Allier géostatistique et algorithme EM pour cartographier la date d'apparition des hommes

Edith Gabriel [1,2] & Frédérik Saltré [3] & Joël Chadœuf [4] & Corey J. A. Bradshaw [5]

[1] *Laboratoire de Mathématiques d'Avignon, F-84000 Avignon, France*
[2] *Unité Biostatistique et Processus Spatiaux, INRA, F-84000 Avignon, France*
[3] *School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia*
[4] *Statistique-GAFL, INRA F-84000, Avignon, France*
[5] *School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, South Australia 5001, Australia*

Résumé. Afin de cartographier les dates d'apparition d'espèces sur la base des seules données fossiles et/ou archéologiques, et donc sur la base d'observations de présence et non de dates d'arrivées aux points d'observation, nous proposons de combiner le modèle de dynamique de population de Verhulst à la géostatistique et d'utiliser l'algorithme EM pour l'estimation de ses paramètres, dont un est le champ aléatoire d'intérêt. Les propriétés de la méthode sont illustrées sur simulations et l'approche est utilisée pour estimer et cartographier les dates d'apparition des premiers hommes modernes (*Homo sapiens*) dans le territoire australien à partir de données archéologiques.

Mots-clés. Algorithme EM, Géostatistique, Modèle de dynamique de population

Abstract. We aimed to map dates of appearance of new species from fossils and/or archaeological data only, by combining the Verhulst population-dynamic model to geostatistical approaches, and to use the EM algorithm to estimate its parameters. We illustrated the properties of our approach by inferring and mapping the date of appearance of *Homo sapiens* in Australia from archeaological records.

Keywords. EM algorithm, Geostatistics, Population dynamics

1 Introduction

Nous nous proposons de cartographier les dates d'apparition dans le territoire australien des premiers hommes modernes (*Homo sapiens*) sur la base des seules données archéologiques. La plupart des études basées sur ces données considèrent que la date d'apparition est équivalente à la plus ancienne date (ou la plus récente dans le cadre des extinctions) de présence observées négligeant tous les biais relatifs aux processus taphonomiques. Si plusieurs méthodes corrigeant ces biais ont été développées pour estimer ces dates d'apparition ou d'extinction dans un contexte non-spatial (cf revue de Wang & Marshall 2016), aucune d'entre elles n'a été adaptée dans un cadre spatial. Quelques méthodes alternatives reposent généralement sur l'usage de données génétiques (Pagani *et al.*, 2016, Reyes-Centeno, 2016), de covariables climatiques (Timmermann & Friedrich, 2016) ou environnementales (voir par exemple MacDonald (2012), Lorenzen (2011) pour l'estimation de la date de dispartition du mamouth). Bird *et al.* (2016) de leur coté estiment la dynamique de colonisation humaine en Australie en utilisant la connectivité des points d'eau. Maier (2016) associe données d'environnement et données éthnographiques pour estimer des probabilités de présence humaine.

Cartographier des dates d'apparition (ou d'extinction) d'espèces sans *a priori* sur les variables qui les dirigent permettrait de choisir objectivement les variables les plus à même à diriger les dates d'apparition dans l'espace, ainsi que le lien statistique entre apparition et variable explicative. L'approche géostatistique (Wackernagel 2003, Cressie 2015) est souvent privilégiée dans ce cadre. Elle allie en effet simplicité (l'estimateur est linéaire), robustesse et bonnes propriétés statistiques (estimateur linéaire de

variance minimale). Son application dans notre cadre se heurte cependant à un problème majeur puisque nous disposons d'observations de présence et non de dates d'arrivée aux points d'observations.

Nous proposons alors de combiner un modèle de développement de population simple, le modèle de Verhulst (Verhuslt, 1838), à l'approche géostatistique. L'estimation du modèle est faite via l'algorithme EM (Dempster *et al.*, 1977) : une étape d'estimation où l'on estime les dates d'apparition humaine conditionnellement à la connaissance du taux de croissance, une étape de maximisation où l'on estime le taux de croissance conditionnellement aux dates d'apparition.

Dans la suite, nous présentons le modèle et sa méthode d'estimation, puis les propriétés de l'approche proposée sur un exemple simulé. L'estimation des dates d'apparition d'*Homo sapiens* en Australie est ensuite présentée. Enfin, nous discutons des possibilités d'extension de la méthode.

2 Estimation

Soit $W \in \mathbb{R}^2$ un compact et $T_m(x)$ le champ aléatoire représentant les dates d'apparition des humains. Nous supposons que ce champ est stationnaire. Nous considérons que la dynamique de population en un point donné suit le modèle de Verhulst. Si $d_o(x)$ est la densité initiale d'humains en $x \in W$ à la date $t = T_m(x)$, la densité $d(x,t)$ d'humains à la date t vérifie

$$d'(x,t) = \beta d(x,t)\left(1 - \frac{d(x,t)}{d_e(x)}\right),$$

où $d_e(x)$ est la capacité d'accueil (densité à l'équilibre) au point x et $d'(x,t)$ désigne la dérivée de $d(x,t)$ par rapport à t. Ainsi,

$$d(x,t) = \frac{d_e(x)d_o(x)}{d_e(x) - d_o(x)} \frac{e^{\beta(t-T_m(x))}}{1 + \frac{d_o(x)}{d_e(x)-d_o(x)} e^{\beta(t-T_m(x))}}.$$

Nous supposons ici que le taux de croissance β en l'absence de compétition est constant dans l'espace.

La probabilité de trouver à la date 0, en un point x, un individu d'âge a est alors

$$p_x(a) = \frac{1}{K(x)} \frac{e^{\beta(T_m(x)-a)}}{1 + \frac{d_o(x)}{d_e(x)-d_o(x)} e^{\beta(T_m(x)-a)}}$$

où $K(x)$ est une constante de normalisation.

En chaque point d'observation x_i on dispose d'un fossile d'âge A_i donné et non d'un âge moyen observé $\tilde{A}_i$. Le champ T_m étant stationnaire, le champ $\tilde{A}$ des âges moyens en x est stationnaire et est estimé par krigeage.

Définissant la date d'arrivée des humains comme celle telle que $\frac{d_o(x)}{d_e(x)-d_o(x)} = 10^{-2}$, l'estimation du modèle va se faire via un algorithme EM. En supposant que les âges sont uniformément distribués sur $[0, T_m(x)]$, nous pouvons obtenir une première estimation de la date d'occupation en doublant le champ moyen, ce qui nous permet d'initialiser l'algorithme. Ensuite, à chaque itération,

- la phase de maximisation correspond à l'estimation par maximum de vraisemblance du paramètre β en supposant connu le champ $T_m(x)$ aux points d'observations,
- la phase d'estimation correspond à l'estimation du champ T_m aux points d'observations sachant le paramètre β par minimisation de l'écart quadratique entre l'âge moyen des fossiles observé en x, $\tilde{A}(x)$, et l'âge moyen théorique $\int_0^{T_m(x)} a p_x(a)\,\mathrm{d}a$.

Notre critère d'arrêt est basé sur la précision obtenue lors de la minimisation des écarts entre âges moyens théoriques et estimés.

3 Exemples simulés

Dans un premier temps nous avons simulé des âges associés à $n = 300$ individus répartis aléatoirement dans le carré unité, à partir du champ aléatoire log-normal $T_m(x)$, de covariance exponentielle de portée 0.25 et de palier 0.05. Dans un second temps, la méthode a été testée dans le cadre d'un gradient linéaire afin de mesurer la robustesse de l'estimation à des écarts à la stationnarité. Pour cela nous avons considéré le champ déterministe $T_m(x_1, x_2) = 0.3 + (x_1^2 + x_2^2)/4$ sur le carré unité, $x = x(x_1, x_2)$. Dans les deux cas nous avons fixé le taux de croissance initial à $\beta = 2$.

La première colonne de la Figure 1 présente la carte des dates d'occupation estimées $\widehat{T}_m(x)$. On retrouve globalement l'allure du champ théorique $T_m(x)$ dans le cas stationnaire (première ligne) et non-stationnaire (deuxième ligne). Le graphique en bas à droite de la Figure 1, où $\hat{T}_m(x)$ est représenté en fonction de $T_m(x)$, confirme ce résultat avec une corrélation de 0.91. Cependant l'estimation se dégrade pour les fortes valeurs de dates d'occupation. Les points en rouge, qui correspondent aux positions des points observés, sont répartis dans tout le nuage. Cela traduit l'absence d'une surdispersion de la prédiction en des points non-observés par rapport aux points observés.

La deuxième colonne de la Figure 1 présente les dynamiques de population théorique et estimée en un site donné occupé il y a 10 ans, dans le cadre stationnaire (première ligne) et non-stationnaire (deuxième ligne). La courbe estimée (en rouge) correspond à une estimation $\hat{\beta}$ =1.91 dans le cas stationnairee et $\hat{\beta}$ =1.65 dans le cas non-stationnaire qui peut se traduire par une montée moins rapide de la population estimée.

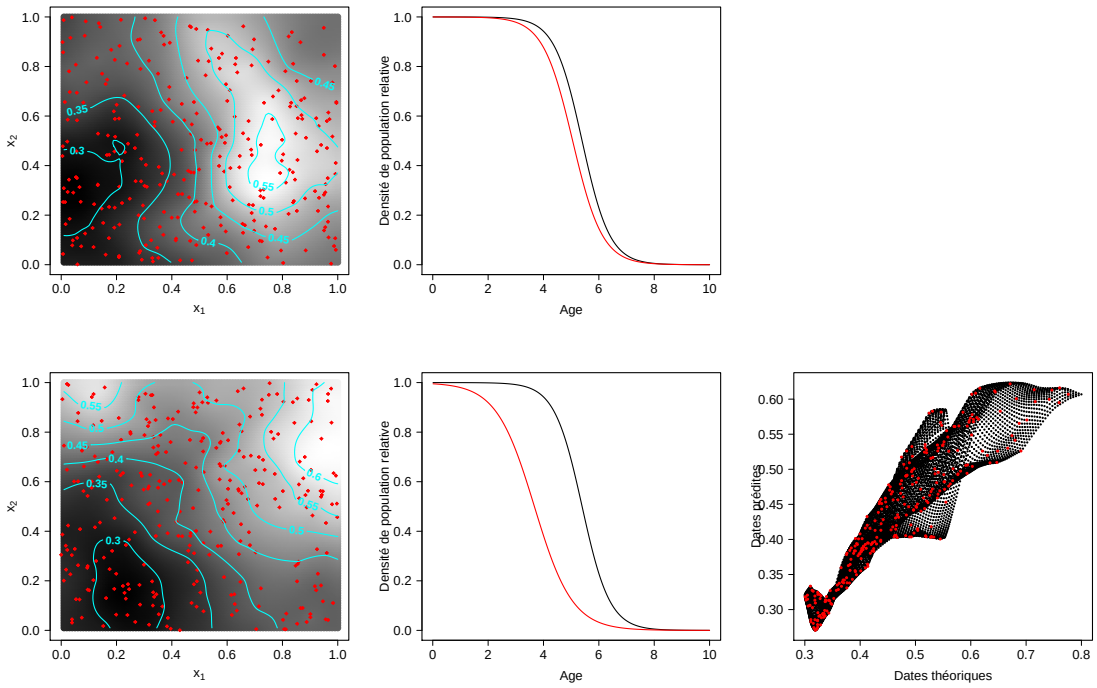

FIGURE 1 – Gauche : dates d'occupation estimées ; Centre : dynamiques de population ; Droite : relation en champ théorique et champ estimé. Première ligne : cas d'un cha aléatoire log-normal ; Deuxième ligne : cas d'un champ déterministe avec présence d'un gradient.

4 Estimation de la date d'occupation de l'Australie par *Homo sapiens*

Nous avons utilisé 283 données archéologiques datées et positionnées attestant de la présence d'*Homo sapiens* en Australie. Ces données sont extraites de la base données FosSahul (Rodríguez-Rey *et al.* 2016). Leur âge s'étend de 73 à 55500 ans. La Figure 2 (gauche) qui présente la distribution des âges montre une forte présence d'individus récents, une proportion relativement constante entre 5000 et 45000 ans, suivie d'une chute de cette proportion pour les individus très âgés.

L'estimation de la date d'occupation, basée sur $\frac{d_o(x)}{d_e(x)-d_o(x)} = 10^{-2}$ donne une date d'occupation allant de 28000 à 100000 ans. La distribution spatiale du champ des âges d'occupation est présentée au centre de la Figure 2. *Homo sapiens* seraient apparus dans l'ouest de l'australie il y a 100000 ans et seraient installés progressivement dans tout le continent, avec une introduction secondaire il y a 80000 ans dans le nord. Notons cependant que peu de fossiles sont disponibles dans l'intérieur du continent ; de ce fait la carte obtenue n'est pas incompatible avec une avancée des humains depuis les bords de l'Australie avec une occupation décalée de l'intérieur.

Le taux de croissance estimé $\hat{\beta}$ est de 1.6, ce qui conduit à la dynamique de population présentée en Figure 2 droite. Une fois installée, la population locale mettrait environ 60000 années à atteindre l'équilibre.

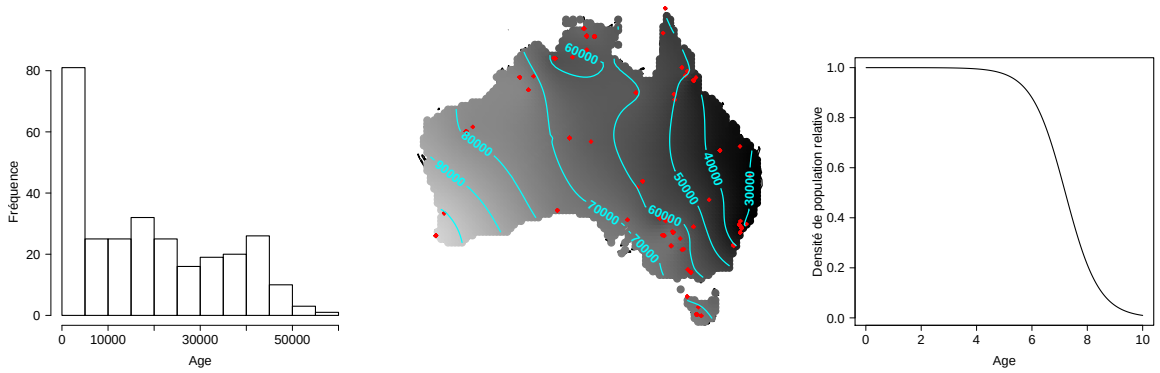

FIGURE 2 – Gauche : distribution des âges des données archéologiques ; Centre : estimation des dates d'occupation d'*Homo sapiens* ; Droite : dynamique de population.

5 Discussion

La méthode proposée permet d'obtenir une estimation sur la base des seules données archéologiques. On dispose ainsi d'un estimateur permettant de choisir objectivement les variables qui dirigent la dynamique d'occupation et peut être leur liaison. On obtient en parallèle une mesure de la dynamique locale des populations. Ce n'est qu'une dynamique apparente car elle ignore les apports par diffusion. Elle est cependant souple car on peut y intégrer des informations complémentaires (modélisation de la capacité d'accueil en fonction des variables d'environnement par exemple) ou des modèles plus réalistes comme la probabilité qu'un fossile arrive jusqu'à nous en fonction de son âge.

La méthode peut également se transposer à la disparition d'espèces mais en adaptant le modèle de dynamique locale. Modéliser la probabilité qu'un fossile d'âge donné arrive jusqu'à nous sera alors

nettement plus important car les âges des fossiles peuvent être nettement plus grands et l'effet de confusion entre dynamique et probabilité de ne pas disparaître (présent dans le cas de l'apparition des humains en Australie) n'existe pas dans ce cas.

Un des points délicats de la méthode proposée est celui de la définition de la date d'apparition. Nous l'avons définie comme la date où le rapport entre densités initiale et à l'équilibre vaut 0.01. La valeur absolue des dates d'apparition est directement liée à la valeur à laquelle on fixe ce taux. Cette définition est cruciale est pose la question de savoir ce qu'on appelle occuper un territoire.

L'utilisation de l'EM en géostatistique n'est pas nouvelle. Zhang (2007) l'a développé pour des variables gaussiennes, Ferrari & Minozzo (2013) à des données de comptage par exemple passant alors à un MCEM. Dans les deux cas le problème posé est celui de l'estimation de la structure de covariance dans des cas multivariables. Le cas qui nous intéresse diffère sur deux points : un des paramètres d'intérêt est un champ complet, et la variable observée n'a pas une relation simple (au sens statistique) avec les paramètres à estimer. De ce fait, au-delà de l'obtention d'un estimateur intuitif, les propriétés statistiques de ce dernier restent à établir. Celles-ci vont dépendre

- de la loi de l'estimateur de l'espérance de la moyenne basée sur le krigeage des âges observés,
- de la loi de l'estimateur du taux de croissance obtenu par EM basé sur les âges observés, dont les lois sont stationnaires mais non-iid,
- de l'erreur de mesure liée aux méthodes de datation.

Bibliographie

[1] Bird MI, O'Grady D & Ulm S (2016). Humans, water, and the colonization of Australia. *Proceedings of the National Academy of the USA*, 113(41), 11477–11482.

[2] Cressie N (2015). *Statistics for Spatial Data*, Revised Edition. Wiley Series in Probability and Statistics.

[3] Dempster AP, Laird NM & Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1–38.

[4] Ferrari, C & Minozzo M (2013). Multivariate geostatistical mapping of radioactive contamination in the Maddalena Archipelago (Sardinia, Italy) : spatial special issue. *AStA Advances in Statistical Analysis*, 97(2), 195–213.

[5] Lorenzen ED, Nogues-Bravo D, Orlando L, Weinstock J, Binladen J, Marske KA, Ugan A, Borregaard MK, Gilbert MTP, Nielsen R, Ho SYW, Goebel T, Graf KE, Byers D, Stenderup JT, Rasmussen M, Campos PF, Leonard JA, Koepfli KP, Froese D, Zazula G, Stafford Jr TW, Aaris-Sorensen K, Batra P, Haywood AM, Singarayer JS, Valdes PJ, Boeskorov G, Burns JA, Davydov SP, Haile J, Jenkins DL, Kosintsev P, Kuznetsova T, Lai X, Martin LD, McDonald HG, Mol D, Meldgaard M, Munch K, Stephan E, Sablin M, Sommer RS, Sipko T, Scott E, Suchard MA, Tikhonov A, Willerslev R, Wayne RK, Cooper A, Hofreiter M, Sher A, Shapiro B, Rahbek C & Willerslev E. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, 479, 359–364.

[6] MacDonald GM, Beilman DW, Kuzmin YV, Orlova LA, Kremenetski KV, Shapiro B, Wayne RK & Van Valkenburgh B (2012). Pattern of extinction of the woolly mammoth in Beringia. *Nature Communications*. 3(893), 1–8.

[7] Maier A, Lehmkuhl F, Ludwig P, Melles M, Schmidt I, Shao Y, Zeeden C & Zimmermann A (2016). Demographic estimates of hunteregatherers during the Last Glacial Maximum in Europe against the background of palaeoenvironmental data. *Quaternary International*, 425, 49–61.

[8] Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, Wall JD, Cardona A, Mägi R, Sayres MAW, Kaewert S, Inchley C, Scheib CL, Järve M, Karmin M, Jacobs GS, Antao T, Iliescu FM, Kushniarevich A, Ayub Q, Tyler-Smith C, Xue Y, Yunusbayev B, Tambets K, Mallick CB, Saag L, Pocheshkhova E, Andriadze G, Muller C, Westaway MC, Lambert DM, Zoraqi G, Turdikulova S, Dalimova D, Sabitov Z, Sultana GNN, Lachance J, Tishkoff S, Momynaliev K, Isakova J, Damba LD, Gubina M, Nymadawa P, Evseeva I, Atramentova L, Utevska

O, Ricaut FX, Brucato N, Sudoyo H, Letellier T, Cox MP, Barashkov NA, Škaro V, Mulahasanovic´ L, Primorac D, Sahakyan H, Mormina M, Eichstaedt CA, Lichman DV, Abdullah S, Chaubey G, Wee JTS, Mihailov E, Karunas A, Litvinov S, Khusainova R, Ekomasova N, Akhmetova V, Khidiyatova I, Marjanović D, Yepiskoposyan L, Behar DM, Balanovska E, Metspalu A, Derenko M, Malyarchuk B, Voevoda M, Fedorova SA, Osipova LP, Lahr MM, Gerbault P, Leavesley M, Migliano AB, Petraglia M, Balanovsky O, Khusnutdinova EK, Metspalu E, Thomas MG, Manica A, Nielsen R, Villems R, Willerslev E, Kivisild T & M. Metspalu M (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538, 238–242.

[9] Reyes-Centeno H (2016). Out of Africa and into Asia : fossil and genetic evidence on modern human origins and dispersals. *Quaternary international*, 416, 249–262.

[10] Rodríguez-Rey M, Herrando-Pérez S, Brook BW, Saltré F, Alroy J, Beeton N, Bird MI, Cooper A, Gillespie R, Jacobs Z, Johnson CN, Miller GH, Prideaux GJ, Roberts RG, Turney CSM & Bradshaw CJA(2016). FosSahul : a comprehensive database of quality-rated fossil ages for Sahul's Quaternary vertebrates. *Scientific Data*, 3, 160053

[11] Timmermann A & Friedrich T (2016). Late Pleistocene climate drivers of early human migration *Nature*, 538, 92–95.

[12] Verhulst PF (1838). Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathématique et physique*, 10, 113–121.

[13] Wang S & Marshall C (2016). Estimating times of extinction in the fossil record *Biology Letters*, 12(4), 1–5.

[14] Wackernagel H (2003). *Multivariate Geostatistics : An Introduction with Applications*. Springer.

[15] Zhang H (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, 18, 125–139.