Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte

Ibrahim Bouzalmat, Benoîte de Saporta, Solym Manou-Abi

Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte

Ibrahim Bouzalmat [1] & Benoîte de Saporta [1] & Solym Manou-Abi [2]

[1] *IMAG, Univ Montpellier, CNRS, Montpellier, France,*
ibrahim.bouzalmat@umontpellier.fr , benoite.de-saporta@umontpellier.fr
[2] *IMAG, Univ Montpellier, CNRS, CUFR Mayotte, France,*
solym.manou-abi@univ-mayotte.fr

Résumé. L'objectif de ce travail est de modéliser la propagation de la fièvre typhoïde à Mayotte à partir d'un jeu de données d'hospitalisations fourni par l'Agence Régionale de Santé. Nous utilisons un processus de naissance et mort linéaire avec immigration comptabilisant les personnes infectées par la maladie. L'objectif est alors d'estimer les taux de contamination de personne à personne, contamination par l'environnement et guérison à partir des données. L'originalité de notre approche et la difficulté du problème provient de deux sources. D'une part, les observations ne sont pas disponibles en temps continu mais seulement à des dates fixes (hospitalisations journalières), et d'autre part à ces dates le nombre total de personnes infectées n'est pas observé directement, seuls les nouveaux cas depuis la date précédente sont comptabilisés. Pour traiter ces spécificités, nous obtenons d'abord une expression explicite des lois de transition à pas fixe pour construire des estimateurs de nos paramètres basés sur les fréquences des transitions pour le nombre d'infectés. Ensuite nous nous plaçons dans le cadre des chaînes de Markov cachées et adaptons l'algorithme de Baum-Welch à notre cas.

Mots-clés. Processus de naissance et mort linéaire avec immigration, Modèle de Markov Caché, Algorithme de Baum-Welch, Estimation paramétrique.

Abstract. The objective of this work is to model the spread of typhoid fever in Mayotte from a set of hospitalization data provided by the Regional Health Agency. We use a linear birth and death process with immigration counting people infected with the disease. The objective is then to estimate the rates of contamination from person to person, contamination by the environment and recovery from the data. The originality of our approach and the difficulty of the problem comes from two sources. On the one hand, the observations are not available in continuous time but only at fixed dates (daily hospitalizations), and on the other hand at these dates the total number of infected people is not observed directly, only the new cases since the previous date are counted. To deal with these specificities, we first obtain an explicit expression of the fixed-step transition distributions to construct estimators of our parameters based on the frequencies of the transitions for the number of infected. Then we place ourselves in the framework of hidden Markov chains and adapt the Baum-Welch algorithm to our case.

Keywords. Linear birth-death process with immigration, Hidden Markov Model, Baum-Welch algorithm, Parametric estimation.

1 Introduction

L'objectif de ce travail est de modéliser la propagation de la fièvre typhoïde à Mayotte à partir d'un jeu de données d'hospitalisations entre 2018 et 2020 fourni par l'Agence Régionale de Santé. La fièvre typhoïde est une maladie à déclaration obligatoire, endémique à Mayotte avec une trentaine de cas par an. Elle se transmet par contact de personne à personne ou par ingestion d'eau ou d'aliments contaminés. Comme le nombre de personnes infectées simultanément est faible, et que l'environnement peut être source de contamination, nous modélisons sa propagation à l'aide d'un processus de naissance et mort linéaire avec immigration comptabilisant les personnes infectées par la maladie. L'objectif est alors d'estimer les taux de contamination de personne à personne (naissance), contamination par l'environnement (immigration) et guérison (mort) à partir des données.

L'originalité de notre approche et la difficulté du problème provient de deux sources. D'une part, les observations ne sont pas disponibles en temps continu mais seulement à des dates fixes (hospitalisations journalières). De plus, à ces dates le nombre total de personnes infectées n'est pas observé, seuls les nouveaux cas depuis la date précédente sont comptabilisés. L'inférence de paramètres pour des processus de naissance et mort linéaires est facile à mener par maximum de vraisemblance lorsqu'on observe les dates de survenue des naissances et des morts et en l'absence d'immigration puisqu'on peut alors utiliser une propriété de branchement (Keiding et al. (1975)). Pour des processus de naissance et mort plus généraux, Crawford et Suchard (2012) et Crawford et al (2014) proposent d'utiliser la théorie des fractions continues et l'algorithme espérance-maximisation (EM), Xu et al (2015) utilisent l'algorithme EM pour l'inférence de processus de branchement multi-types. Dans tous ces travaux, les dates de naissance et de mort sont toujours observées, ce qui les rend inapplicables dans notre cas. Dans une première étape, pour prendre en compte le fait que les observations sont discrètes, nous utilisons les équations de Kolmogorov et les fonctions génératrices des moments des lois de transition à pas fixe pour calculer l'expression explicite de ces probabilités de transition en fonction de nos trois paramètres d'intérêt. Ces formules peuvent ensuite être inversées pour construire des estimateurs basés sur les fréquences des transitions pour le nombre d'infectés. Dans une deuxième étape, pour prendre en compte le fait que seules les nouvelles infections sont comptabilisées, nous nous plaçons dans le cadre des chaînes de Markov cachées et nous adaptons l'algorithme de Baum-Welch (Baum et Petrie (1966)) à nos données. Ce problème est non standard car la loi d'émission des observations ne dépend pas que de l'état courant de la chaîne mais de toute sa trajectoire (continue) depuis la dernière observation.

Ce document est organisé comme suit. Dans la section 2, nous modélisons la dynamique de transmission de la typhoïde par un modèle de naissance et mort avec immigration. Dans la section 3, nous dérivons la procédure d'estimation des paramètres de notre modèle en utilisant la méthode de Markov caché et un algorithme de Baum-Welch adapté à nos observations. Dans la section 4, nous appliquons notre approche d'estimation aux données de Mayotte pour obtenir les estimations désirées et discutons les résultats obtenus.

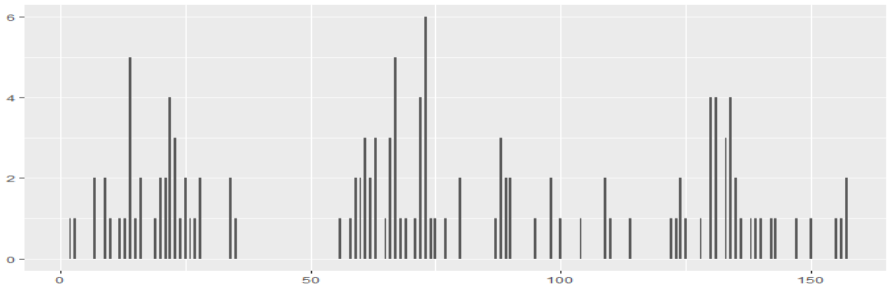

FIGURE 1 – Nombre hebdomadaire de nouveaux cas confirmés de fièvre typhoïde entre 2018 et 2020 à Mayotte (données Agence Régionale de Santé).

2 Modélisation de la dynamique de la typhoïde

La prévalence de la fièvre typhoïde à Mayotte est caractérisée par des niveaux faibles et fluctuants, et des résurgences de la maladie après des plages sans nouvelles infections, voir Fig. 1. Un modèle stochastique de type processus de comptage est donc bien adapté pour modéliser ces données. Parmi ces modèles, l'un des plus simples et pertinent, avec peu de paramètres est le processus de naissance et mort linéaire avec immigration. Un tel processus $(X_t)_{t\geq 0}$ comptabilise le nombre d'individus infectés. Chaque individu contamine un nouvel individu à un taux (de naissance) $\lambda > 0$ et guérit avec un taux (de mort) $\mu > 0$ indépendamment des autres individus, ce qui génère des taux de naissance et mort linéaires pour la population globale. La contamination via l'environnement génère également de nouveaux infectés via un processus de Poisson de taux (d'immigration) $\nu > 0$. Notre objectif est d'estimer ces trois paramètres λ, μ, ν.

Le nombre d'infectés X_t n'est pas directement observé. Nous avons à notre disposition dans le jeu de données uniquement les cumuls de nouvelles infections par jour. Rappelons qu'un processus de naissance et mort ne peut faire que des sauts d'amplitude $+1$ (nouvelle contamination) ou -1 (nouvelle guérison). Notons $N_{]s,t]}$ le nombre des sauts d'amplitude $+1$ du processus $(X_t)_{t\geq 0}$ sur l'intervalle $]s,t]$. Alors le processus joint $(X_t, N_{]0,t]})_{t\geq 0}$ est un processus de Markov de saut pur et les probabilités $p_{i,(j,n)}(t) := \mathbb{P}\left(X_t = j, N_{]0,t]} = n \mid X_0 = i\right)$ satisfont l'équation de Kolmogorov suivante : si $j > i + n$, alors $p_{i,(j,n)}(t) = 0$ et si $j \leq i + n$ alors

$$\frac{dp_{i,(j,n)}(t)}{dt} = -((\lambda+\mu)i+\nu)p_{i,(j,n)}(t) + (\lambda i+\nu)p_{i+1,(j,n-1)}(t) + (\mu i)p_{i-1,(j,n)}(t), \qquad (1)$$

avec la condition initiale $p_{i,(j,n)}(0) = \delta_{i=j}\delta_{n=0}$. Cette équation n'a pas de solution analytique. Nous ne pouvons pas exprimer la distribution marginale des observations en fonction de λ, μ et ν. Ainsi, il n'est pas possible de mettre en oeuvre une méthode d'estimation directe par maximum de vraisemblance avec ce schéma d'observations. Pour pallier ce

problème, dans un premier temps nous allons considérer le problème d'estimation lorsque X_t est observé à des dates régulières. Dans un second temps, nous replacerons notre cadre d'observation dans la famille des modèles de Markov cachés.

Soit Δt un pas de temps fixé ($\Delta t = 1$ jour pour nos données). Alors, le processus $(X_{n\Delta t})_{n\in\mathbb{N}}$ est une chaîne de Markov de matrice de transition $P = (p_{i,j})$. Les coefficients de P peuvent être calculés explicitement en fonction de λ, μ, ν en résolvant une équation de Kolmogorov et en passant par les fonctions génératrices des moments. Comme nous avons 3 paramètres à estimer, et que les nombres de nouveaux cas journaliers observés sont fréquemment nuls, nous utilisons l'expression explicite des paramètres λ, μ et ν en fonction uniquement des probabilités de transition $p_{0,0}, p_{0,1}$ et $p_{1,0}$:

$$\lambda = \frac{ln\left(\frac{u}{p}\right)(p-1)}{\Delta t\,(p-u)}, \quad \mu = \frac{ln\left(\frac{u}{p}\right)(u-1)}{\Delta t\,(p-u)}, \quad \nu = r\lambda, \tag{2}$$

avec

$$p = \frac{p_{0,1}}{p_{0,0}ln(p_{0,0})} W\left(\frac{p_{0,0}^{\left(\frac{p_{0,0}}{p_{0,1}}+1\right)} ln(p_{0,0})}{p_{0,1}}\right), \quad r = \frac{ln(p_{0,0})}{ln(p)}, \quad u = 1 - \frac{p_{1,0}}{p_{0,0}},$$

où W est la fonction W de Lambert. Ainsi, avec une estimation des probabilités de transition $p_{0,0}, p_{0,1}$ et $p_{1,0}$, on obtient une estimation de nos paramètres d'intérêt par plug-in.

3 Estimation des taux de contamination et de guérison

Notons $O_n := N_{[(n-1)\Delta t, n\Delta t[}$ pour $n \in \mathbb{N}^*$ le nombre de nouveaux infectés sur l'intervalle de temps $[(n-1)\Delta t, n\Delta t[$ qui correspond aux observations disponibles. Ici, O_n dépend de toute la trajectoire continue de X_t sur l'intervalle $[(n-1)\Delta t, n\Delta t[$, et pas seulement de $X_{n\Delta t}$. Le schéma d'émission est donc non standard, et l'approche classique doit être adaptée en conséquence. Par la propriété de Markov, on peut montrer que la loi de O_n dépend en fait de $X_{n-1\Delta t}$ et $X_{n\Delta t}$, et que conditionnellement aux $(X_{k\Delta t})_{k\leq n}$, les $(O_k)_{k\leq n}$ sont indépendants, ce qui donne le schéma d'émission de la Figure 2. Posons $Z_n = (X_{(n-1)\Delta t}, X_{n\Delta t})$ pour $n \in \mathbb{N}^*$. Alors $(Z_n)_{n\geq 1}$ est une chaîne de Markov à valeurs dans $\mathbb{N}^2$, et pour cette chaîne, le schéma d'émission des observations (O_n) redevient standard puisque O_n ne dépend que de Z_n. Soit $M = (Q, \psi, \rho)$ les paramètres de ce modèle de Markov caché où Q est la matrice de transition du processus d'état (Z_n), calculée à partir de la propriété de Markov sur le processus $(X_{n\Delta t})$: $Q_{(i,j),(i',j')} = p_{i',j'}\delta_{i'=j}$, ψ est la probabilité d'émission du processus O sachant Z

$$\psi_{(i,j)}(o) = \mathbb{P}(O_n = o | Z_n = (i,j)) = \frac{p_{i,(j,o)}(\Delta t)}{p_{i,j}},$$

et ρ est la loi de l'état initial Z_1 : $\rho_{i,j} = p_{i,j}\pi_i$, où $\pi_i := \mathbb{P}(X_0 = i)$ est la loi initiale de (X_t). En résolvant numériquement (1), nous pouvons calculer la probabilité $p_{i,(j,o)}(\Delta t)$ et en déduire la probabilité d'émission $\psi_{(i,j)}(o)$.

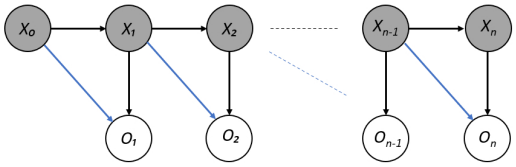

FIGURE 2 – Diagramme d'émission entre les nombre d'infectés X_k à la date $k\Delta t$ et les nombre de nouveaux cas cumulés O_k entre les dates $(k-1)\Delta t$ et $k\Delta t$.

Étant donné une suite d'observations $(O_n)_{0\leq n\leq T}$, nous cherchons à déterminer par maximum de vraisemblance les paramètres $M = (Q, \psi, \rho)$ du modèle, et plus précisément les trois coefficients de la matrice de transition P qui apparaissant dans les équations (2). Malheureusement, l'utilisation directe de l'algorithme de Baum-Welsh sur la chaîne (Z_n) sans tenir compte de sa structure particulière ne donne pas de résultats cohérents sur données simulées. Nous avons donc récrit toutes les formules de récurrence pour les quantités d'intérêt dans le cas particulier où Z_n correspond à deux états successifs d'une chaîne de Markov. Finalement, partant d'un modèle initial $M^{(0)} = (Q^{(0)}, \psi^{(0)}, \rho^{(0)})$, on obtient à chaque étape une nouvelle estimation des paramètres du modèle par les formules

$$p_{i,j}^{(n)} = \frac{1}{N}\sum_{i'=0}^{N} Q_{(i',i),(i,j)}^{(n)},$$

où

$$Q_{(i,j),(i',j')}^{(n)} = \frac{\sum_{t=0}^{T-1}\xi_{(i,j),(i',j')}(t)}{\sum_{t=0}^{T-1}\gamma_{(i,j)}(t)}\delta_{i'=j}, \quad \psi_{(i,j)}^{(n)}(o) = \frac{\sum_{t=0}^{T}\mathbb{1}_{\{O_t=o\}}\gamma_{(i,j)}(t)}{\sum_{t=0}^{T}\gamma_{(i,j)}(t)}, \quad \rho_{i,j}^{(n)} = \gamma_{(i,j)}(0),$$

avec

$$\gamma_{(i,j)}(t) = \frac{\alpha_{(i,j)}(t)\beta_{(i,j)}(t)}{\mathbb{P}(O_{0:T} = o_{0:T}|M^{n-1})}, \quad \xi_{(i,j),(i',j')}(t) = \frac{\alpha_{(i,j)}(t)p_{(i',j')}^{(n-1)}\psi_{(i',j')}^{(n-1)}(o_{t+1})\beta_{(i',j')}(t+1)}{\mathbb{P}(O_{0:T} = o_{0:T}|M^{(n-1)})}\delta_{i'=j}$$

$$\mathbb{P}(O_{0:T} = o_{0:T}|M^{(n-1)}) = \sum_{i=0}^{N}\sum_{j=0}^{N}\alpha_{(i,j)}(t)\beta_{(i,j)}(t)$$

$$\alpha_{(i,j)}(0) = p_{i,j}\pi_i\psi_{(i,j)}^{(n-1)}(o_0), \quad \alpha_{(i,j)}(t) = \psi_{(i,j)}^{(n-1)}(o_t)\, p_{i,j}^{(n-1)}\sum_{i'=0}^{N}\alpha_{(i',i)}(t-1) \quad \text{(Forward probability)}$$

$$\beta_{(i,j)}(T) = 1, \quad \beta_{(i,j)}(t) = \sum_{j'=0}^{N} p_{j,j'}\psi_{(j,j')}^{(n-1)}(o_{t+1})\,\beta_{(j,j')}(t+1) \quad \text{(backward probability)}.$$

L'algorithme recalcule les opérations de ré-estimation M fois jusqu'à un certain critère d'arrêt. Les estimateurs des paramètres λ, μ et ν sont obtenues par le plug-in à partir de $p_{0,0}^{(M)}, p_{0,1}^{(M)}$ et $p_{1,0}^{(M)}$ et des équations (2).

4 Résultats sur les données d'hospitalisation à Mayotte

Nous disposons de 1086 cumuls journaliers de nouveaux cas positifs de la fièvre typhoïde à Mayotte pour les années 2018 à 2020. Ces cumuls compris entre 0 et 3. Comme il s'agit d'une maladie à déclaration obligatoire, nous considérons que tous les nouveaux cas positifs sont bien observés. A partir de la littérature, nous choisissons les paramètres initiaux dans les intervalles suivants $\lambda^{(0)} \in [0.05, 0.08], \mu^{(0)} \in [0.11, 0.25]$ et $\nu^{(0)} \in [0.015, 0.03]$. Pour l'initialisation de $\rho^{(0)}$, nous avons supposé arbitrairement que X_0 suit la loi invariante du processus de naissance et mort de paramètres $\lambda^{(0)}, \mu^{(0)}$ et $\nu^{(0)}$. Les valeurs des paramètres maximisant la vraisemblance des observations pour toutes les valeurs initiales testées sont $\hat{\lambda} = 0.05, \hat{\mu} = 0.11$ et $\hat{\nu} = 0.015$. En particulier, on a $\hat{\lambda} < \hat{\mu}$ qui correspond à un régime récurrent positif sous lequel nos estimations sont valables, et où le processus retourne à 0 infiniment souvent ce qui est cohérent avec les observations. Le temps moyen de guérison $1/\hat{\mu} = 9$ est cohérent avec les durées d'hospitalisations et de traitements, alors que celles-ci n'ont pas été utilisées dans l'estimation.

En conclusion, nous avons présenté une méthode d'estimation générique pour les paramètres d'un processus de naissance et mort avec immigration lorsque seuls les cumuls de nouveaux cas à date périodiques sont observés. Nous avons implémenté cette méthode sur des données de fièvre typhoïde à Mayotte et obtenu des résultats cohérents avec la réalité du terrain. Ces estimations pourront permettre par exemple de déterminer des seuils critiques de nombres de malades à partir desquels une saturation hospitalière risque d'arriver avec forte probabilité dans une fenêtre de temps donnée.

Bibliographie

L. E. Baum and T. Petrie (1966), Statistical inference for probabilistic functions of finite state markov chains, *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554-1563.

N. Keiding et al. (1975), Maximum likelihood estimation in the birth-and-death process, *The Annals of Statistics*, vol. 3, no. 2, pp. 363-372.

F. W. Crawford and M. A. Suchard (2012), Transition probabilities for general birth ?death processes with applications in ecology, genetics, and evolution, *J. of mathematical biology*, vol. 65, no. 3, pp. 553-580.

F. W. Crawford, V. N. Minin, and M. A. Suchard (2014), Estimation for general birth-death processes, *J.of the American Statistical Association*, vol. 109, no. 506, pp. 730-747.

J. Xu, P. Guttorp, M. Kato-Maeda, and V. N. Minin (2015), Likelihood-based inference for discretely observed birth-death-shift processes, with applications to evolution of mobile genetic elements, *Biometrics*, vol. 71, no. 4, pp. 1009-1021.