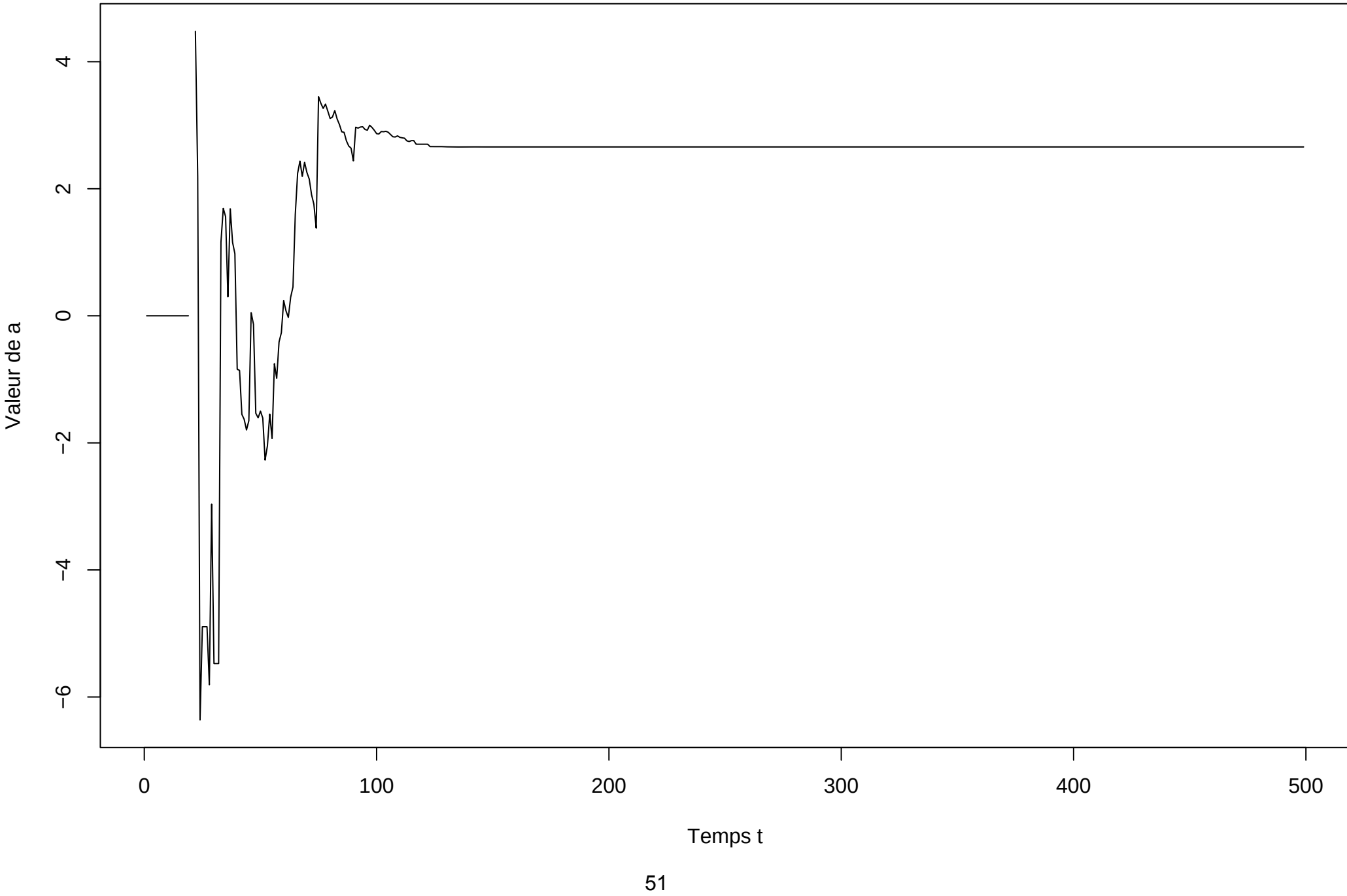

Consommations vraies et estimees
Consommation
120000
100000
80000
60000
40000
0
100
200
300
400
500
Temps t0
Vraie Consommation
Estimateur simple
Estimateur stratifie
Estimateur prob. inegales
Estimateur regression

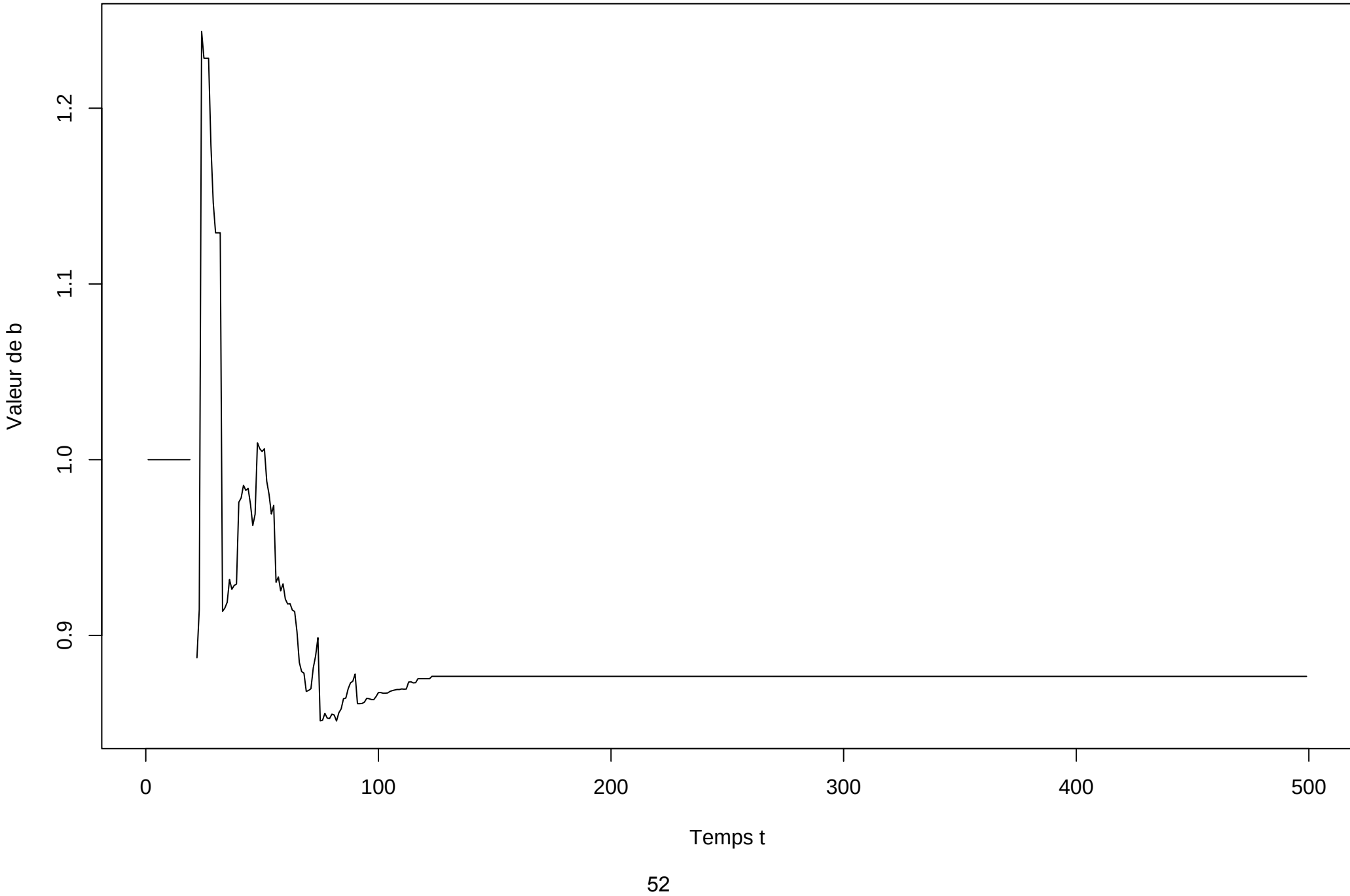

Erreur de l’estimateur par sondage simple
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur par sondage stratifie
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur avec probabilite inegales
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur de la regression
100
80
60
40
20
0
Temps t0


Valeur de a
4
2
0
−2
−4
−6
0
100
200
300
400
500
Temps t

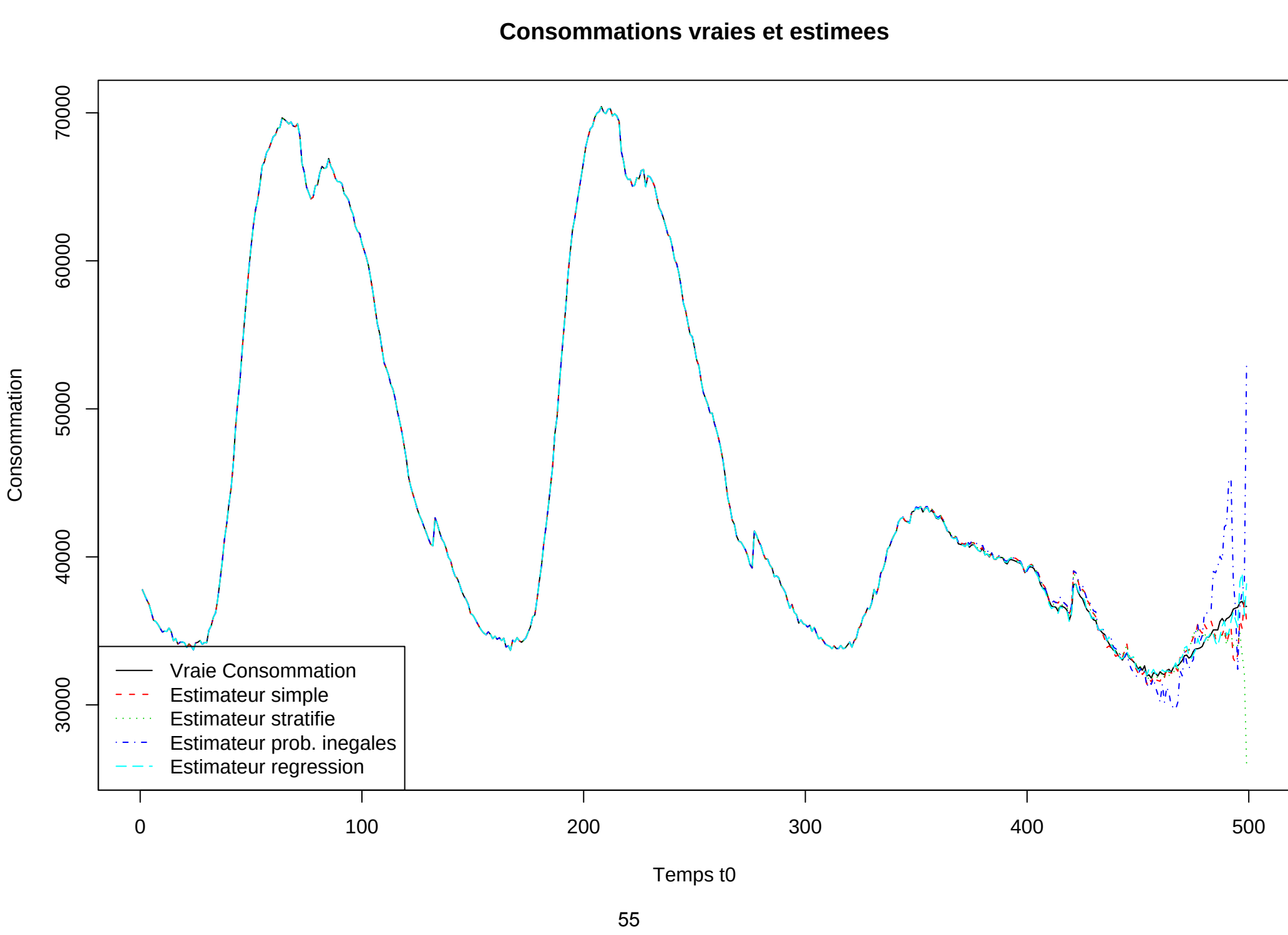

Valeur de b
0.9
1.0
1.1
1.2
0
100
200
300
400
500
Temps t

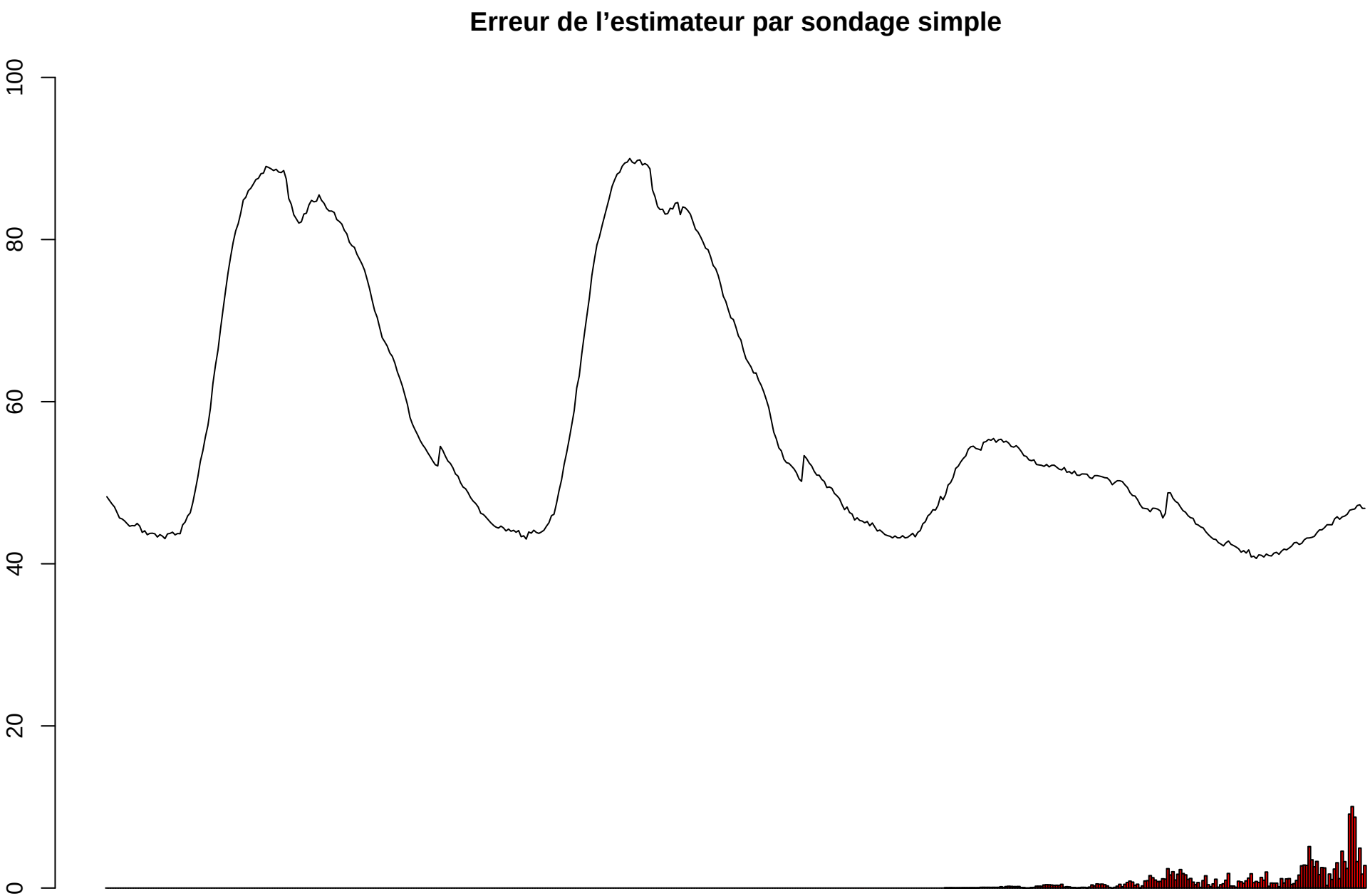
RESULTATS DES SIMULATIONS

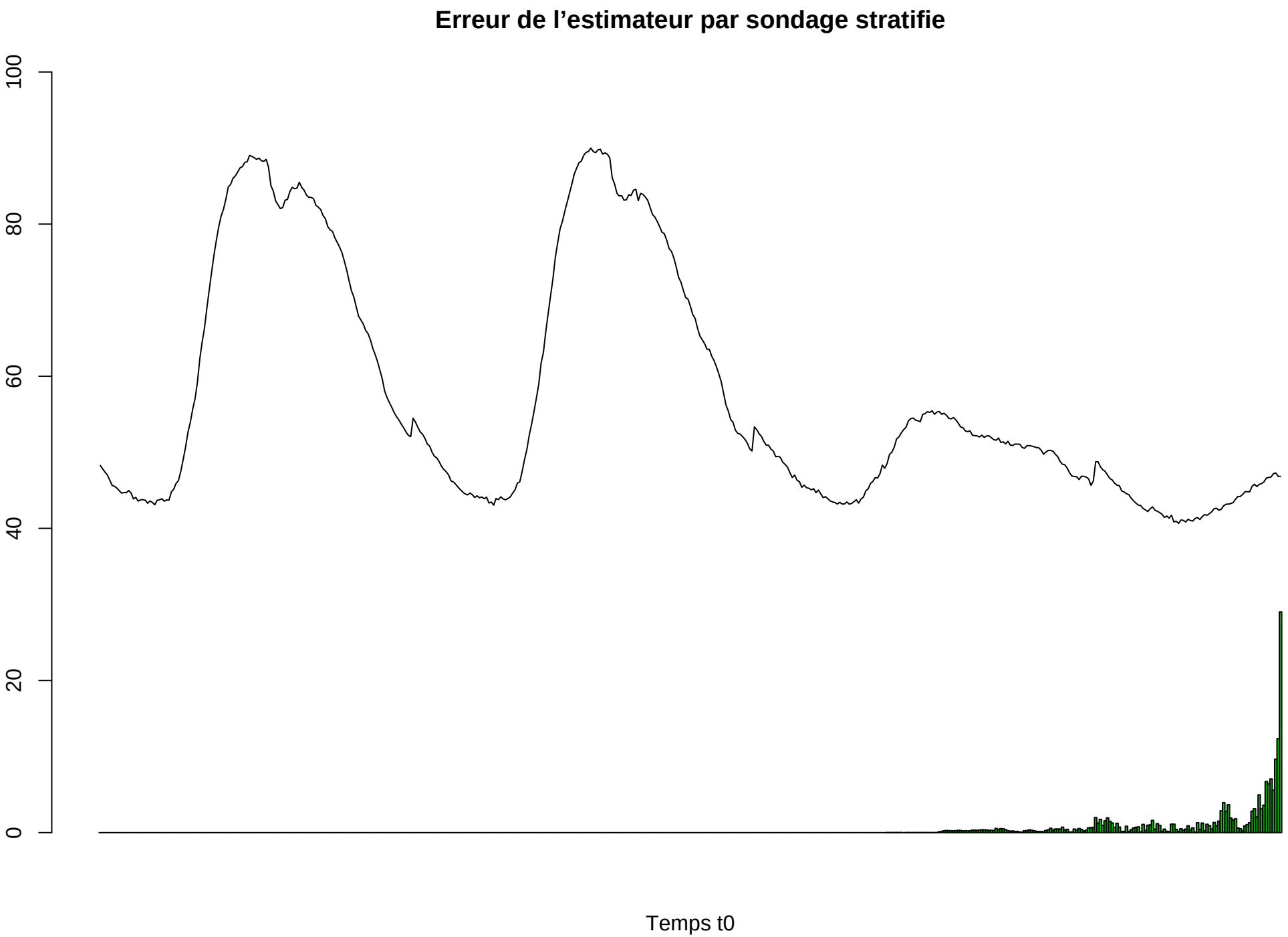
Loi Normale pour les retards

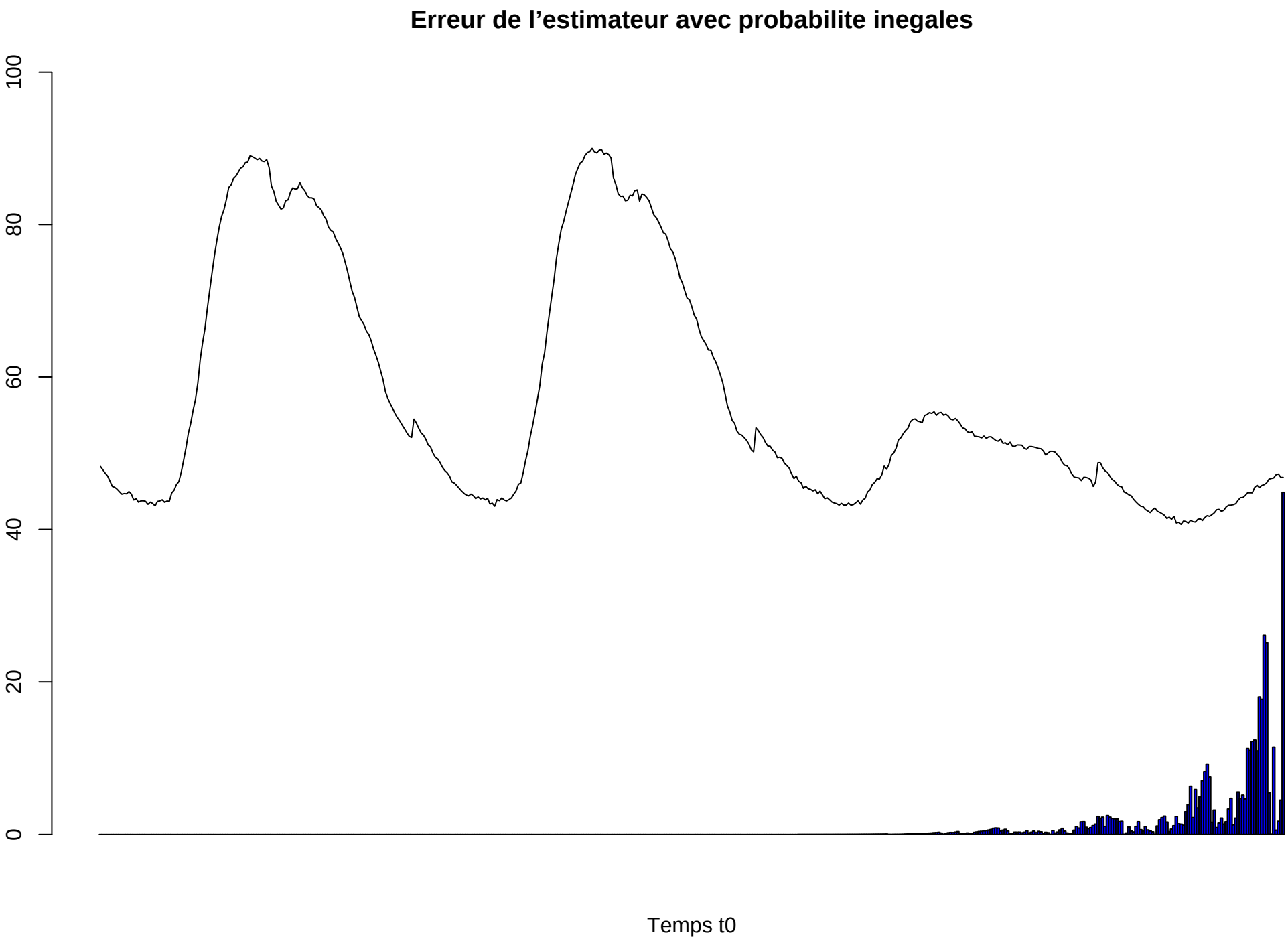
Tous les compteurs suivent la même loi : moyenne 12h, écart-type 24h

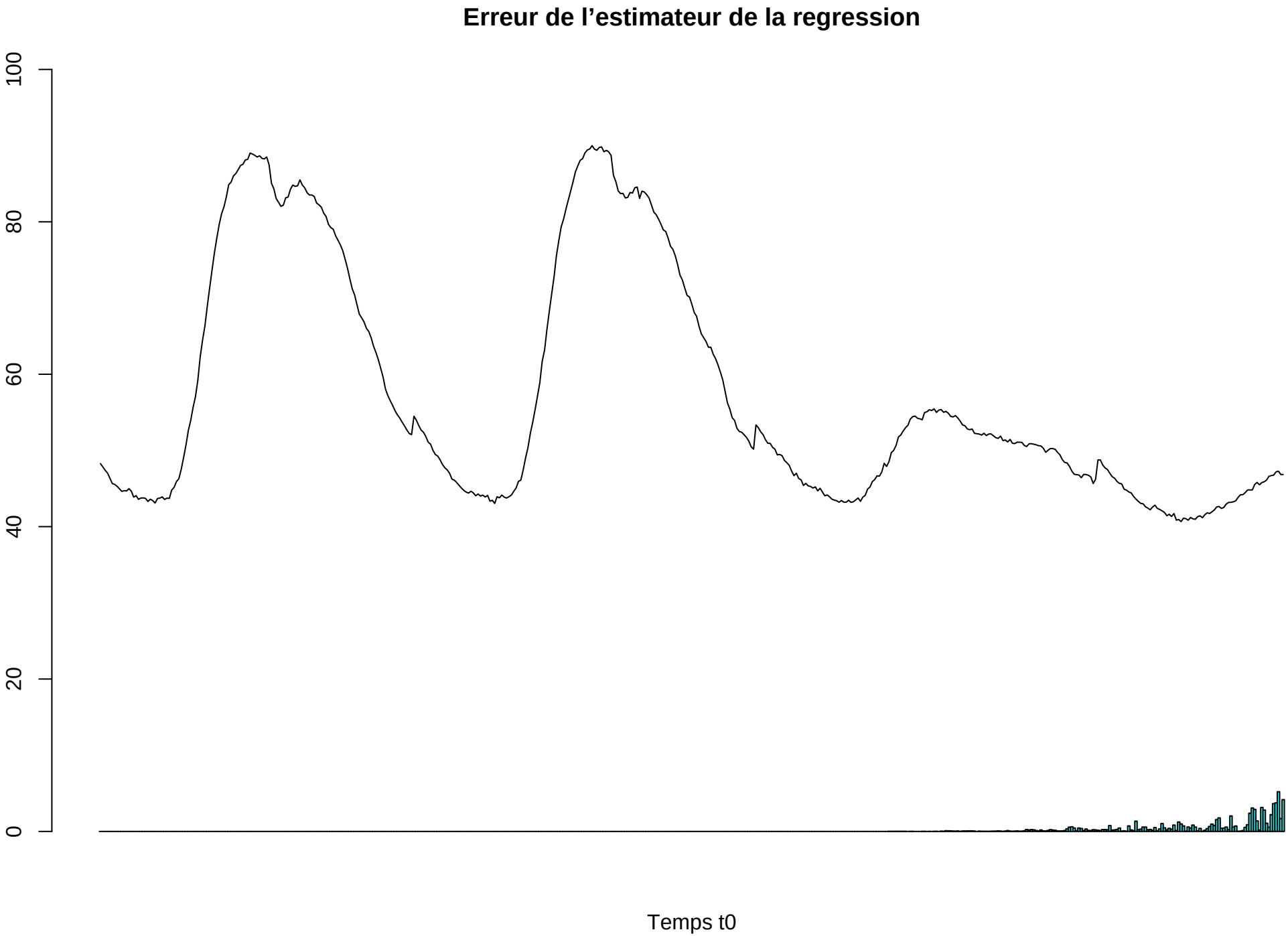

Tailles des echantillons
1000
800
600
400
200
0
Temps

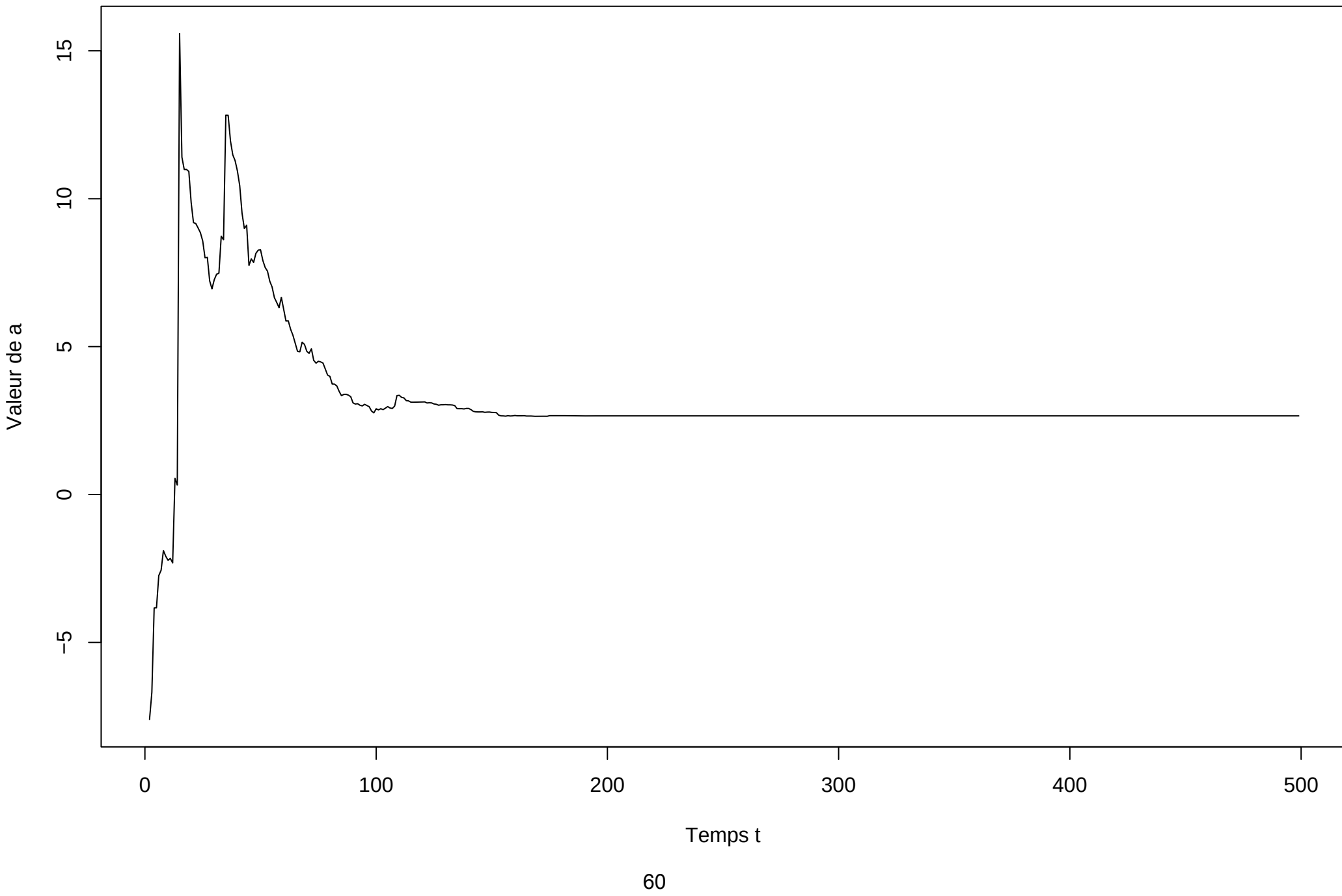

Consommations vraies et estimees
Consommation
70000
60000
50000
40000
30000
0
100
200
300
400
500
Temps t0
Vraie Consommation
Estimateur simple
Estimateur stratifie
Estimateur prob. inegales
Estimateur regression

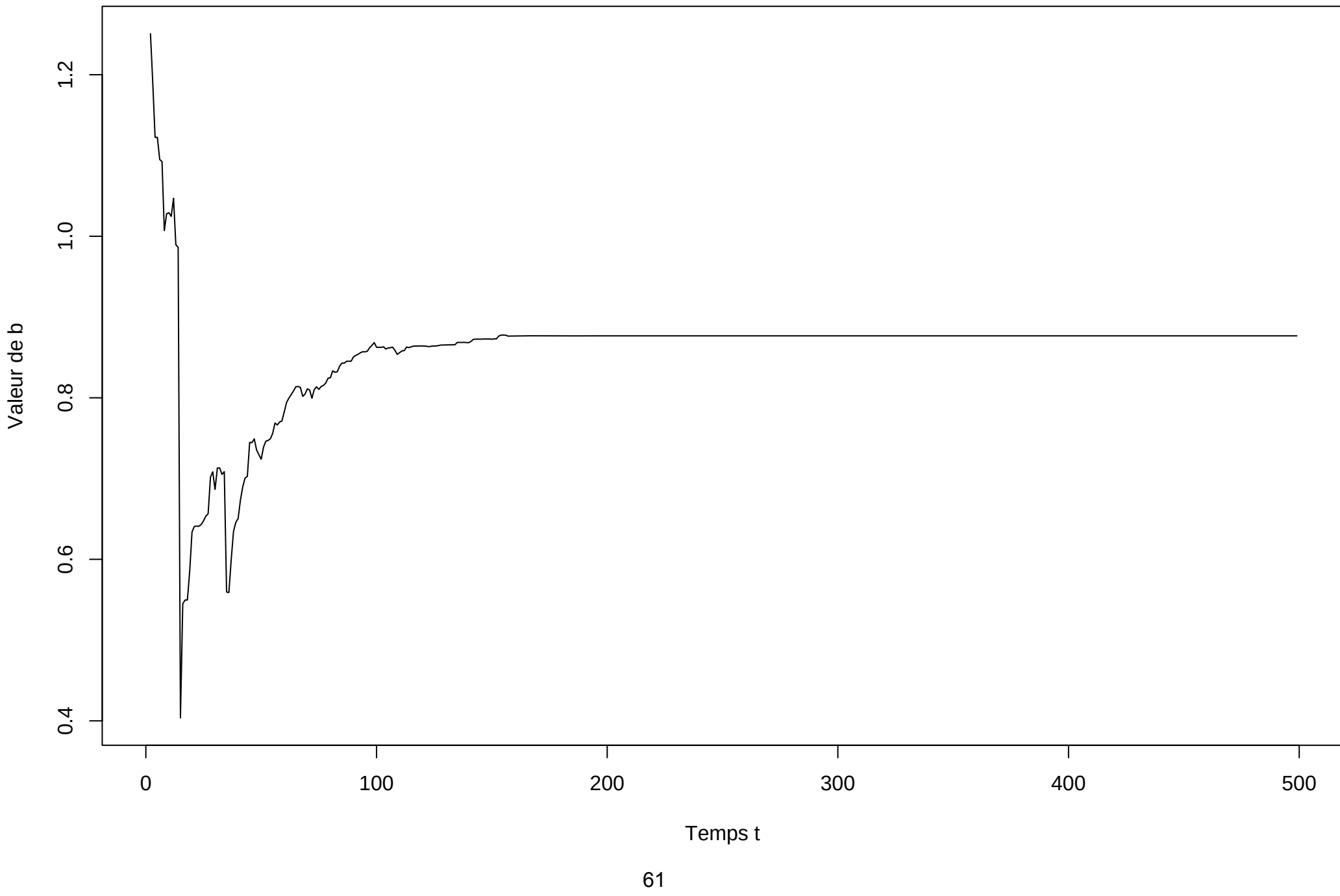

Erreur de l’estimateur par sondage simple
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur par sondage stratifie
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur avec probabilite inegales
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur de la regression
100
80
60
40
20
0
Temps t0


Valeur de a
15
10
5
0
−5
0
100
200
300
400
500
Temps t

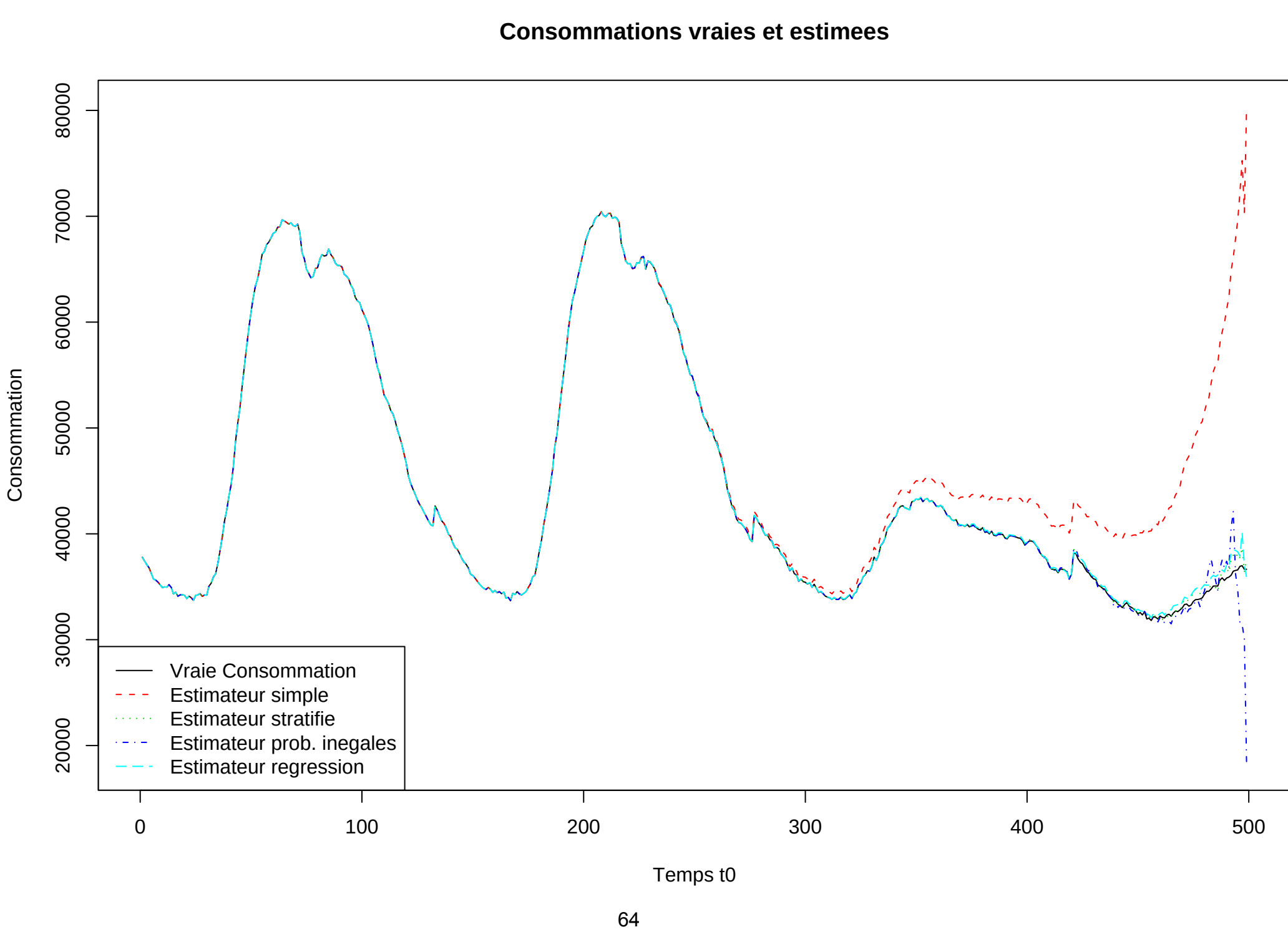

Valeur de b
0.4
0.6
0.8
1.0
1.2
0
100
200
300
400
500
Temps t

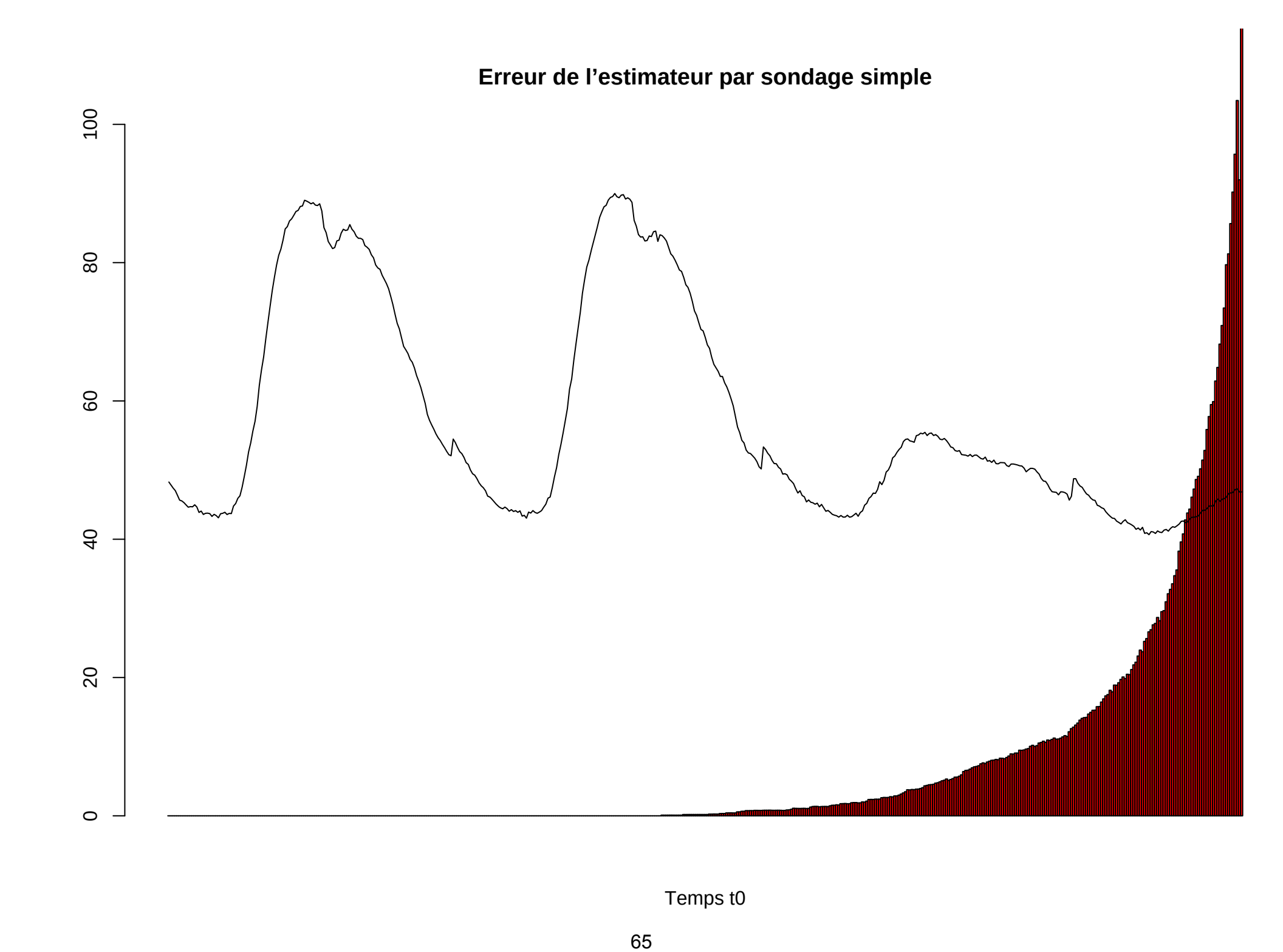
RESULTATS DES SIMULATIONS

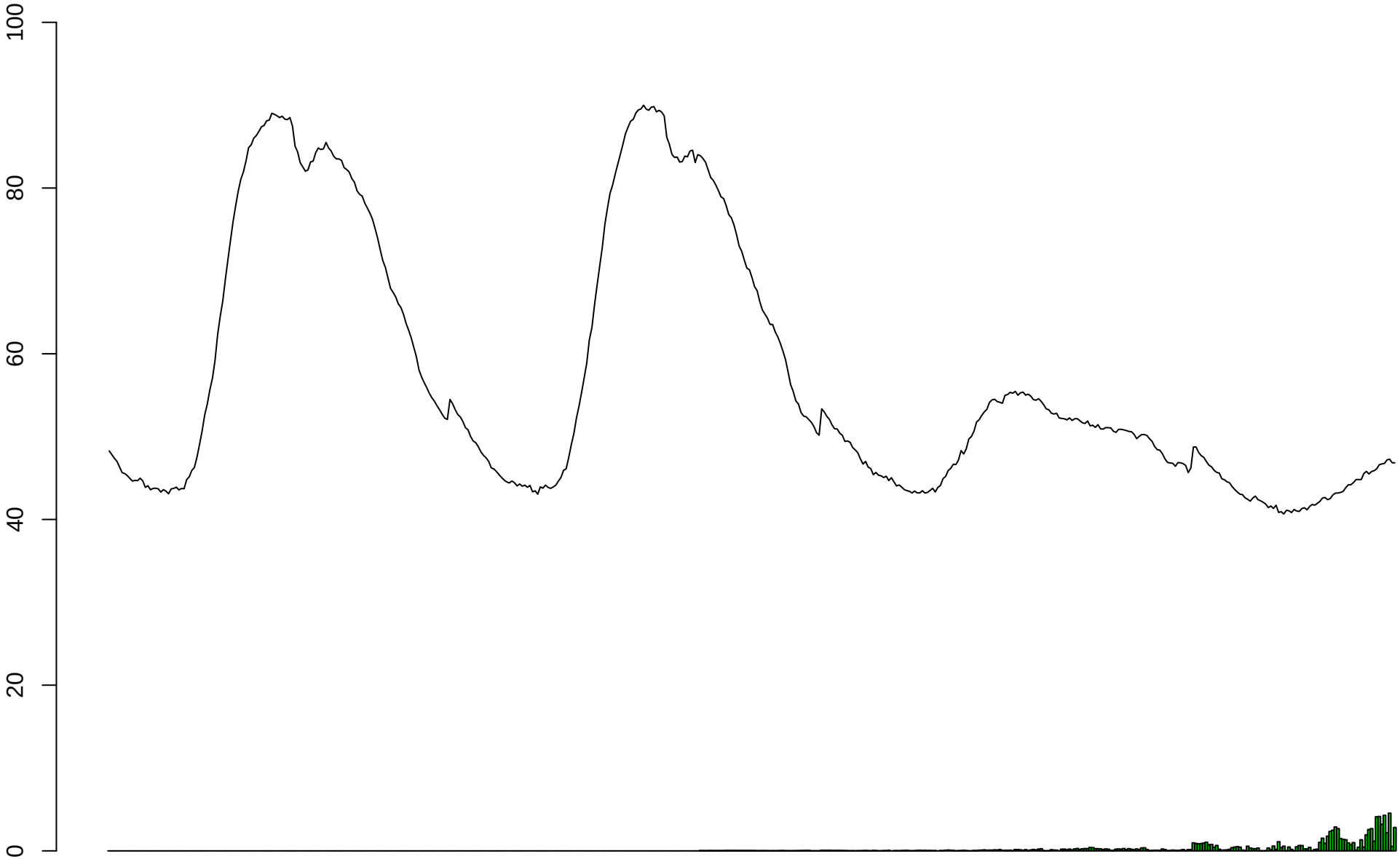
Loi Normale pour les retards

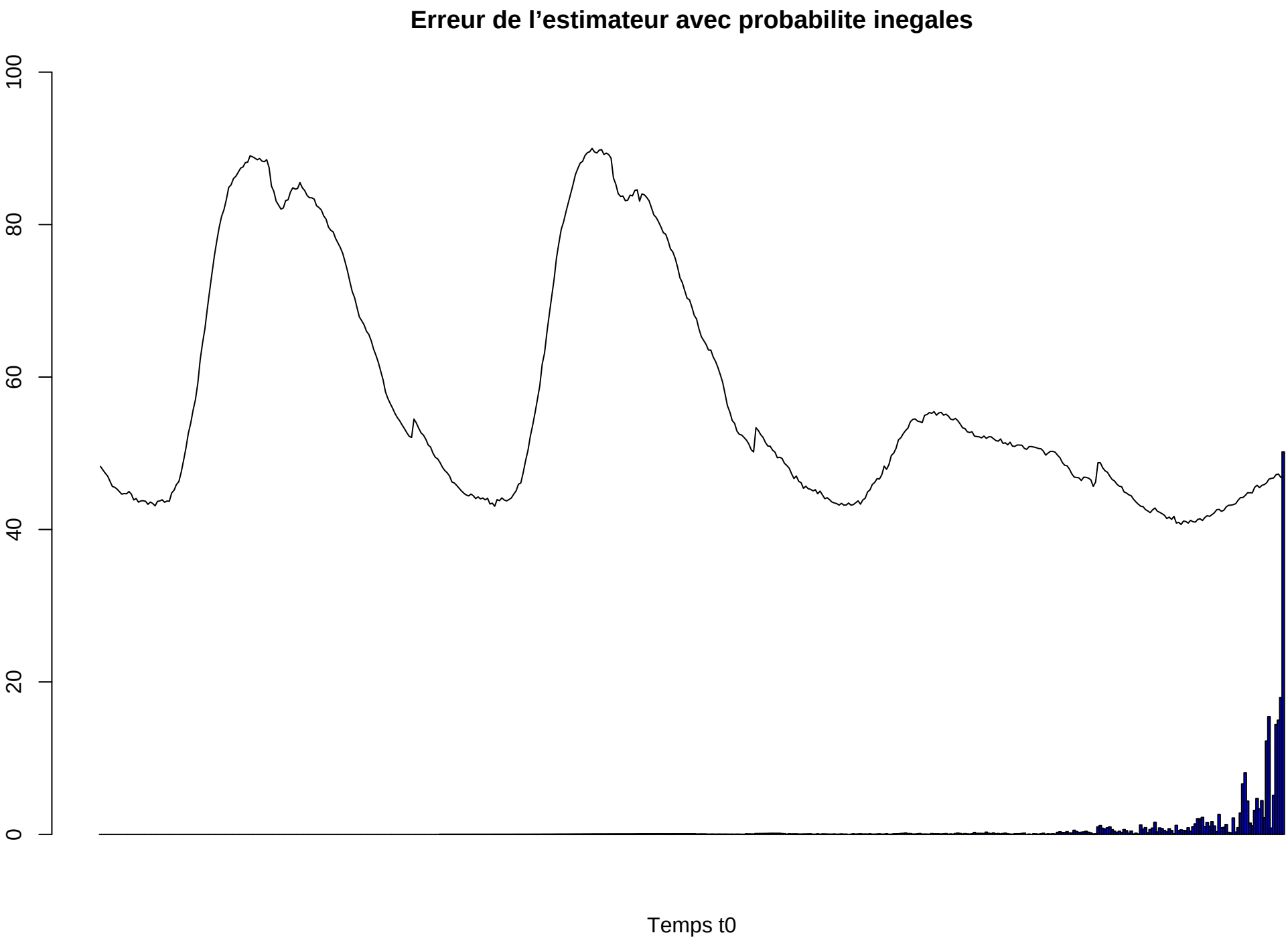
Retards différenciés suivant 5 classes de compteurs

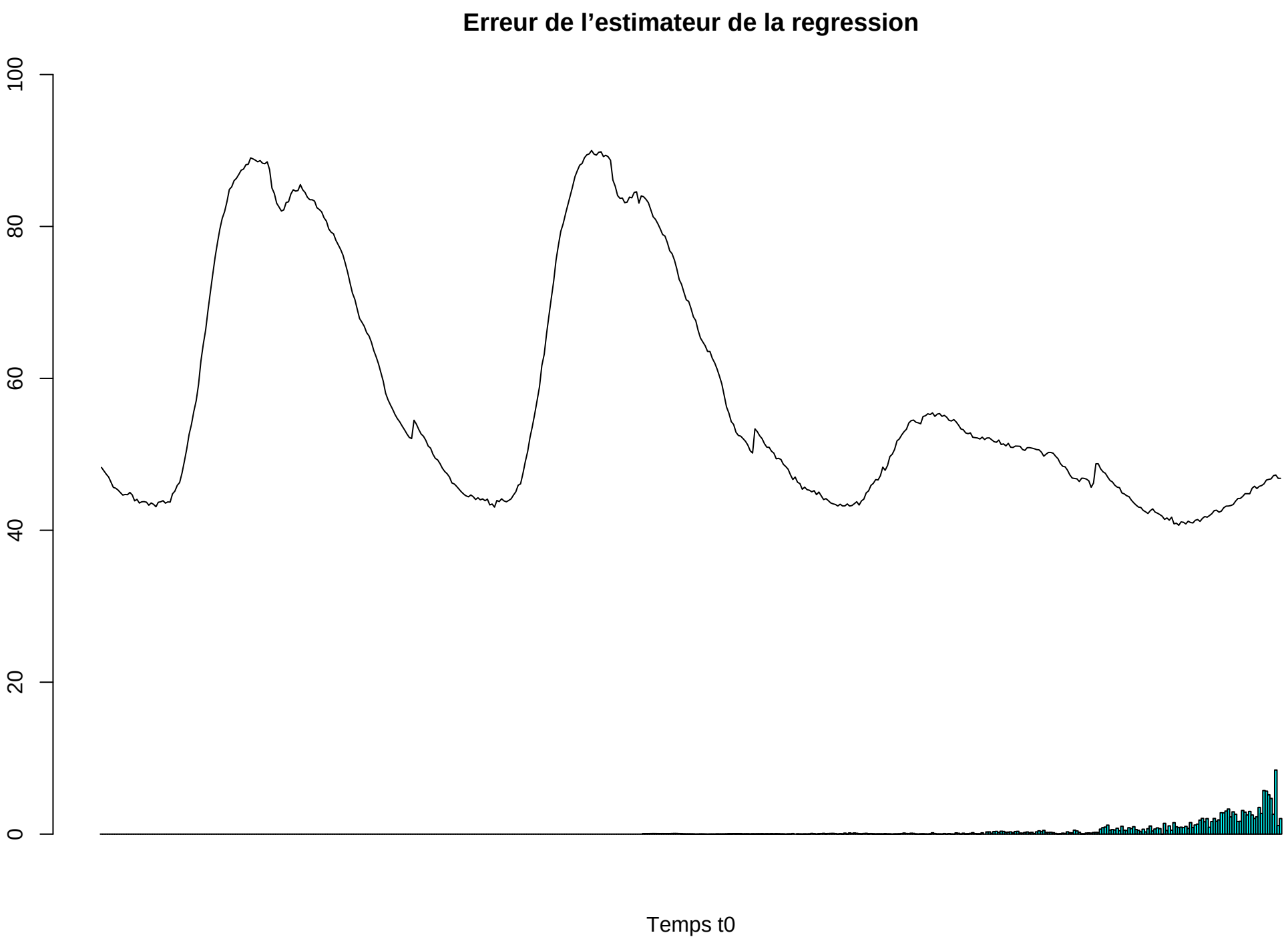

Tailles des echantillons
1000
800
600
400
200
0
Temps

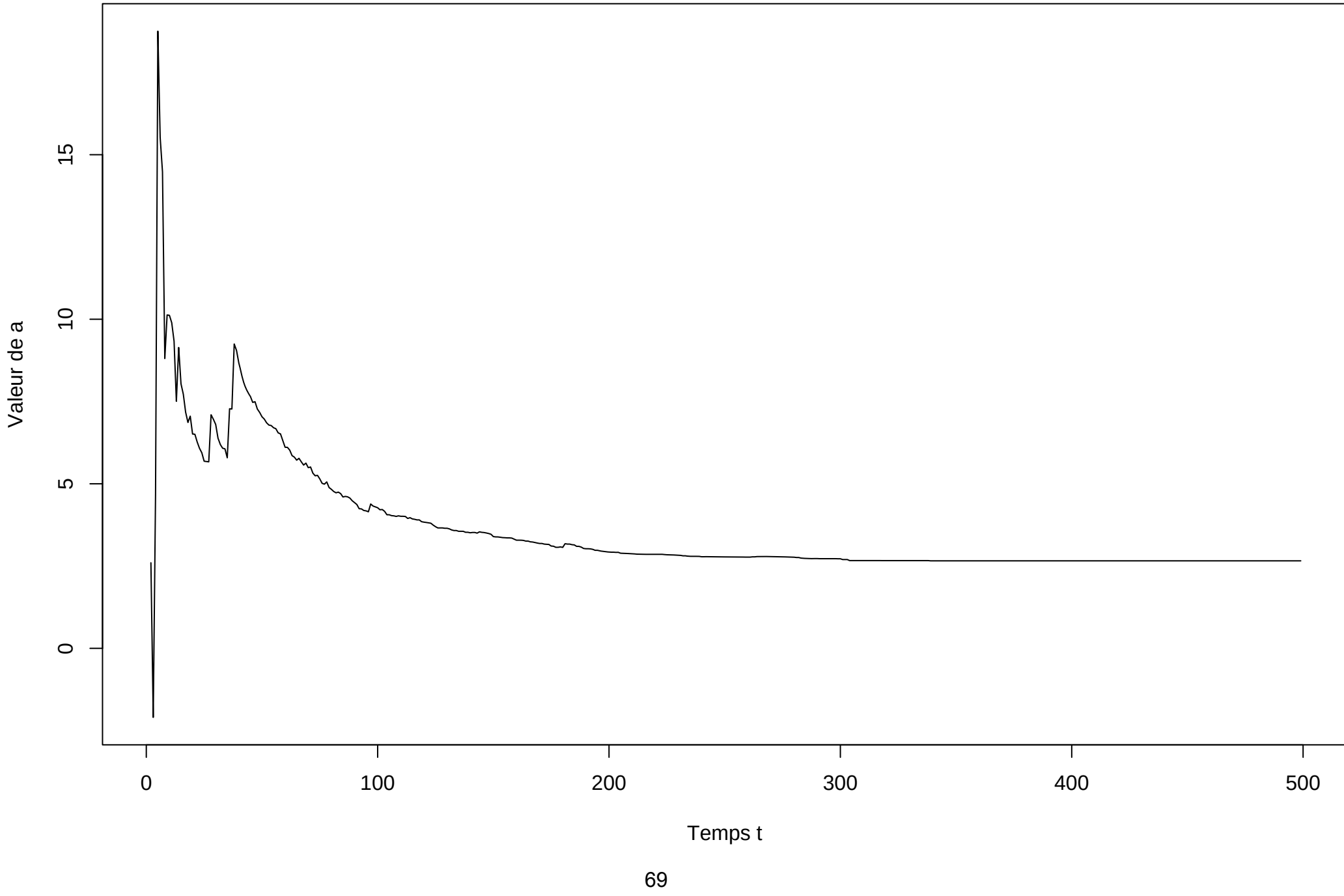

Consommations vraies et estimees
Consommation
20000
30000
40000
50000
60000
70000
80000
0
100
200
300
400
500
Temps t0
Vraie Consommation
Estimateur simple
Estimateur stratifie
Estimateur prob. inegales
Estimateur regression

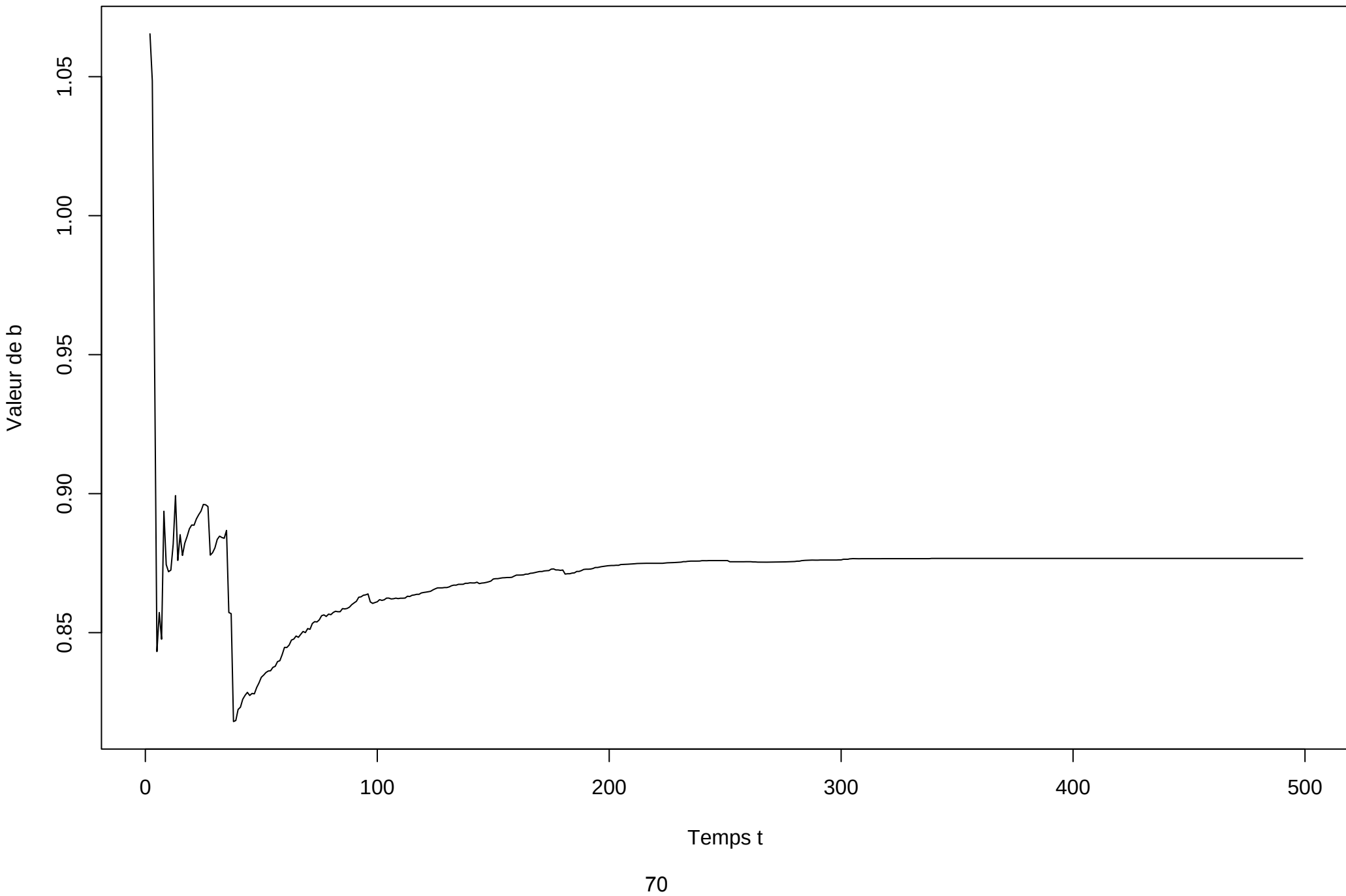

Erreur de l’estimateur par sondage simple
0
20
40
60
80
100
Temps t0


Erreur de l’estimateur par sondage stratifie
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur avec probabilite inegales
100
80
60
40
20
0
Temps t0


Erreur de l’estimateur de la regression
100
80
60
40
20
0
Temps t0


Valeur de a
0
5
10
15
0
100
200
300
400
500
Temps t


Valeur de b
1.05
1.00
0.95
0.90
0.85
0
100
200
300
400
500
Temps t

Dépôt légal : 2008 – 4ème trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

TELECOM ParisTech
Institut TELECOM - membre de ParisTech
46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr
Département INFRES