Estimation indirecte de l'âge en paléodémographie : approche bayésienne

Henri Caussinus, Daniel Courgeau, Isabelle Séguy, Luc Buchet

Estimation indirecte de l'âge en paléodémographie : approche bayésienne

Henri Caussinus, Institut de Mathématiques, Université Paul Sabatier, Toulouse, henri.caussinus@math.univ-toulouse.fr,

Daniel Courgeau, Ined, Paris, courgeau@ined.fr.

Isabelle Séguy, INED /CEPAM, seguy@ined.fr,
Luc Buchet, CEPAM (UMR 6130, Université de Nice Sophia-Antipolis-CNRS) /INED, Valbonne, buchet@cepam.cnrs.fr,

Résumé. En vue d'estimer la structure par âge des populations du passé en ne disposant que d'indicateurs biologiques, les paléodémographes ont développé un certain nombre de méthodes, utilisant une population de référence pour apprécier les probabilités conditionnelles de l'âge connaissant l'indicateur. Compte tenu du faible nombre de données disponibles et du caractère instable du problème, ces méthodes sont en général décevantes. Nous montrons comment les améliorer en introduisant une méthode bayésienne simple intégrant un maximum d'informations non réductibles aux données proprement dites.

Abstract. In paleodemography, the age at death can only be estimated by means of biological indicators and reference data which provide estimates of the conditional distribution of age knowing an indicator. Several methods have been proposed for the estimation of age-structure from these data. Unfortunately, they do not work well in most cases, due to the basically unstable structure of the problem as well as the small size of available samples. We show how to improve the estimation by introducing a simple Bayesian method to incorporate useful prior information on the unknown parameters.

Mots clés : Estimation, méthodes bayésiennes, paléodémographie, lois de Dirichlet, loi de mortalité.

L'âge au décès, notion fondamentale en démographie, ne peut être directement mesuré pour la plupart des populations du passé car elles ne connaissent pas l'état civil. On peut seulement l'estimer d'après des indicateurs biologiques de croissance, pour les sujets immatures, ou de vieillissement, pour les adultes, mesurés sur un échantillon de squelettes appartenant à une population cible donnée. Mais ces indicateurs ne donnent qu'une indication très imprécise de l'âge au décès d'un sujet, car leur corrélation avec l'âge est malheureusement assez faible.

Pour tirer le maximum d'information des données ostéologiques, les paléodémographes utilisent une population, dite de référence, sur laquelle on dispose simultanément pour chaque individu d'une mesure précise de l'âge et de la mesure d'un (ou plusieurs) indicateur(s) biologique(s) (Séguy et Buchet, 2010), souvent prise sur l'os, mais pas exclusivement. Ils partent ensuite d'une « hypothèse d'invariance » selon laquelle la distribution conditionnelle des stades osseux sachant l'âge est la même pour toutes les populations étudiées (au moins en première approximation)[1].

[1] Cette hypothèse est vivement débattue depuis une quinzaine d'années au sein de la communauté des paléodémographes. Séguy et Buchet (2010) concluent : « *Bien que la possibilité d'une dérive séculaire des indicateurs biologiques d'âge ne puisse être écartée, les paléodémographes ont été conduits à la négliger, faute*

1. Formalisation du problème

On traitera ici le cas discret qui est le plus communément considéré, les stades osseux mesurables étant classés en *l* catégories et les âges répartis en *c* classes. La population de référence fournit alors des données n_{ij}, nombre d'individus d'âge *i* et stade *j* (*i = 1,…,l ; j = 1,…,c*). Sur la population cible, on dispose de la répartition des stades osseux : *m* individus sont observés au total dont m_i sont dans le stade *i*.

On désignera par p_{ij} la probabilité qu'un individu pris au hasard dans la population étudiée soit dans le stade *i* et dans la classe d'âge *j ;* la somme sur *i* des p_{ij} sera notée $p_{.j}$ ou simplement p_j (c'est la probabilité qu'un individu soit d'âge *j*) ; la somme sur *j* des p_{ij} sera notée $p_{i.}$ ou simplement π_i (c'est la probabilité qu'un individu soit de stade *i*) ; la probabilité conditionnelle du stade *i* connaissant l'âge *j* sera notée $p_{i|j}$. Ces diverses probabilités sont positives et vérifient les relations : $\sum_i \pi_i = \sum_j p_j = 1$ et $\sum_i p_{i|j} = 1$ pour tout *j*. Elles sont liées d'autre part par la relation suivante :

$$\sum_j p_j \; p_{i|j} = \pi_i \qquad \text{pour tout } i = 1, \ldots, l \qquad [1]$$

Il faut estimer les probabilités p_j, c'est-à-dire la structure par âge de la population cible, au moyen des données décrites plus haut en tenant compte de l'hypothèse d'invariance. Il s'agit donc a priori d'un problème très standard, les données de référence permettant d'estimer les probabilités $p_{i|j}$ et les π_i s'estimant par $\frac{m_i}{m}$. En remplaçant dans [1], on peut obtenir des estimations des p_j par régression (Courgeau, 2010), peut-être convenablement pondérée (il faut aussi noter que les coefficients du premier membre sont en principe entachés d'erreur). On peut aussi utiliser le fait que la distribution des données est modélisable par des lois multinomiales ; on est alors conduit à l'estimation dite IALK (Iterative Age Length Key, Kimura et Chikuni, 1987) si l'on ne tient pas compte des erreurs sur l'estimation des données. On peut aussi tenir compte d'une loi multinomiale sur ces erreurs (Caussinus et Courgeau, 2010). Ces diverses méthodes donnent des résultats décevants pour des raisons variées, la plus importante étant la petite taille des données disponibles, assortie d'une situation intrinsèquement peu stable dans la mesure où la corrélation entre stades et âges est souvent assez faible. On peut cependant améliorer la situation, si l'on tient compte d'informations non exprimées par les données n_{ij} et m_i. On sait en effet que les probabilités p_j à estimer expriment un profil de mortalité qui ne saurait être « n'importe quoi ». Certains auteurs (voir l'ouvrage de Hoppa et Vaupel, 2002) contraignent ainsi les probabilités p_j à satisfaire un modèle paramétrique usuel chez les démographes (comme le modèle de Gompertz ou celui de Gompertz-Makeham, etc.) ; réduire ainsi l'espace paramétrique est évidemment susceptible de réduire la variance des estimateurs, mais peut introduire un biais substantiel. Bocquet-Appel et Bacro (2008) proposent de réduire l'espace paramétrique de façon moins drastique en introduisant un grand nombre de vecteurs de probabilité « de base » (756) constituant une famille de lois de mortalité envisageables ; ils cherchent ensuite à résoudre une régression de type [1] en pondérant convenablement et en *bootstrapant* les données de référence afin de

---

de pouvoir la mesurer, tout en espérant que les éventuelles divergences ne soient pas trop profondes (Pour s'en prémunir ils utilisent des populations de référence préindustrielles qui n'ont pas, ou peu, commencé leur transition démographique).

tenir compte de leur caractère aléatoire. Ils obtiennent ainsi une solution appartenant à l'enveloppe convexe des vecteurs de base. En tant qu'estimation ponctuelle, cette solution est intéressante ; mais il faut noter que les intervalles de confiance associés sont beaucoup trop optimistes car ignorant une part des incertitudes pesant sur les données.

Toutes ces dernières approches consistant à introduire des informations a priori non présentes dans les données, le plus adapté semble d'utiliser une méthode explicitement bayésienne ; c'est l'approche que nous décrivons maintenant.

2. Une nouvelle méthode d'estimation

2.1 Modèle et principe de la méthode

Il est naturel de considérer que les fréquences m_i *(i*=1,…*l)* observées sur le site pour les divers stades, sont les valeurs observées d'une distribution multinomiale dont les paramètres π_i sont liés aux p_j et aux $p_{i|j}$ selon le système [1]. Nous passerons par ces derniers paramètres pour poursuivre la modélisation.

Notons G la densité a priori des paramètres $p_{i|j}$, i = 1,…, l et j = 1,..., c et supposons que les paramètres p_j (*j*=1,…, *c)* ont une densité a priori g et sont indépendants des $p_{i|j}$. Notons M le vecteur des m_i, P le vecteur des $p_{i|j}$ et p le vecteur des p_j. La densité conjointe de *(M,P,p)* est *f* donnée par :

$$f(M,P,p) = g(p)G(P)\frac{m!}{\prod_i m_i!}\prod_i\left(\sum_j p_j\ p_{i|j}\right)^{m_i}.$$

La densité conditionnelle de *p* sachant *M* est :

$$\frac{\int f(M,P,p)dP}{\iint f(M,P,p)dp\,dP}$$

C'est la densité a posteriori des p_j (*j*=1,…*c*) sur laquelle sera basée l'estimation bayésienne.

L'espérance conditionnelle à M d'une fonction φ de *p* sera donnée par

$$\frac{\iint \varphi(p) f(M,P,p)dp\,dP}{\iint f(M,P,p)dp\,dP} \qquad [2]$$

Des choix convenables de φ donnent les moments a posteriori des p_j (en particulier l'espérance qui peut servir d'estimation ponctuelle), ou leur fonction de répartition a posteriori qui servira à obtenir des intervalles de crédibilité (Robert, 2006). Ces diverses intégrales peuvent être évaluées par une méthode de Monte Carlo une fois précisées G et g.

2.2. Choix des lois a priori.

L'information sur les probabilités conditionnelles $p_{i|j}$ vient exclusivement des données de référence ; si celles-ci sont des données brutes simplement obtenues en relevant des fréquences de stades sur un échantillon de squelettes d'âges connus, il est naturel d'admettre

que, pour chaque âge j $(j=1,\ldots,c)$, les fréquences n_{ij} sont les valeurs observées d'une distribution multinomiale de total n_j et probabilités $p_{i|j}$ $(i=1,\ldots,l)$. Comme il y a peu d'information complémentaire sur ces probabilités $p_{i|j}$ au-delà de celle contenue dans les données de référence, il est naturel d'adopter comme loi a priori des $p_{i|j}$, pour chaque *j*, une loi uniforme. Pour un *j* donné, on trouve alors pour loi a posteriori des $p_{i|j}$ une loi de Dirichlet de paramètres $\alpha_{ij} = n_{ij} + 1$ (i=1,..,l). Le produit de ces *c* lois de Dirichlet est alors pris comme loi a priori de *P*, ce qui fournit la densité G.

En pratique, le caractère multinomial des données de référence n'est invoqué que pour fixer les idées : il est seulement indicatif et non indispensable pour « justifier » cette distribution a priori G. On peut d'ailleurs chercher à raffiner ce choix mais cela ne semble pas déterminant en pratique (Caussinus et Courgeau, 2010).

Le choix de la loi a priori des paramètres p_j est plus délicat. Comme la « classe » de lois dans laquelle il convient de chercher la loi a priori ne semble pas s'imposer de façon particulière, le plus naturel (et le plus simple) est d'opter pour une loi de Dirichlet. Se pose alors le seul problème du choix des paramètres, disons $(\beta_1,\ldots,\beta_c)$, de cette loi. Comme il s'agit d'estimer une loi de mortalité et que les paléodémographes ont établi des lois de mortalité standards pour la période préindustrielle (Séguy *et al*, 2008 ; Séguy et Buchet, 2010) une procédure simple consiste à choisir les probabilités correspondantes des diverses classes d'âge pour moyennes de la loi a priori, ce qui donne les paramètres β_j à un coefficient de proportionnalité près. Quelques considérations théoriques et pratiques (simulations) laissent entendre qu'un choix raisonnable est de prendre la somme $\beta_.$ des β_j égale à *c* (mais chercher un meilleur $\beta_.$ autour de cette valeur est évidemment envisageable). Dans certaines situations, il est possible que le paléodémographe possède des informations particulières. Ainsi, dans l'étude d'un cimetière monastique (Maubuisson, Val d'Oise), nous avons tenu compte du fait qu'il s'agissait de femmes a priori en meilleure santé que la moyenne de la population et non soumises à certains risques importants de mortalité pour les plus jeunes, en particulier la mortalité en couche. Notons que le problème essentiel du démographe est de détecter les spécificités de certains sites ; le choix ci-dessus de la loi a priori est alors particulièrement utile, permettant d'apprécier en quoi les données d'un site particulier font évoluer la loi a priori et donc en quoi ce site diffère ou non d'un site que l'on pourrait qualifier de « moyen ».

Au-delà de quelques variantes de l'approche ci-dessus, des approches assez différentes sont évidemment possibles. Par exemple, un peu dans l'esprit des propositions de Bocquet-Appel et Bacro (2008), on peut définir la loi a priori comme une loi uniforme sur un ensemble discret de lois correspondant à des lois types de mortalité. C'est une solution plus lourde à mettre en œuvre que la précédente, mais elle devient d'utilisation aisée lorsque l'effort de construction de tels ensembles de vecteurs a déjà été fait.

2.3. Comparaison aux méthodes antérieures

Notons d'abord que les méthodes fréquentistes ne sont pas utilisables (non identifiabilité) si le nombre *l* de stades est inférieur au nombre *c* des classes d'âge. La méthode proposée ici reste utilisable dans ce cas et peut même s'avérer plus performante avec *c* assez grand (cf. exemple plus bas). Les comparaisons par simulation (Caussinus et Courgeau, 2010a, 2010b) dans des conditions cherchant à représenter le mieux possible les situations réelles ont montré que l'estimation ponctuelle par la moyenne a posteriori était en général bien meilleure que l'estimation fournie par les méthodes antérieures, seule la méthode de Bocquet-Appel et

Bacro (2008) pouvant s'avérer presque comparable, mais étant elle-même supplantée par notre méthode assortie de la loi a priori déduite du travail de ces auteurs (voir plus haut : second choix de la loi a priori des p_j). Cela en dehors des divers autres avantages liés à la considération des lois a posteriori, dont certains sont illustrés dans l'application qui suit.

3. Application

Nous considérons la nécropole de Frénouville (Calvados, époque mérovingienne) pour laquelle les stades osseux de 200 crânes ont été répartis en cinq classes dont les fréquences observées sont $(92 \quad 29 \quad 22 \quad 27 \quad 30)$. L'étude de la répartition par âge a été effectuée pour deux subdivisions. Dans les deux situations, la première classe est 18-19 ans, la dernière classe est 80 ans ou plus, la population des 20-79 ans étant respectivement divisée en classes décennales (8 classes au total) ou quinquennales (14 classes en tout).

La méthode bayésienne proposée a été utilisée avec une loi a priori de Dirichlet pour les p_j qui permet de considérer facilement toute répartition en classes d'âge. Il n'y a ici aucune indication particulière sur la population concernée : les β_j ont donc été choisis proportionnels aux probabilités du standard préindustriel (hommes et femmes réunis), soit pour respectivement 8 et 14 classes :

$$(0{,}02 \quad 0{,}10 \quad 0{,}11 \quad 0{,}13 \quad 0{,}16 \quad 0{,}20 \quad 0{,}19 \quad 0{,}09),$$

$$(0{,}02 \quad 0{,}05 \quad 0{,}05 \quad 0{,}05 \quad 0{,}06 \quad 0{,}06 \quad 0{,}07 \quad 0{,}07 \quad 0{,}09 \quad 0{,}10 \quad 0{,}11 \quad 0{,}11 \quad 0{,}09 \quad 0{,}09),$$

la somme des β_j étant égale au nombre de classes.

Avec 8 classes d'âge nous avons obtenu une loi a posteriori des p_j dont quelques caractéristiques sont indiquées dans le tableau 1. L'écart interquartile est une façon de mesurer la dispersion qui complète très utilement l'écart-type, surtout dans les cas de distributions très dissymétriques comme c'est souvent le cas pour les petites probabilités qui nous intéressent.

Classes	18-19	20-29	30-39	40-49	50-59	60-69	70-79	80 +
moyenne	0.156	0.325	0.067	0.067	0.075	0.135	0.095	0.080
Ecart-type	0.163	0.187	0.068	0.060	0.060	0.079	0.067	0.070
Inter-quart	0.300	0.347	0.072	0.070	0.072	0.116	0.090	0.100
q-05%	0.002	0.020	0.005	0.008	0.010	0.027	0.015	0.004
q-95%	0.446	0.583	0.208	0.186	0.192	0.289	0.227	0.229
q-25%	0.003	0.140	0.019	0.025	0.032	0.075	0.044	0.024
q-75%	0.303	0.487	0.091	0.095	0.104	0.191	0.134	0.124
mi-quart	0.153	0.313	0.055	0.060	0.068	0.133	0.089	0.074
médiane	0.095	0.366	0.046	0.051	0.061	0.125	0.080	0.064

Tableau 1. Exemple de Frénouville, 8 classes d'âge. Quelques caractéristiques de la loi a posteriori.

Les espérances a posteriori (ligne 1 des résultats du tableau) conduisent à une forte révision à la hausse de la mortalité des deux classes les plus jeunes par rapport aux valeurs « a priori » (avec cependant une grande incertitude dont témoignent les écarts-types de la ligne 2 aussi bien que les écarts interquartiles de la ligne 3) et la baisse consécutive pour les autres classes, avec une incertitude plus faible.

Cela peut être précisé en examinant les 8 densités a posteriori comparées aux densités a priori (graphiques non présentés ici). Pour les classes 30-39 à 70-79, les densités a posteriori

sont relativement « pointues » et symétriques, montrant une bonne fiabilité des estimations par la moyenne a posteriori, estimations qui suggèrent une nette diminution de la probabilité de ces classes par rapport au standard préindustriel (impression renforcée par le fait que la légère dissymétrie conduit à des médianes toujours inférieures aux moyennes). Pour la dernière classe, l'estimation est pratiquement celle du standard préindustriel (les lois a priori et a posteriori sont très proches). Pour les deux premières classes, on voit d'abord que les lois a posteriori sont bimodales (un phénomène exceptionnel si l'on en juge par les autres exemples qui ont été traités). On note ensuite une corrélation négative élevée (r = -0,912) si bien que la somme des deux estimations par la moyenne a posteriori a un écart-type de 0,077, du même ordre que celui obtenu pour les autres classes d'âge. On voit donc une sorte d'effet de bascule entre les deux classes conduisant à une mauvaise estimation de la probabilité de chacune mais une bonne estimation de la probabilité de leur regroupement.

On a repris les calculs pour 14 classes. Parmi d'autres, on note alors deux points importants. En premier lieu, le phénomène de bimodalité et de « bascule » mentionné plus haut se retrouve pour les classes 18-19 et 20-24 mais non pour la classe 25-29, précisant d'où viennent les difficultés. D'autre part, l'agrégation des nouveaux résultats pour retourner aux 8 classes précédentes conduit à :

moyennes a posteriori :	(0.120	0.312	0.087	0.083	0.086	0.136	0.108	0.068)
écarts-types a posteriori :	(0.123	0.147	0.067	0.056	0.057	0.063	0.059	0.047)

Si l'on compare aux estimations obtenues directement pour 8 classes, on voit une bonne cohérence des moyennes et des écarts-types plus faibles. Au total il semble donc judicieux de partir d'un nombre de classes assez élevé, quitte à les regrouper ensuite en fonction des résultats obtenus. Une étude plus approfondie reste cependant à faire.

Références

Bocquet-Appel, J.-P., Bacro, J.-N. (2008). Estimation of an age distribution with its confidence intervals using an iterative bayesian procedure and a bootstrap sampling approach, in *Recent Advances in paleodemography*, Bocquet-Appel, J.-P., ed., Dordrecht: Springer-Verlag, p.63-82.

Caussinus, H., Courgeau, D. (2010a). Estimation de la structure par âge des décès : nouvelles propositions, in *Manuel de paléodémographie*, Séguy, I., Buchet, L. eds, Paris : Ined.

Caussinus, H., Courgeau, D. (2010b). *Estimer l'âge sans le mesurer en paléodémographie*. A paraître dans *Population* (vol.1, 2010).

Courgeau, D. (2010). Critiques sur les méthodes actuellement utilisées pour estimer la structure par âge au décès d'une population archéologique, in *Manuel de paléodémographie*, Séguy, I., Buchet, L. eds, Paris : Ined.

Hoppa, R.D., Vaupel, J.W., eds (2002), *Paleodemography. Age distributions from skeletal samples*, Cambridge: Cambridge University press.

Kimura, D.K., Chikuni, S. (1987). Mixture of empirical distributions: an iterative application of the age-length key, *Biometrics*, 43, p. 23-35.

Robert, C. (2006). *Le choix bayésien. Principes et pratique*, Dordrecht: Springer-Verlag, Paris.

Séguy, I., Buchet, L. eds. (2010). *Manuel de paléodémographie*, Paris : Ined,

Séguy, I., Buchet L., Bringé, A. *et al.* (2008). Model life tables for pre-industrial populations. First applications in paleodemography , in *Recent advances in paleodemography. Data, techniques, Patterns*, Bocquet-Appel, J.-P. ed,, Dordrecht: Springer-Verlag, p. 109-141.