Chapitre 1.3

CeraRdate : un modèle statistique de datation des contextes archéologiques à partir de la céramique

Lise BELLANGER(1), Philippe HUSI(2),
Pascal CHAREILLE(3), Arthur COULON (2)

(1) Nantes Université, Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, 2 rue de la Houssinière BP 92208, 44322 Nantes Cedex 03, France, (2) CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires, 40 rue James Watt, ActiCampus 1, 37200 Tours, France, (3) Université de Tours, Centre Tourangeau d'Histoire et d'études des Sources EA 6298, 3 rue des Tanneurs BP 4103, 37041 Tours Cedex, France.

CeraRdate présenté ici fait partie d'un projet plus vaste nommé SPARTAAS : Statistical Pattern Recognition and daTing using Archeological Artefacts assemblageS (Coulon, Bellanger and Husi, 2021). Il a pour objectif de développer des outils statistiques dédiés à l'analyse des données mobilières volumineuses dans une démarche de modélisation archéo-statistique. Ce projet est adossé à un package R déposé sur le CRAN [1] et à un site internet (https ://spartaas.gitpages.huma-num.fr/r-package). Les outils implémentés concernent la datation *(CeraRdate)*, la périodisation (hclustcompro) et la spatialisation (mapclust). L'outil hclustcompro est une aide à la périodisation des ensembles stratigraphiques et des sites. Il est fondé sur une classification ascendante hiérarchique (CAH) prenant en compte deux sources d'information associées aux mêmes objets (céramique/mobilier et stratigraphie ; céramique/mobilier et datation des ensembles stratigraphiques…). Cette méthode, appelée CAH par compromis (hclustcompro), permet un compromis entre les hiérarchies obtenues à partir de chaque source prise séparément. Une combinaison convexe des dissimilarités associées à chacune des sources est utilisée pour modifier la mesure de dissimilarité dans l'algorithme CAH classique (Bellanger, Coulon and Husi, 2020, 2021a). L'outil mapclust est une aide à la spatialisation des sites via les assemblages céramiques. Cette méthode est fondée sur une classification descendante hiérarchique (CDH) avec des

[1] https ://cran.r-project.org/package=SPARTAAS

contraintes géographiques (Bellanger, Coulon and Husi, 2021b). Afin de mieux comprendre la distribution spatiale des individus (unités statistiques), la construction de patchs spatiaux (Woillez, Rivoirard and Petitgras, 2009) est appliquée dans une classification descendante hiérarchique (CDH) permettant ainsi de visualiser et d'analyser facilement des répartitions géographiques des sites. Cette approche est essentielle à la construction d'aires socio-économiques ou culturelles à partir du mobilier archéologique, source essentielle pour comprendre ces phénomènes dans le temps long. Ces deux derniers outils ont largement été mobilisés pour l'élaboration de la périodisation et la construction d'aires économiques et culturelles du bassin de la Loire moyenne à partir de la céramique entre le 6 et le 19e s. Cette recherche publiée en ligne, au format logiciste permettant une lecture rapide et à plusieurs niveaux du raisonnement, des résultats et de l'argumentaire, est un bon exemple d'utilisation de hclustcompro et mapclust à partir d'un important corpus céramique (Husi, 2022). Nous avons fait le choix de ne présenter dans ce chapitre que l'outil de datation *CeraRdate* pour des données céramiques également issues de sites archéologiques du bassin de la Loire moyenne.

1 Présentation de la problématique

La question du temps, omniprésente en archéologie, met en jeu le rapport à l'objet et au contexte archéologique. Elle peut s'appréhender de trois manières différentes selon qu'on s'interroge sur :

- La durée, à relier au rythme et donc à l'intensité d'occupation ;
- La succession (ou encore l'évolution), à relier à la chronologie relative ;
- La date (ou la datation), à relier à la chronologie absolue.

Ces trois aspects du temps sont souvent, à tort, confondus par l'archéologue, qui les différencie mal en raison de leurs liens étroits (Ferdière, 2007).

Les méthodes de datation sont fondées sur deux types d'objets : (i) ceux – rares – qui sont porteurs de leur propre date (monnaie ou documents épigraphiques par ex.) et (ii) ceux qui ne sont datés que par référence à une chrono-typologie (céramique ou relative à tout autre type de mobilier). L'objet (i) retenu pour dater fournit au contexte étudié un *terminus post quem* (ou borne inférieure). C'est la chronologie relative entre contextes stratigraphiques qui permet de refermer l'intervalle des datations possibles pour un contexte donné, en fournissant le *terminus ante quem* (ou borne supérieure) (Ferdière, 2007).

Bien que, le plus souvent, aucune date ne soit inscrite sur des récipients en terre cuite, la céramique reste en pratique l'un des meilleurs marqueurs pour préciser la chronologie relative des sites archéologiques. Elle dispose en effets de trois atouts majeurs : une présence régulière et importante en fouille (les récipients domestiques d'usage courant sont peu chers, facilement remplaçables, donc massivement rejetés) ; une nature quasi-indestructible au cours du temps (l'argile cuite est ainsi un matériau omniprésent en archéologie) ; des changements typologiques rapides au cours du temps (la vaisselle en terre cuite permet de disposer de jalons chronologiques identifiables dans la longue durée).

Il est acquis que la céramique permet de dater des ensembles stratigraphiques sans relation physique, issus d'un ou plusieurs sites, par la comparaison des assemblages céramiques qui les constituent. Ces assemblages sont composés de tessons de céramique issus de productions (groupes techniques) différentes, avec leur propre cycle de vie, mais qui ne sont pas intrinsèquement datés dans le temps calendaire. Deux types de contextes archéologiques peuvent être distingués : (i) les ensembles dits *actifs* (ou de référence), car datés par un élément externe à la céramique (par exemple la date d'émission de la monnaie la plus récente *in situ*) ; (ii) les ensembles dits *supplémentaires*, sans date puisque sans information en lien avec la chronologie absolue.

CeraRdate est un modèle* statistique de datation des contextes archéologiques qui permet, à partir des ensembles de référence, de prévoir la datation des ensembles supplémentaires. La comparaison statistique des faciès céramiques issus des assemblages quantifiés par groupes techniques (productions) permet de construire ce modèle.[2]

2 Le corpus des données archéologiques

Le choix du corpus est une étape importante qui détermine fortement les interprétations archéologiques ultérieures. Il est indispensable de ne conserver que les contextes archéologiques (ou ensembles) les moins perturbés, donc ceux susceptibles de révéler la plus grande quantité de matériel contemporain de l'action interprétée par l'archéologue, autrement dit, éviter par exemple les contextes de remblais fortement perturbés.

Le corpus choisi ici pour présenter la méthode est issu d'un projet de recherche sur la céramique médiévale et moderne du Centre-Ouest de la France du 6^{e} au 17^{e} s. (Husi, 2003, 2013). Il est composé de 278 contextes archéologiques (ou ensembles), répartis en 95 contextes *actifs* de Tours et 183 contextes *supplémentaires* (qui ne participent donc pas à la construction du modèle) (Bellanger and Husi, 2012).

La construction de ce corpus suppose l'identification, l'inventaire et l'individualisation des groupes techniques pertinents pour la zone d'étude. Seuls 200 parmi ceux qui ont pu être identifiés ont été retenus : ils sont ceux qui reflètent les caractéristiques communes à la totalité des ensembles stratigraphiques étudiés. Leur choix permet de traiter, suivant une méthodologie standardisée, un corpus de 15044 individus (quantification en Nombre Minimum d'Individus (NMI), préférée par exemple au Nombre de Reste (NR), trop sensible à la fragmentation).

Une telle approche suppose de longs efforts d'analyse de la céramique. L'harmonisation des méthodes de traitement est en effet un préalable à la réussite de l'entreprise. Elle a été réalisée par un petit groupe de chercheurs qui travaillent cette question pour mettre en place des outils typologiques et des techniques de quantification communs pour tout l'Ouest de la France. Afin d'élargir le champ d'investigation, le réseau de chercheurs s'est ouvert désormais à l'espace européen francophone. Le groupe est fédéré autour d'un site Internet conçu comme une base de données spatialisées,

[2] Pour plus de détails sur la démarche, le lecteur pourra se référer à : Husi, Tomassone and Chareille, 2000 ; Bellanger, Husi, Tomassone, 2006a,b ; Bellanger, Husi and Tomassone, 2008 ; Bellanger and Husi, 2012)

ICERAMM [3], mettant en ligne les outils typologiques régionaux communs et les notices de sites.

3 Modélisation statistique : construction d'un modèle céramique

Proposer la datation d'un contexte archéologique s'avère délicat et nécessite de ne pas perdre de vue ce que l'on cherche à dater : l'action ponctuelle inscrite dans un temps événementiel, ou le processus d'accumulation inscrit dans une durée.

Dans ce travail, nous proposons des datations résultant d'une modélisation statistique de l'ensemble des données céramiques, plutôt que des datations « traditionnelles » dont la justification est souvent peu argumentée et fondée sur une comparaison intuitive des assemblages céramiques. Il ne s'agit évidemment pas ici de nier l'importance des connaissances du céramologue dans le processus de datation, mais bien au contraire de mieux les prendre en compte en les intégrant dans une démarche systémique, indispensable lorsque la masse de données à analyser devient trop importante.

La méthode repose sur une approche statistique et visuelle. Pour chaque contexte archéologique (ou ensemble), deux courbes de densité estimées sont construites. La première permet d'estimer une date correspondant au *terminus post quem* de l'ensemble, un curseur reflétant le temps événementiel. Nous ne la décrirons ici que succinctement et renvoyons le lecteur aux travaux cités ainsi qu'à l'annexe. La seconde, qui utilise les résultats de la première étape, permet d'estimer le profil chronologique de l'ensemble, image plus proche du temps archéologique (ou temps de l'accumulation).

3.1 Étape 1 : modélisation du temps événementiel (*dateEv)*

Il s'agit d'estimer la date d'une action inscrite dans le sol – pour l'archéologue, d'un contexte archéologique – en fonction de l'assemblage céramique qui le compose. La méthode *CeraRdate* consiste à construire un modèle de régression linéaire liant une date connue dans le temps calendaire – ici celle de l'émission d'une monnaie –, à son faciès céramique.

Dans ce travail, le corpus de référence ayant permis d'ajuster le modèle est constitué des ensembles de la ville de Tours qui possèdent des monnaies et qui présentent une excellente qualité chrono-stratigraphique. Le modèle d'ajustement permet de calculer une prévision ponctuelle et une estimation par intervalle de confiance (par ex. à 95%) de la datation d'ensembles extérieurs au corpus de référence, parfois stratigraphiquement isolés ou mal documentés.

La démarche statistique est donc la suivante :

1. Synthèse de l'information contenue dans le corpus (de données céramiques) de référence : on effectue une *Analyse factorielle des correspondances* (AFC) (*cf.* 6) sur la table de contingence (95x200) croisant les 95 contextes archéologiques (ensembles actifs) et les 200 groupes techniques quantifiés en NMI. Puis, nous conservons uniquement les premiers axes (dans notre exemple, les 10 premiers axes factoriels qui expliquent environ 64% de l'inertie totale du nuage de points). Ainsi nous réduisons

[3] Information sur la CERAmique Médiévale et Moderne : http://iceramm.univ-tours.fr/

la table initiale 95x200 à un tableau 95x10, *reconstitution incomplète mais suffisante des données*[4].

2. Estimation/prévision d'une date pour chaque ensemble : on ajuste un modèle de régression linéaire à partir des ensembles datés par des objets (ici des monnaies en contexte). Parmi les 95 ensembles actifs de Tours, affichant tous une grande fiabilité chrono-stratigraphique (niveaux d'occupation ou ensembles clos issus de séquences stratigraphiques urbaines), seuls 25 peuvent être datés de cette manière. Les biais liés à un décalage entre la date de la monnaie et celle du contexte ont ainsi été minimisés. En outre, ces 25 ensembles jalonnent la totalité de la fourchette chronologique entre le 6$^{\text{e}}$ et le 17$^{\text{e}}$ s. et présentent une bonne répartition au sein des ensembles datés. Un modèle de régression linéaire multiple gaussien (*cf.* équation (1)) est ajusté sur ces 25 ensembles en fonction des composantes factorielles de l'AFC significatives parmi les 10 conservées, soit ici les 8 premières uniquement, notées $\boldsymbol{F}^k (k = 1, \ldots, 8)$. Il peut s'écrire sous la forme :

$$dateEv_i = \beta_0 + \sum_{k=1}^{8} \beta_k F_i^k + \varepsilon_i \quad \forall i = 1, \cdots, 25 \tag{1}$$

où les ε_i sont indépendantes* et Normalement distribuées suivant une loi $\mathcal{N}\left(0; \sigma^2\right)$ et F_i^k est la $i^{\text{ième}}$ coordonnée de la $k^{\text{ième}}$ composante factorielle ($i = 1, \ldots, 25$ et $k = 1, \ldots, 8$).

Après avoir estimé les paramètres β_k du modèle en utilisant les résultats classiques de la régression linéaire multiple, et vérifié que le modèle s'ajustait bien aux données (on obtient ici un R^2 ajusté d'environ 0.9973, donc très proche de 1, et un écart-type résiduel de 23 ans environ !), nous pouvons inférer sur la datation d'un ensemble actif, mais aussi prévoir celle d'un ensemble supplémentaire (dans lequel aucune monnaie n'a été retrouvée, donc non daté). Cette estimation (ou prévision, quand aucune monnaie n'a été retrouvée dans le contexte), notée *dateEv*, se construit uniquement à partir de la part d'information contenue dans le faciès céramique de l'ensemble étudié liée à la variable réponse du modèle de régression (date de la monnaie), c'est-à-dire ici à partir des 8 premières composantes factorielles significatives de l'AFC.

[4]Ce principe est appliqué dans de nombreuses techniques d'analyse factorielle. Il permet de réduire le nombre de variables explicatives prises en compte dans le modèle. Dans notre cas, 8 au lieu de 200 ! (Benzécri, 1973a,b ; Moreau, Doudin and Cazes, 2000 ; Saporta, 2011 ; Bellanger and Tomassone, 2014 ; Greenacre, 2017)

À la fin de cette 1re étape, on obtient donc :

- une estimation ponctuelle et par intervalle de confiance* (par ex. à 95%) du temps événementiel des ensembles actifs ;
- une prévision ponctuelle et par intervalle de confiance (par ex. à 95%) du temps événementiel des ensembles supplémentaires, donc non datés, provenant ou non d'autres sites que ceux utilisés pour la construction du modèle. Les résultats statistiques sont alors validés par des arguments externes tels que la présence de monnaies (non prise en compte dans le modèle), et la logique des arguments stratigraphiques habituels.
- la densité* de probabilité estimée correspondant au temps événementiel de chaque ensemble, construite en utilisant les résultats classiques du modèle linéaire gaussien. Celle-ci peut être approchée par une loi Normale dont les paramètres varient suivant l'ensemble (*cf.* ANNEXE). Elle sera représentée par **une courbe (de densité) orange** sur les figures à suivre. Par définition l'aire sous cette courbe de densité vaut 1 (donc 100%).

Exemple

Pour le contexte archéologique D1 (niveau de jardin utilisé pour rejeter des déchets domestiques) daté par une monnaie de 1341, on obtient un intervalle à 95% pour *dateEv* de [1333 ;1381] et la courbe de densité unimodale suivante (Figure 1).

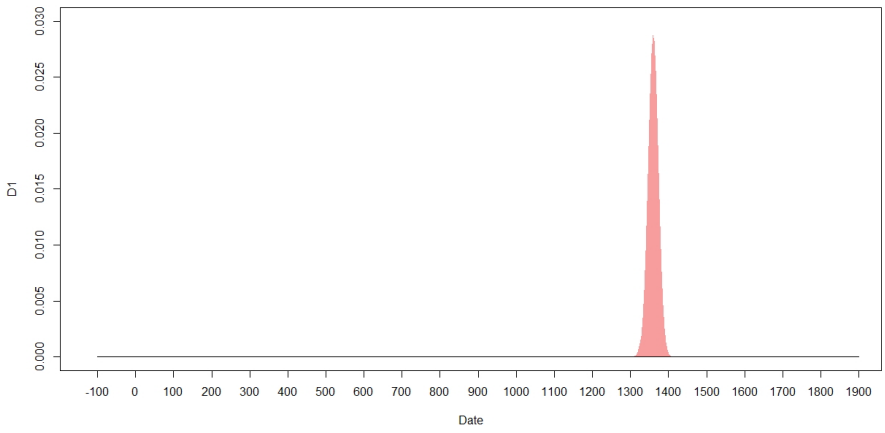


Fig. 1 : Courbe de densité *(dateEv)* du dépotoir D1

La deuxième étape consiste à estimer une datation qui, d'un point de vue archéologique, représente mieux le temps de l'accumulation inscrit dans la matière (Olivier, 2001). Cette datation pourra s'interpréter au mieux comme un processus de formation reflétant une durée (ou une succession) à l'échelle du temps archéologique si la qualité du contexte archéologique le permet, au pire comme une imprécision dans la datation, révélant une forte pollution de l'ensemble stratigraphique par la présence de matériel redéposé ou intrusif.

3.2 Étape 2 : du temps événementiel au temps de l'accumulation (*dateAc*)

Les résultats obtenus à l'étape 1 et les propriétés de l'AFC sont utilisés pour obtenir une estimation ponctuelle de la datation de chaque groupe technique ($dateEv^j$; $j = 1, ..., 200$) ainsi qu'un intervalle de confiance à 95%. On définit alors le *temps archéologique*, noté *dateAc* (ou temps de l'accumulation) d'un ensemble comme la somme pondérée des datations des groupes techniques qui le compose : les poids étant les proportions des Nombre Minimum d'Individus (NMI) de chaque groupe technique présent dans l'ensemble. En supposant que les variables* aléatoires $dateEv^j$ sont indépendantes (par construction des GT, hypothèse réaliste), on peut approcher la loi du temps de l'accumulation d'un ensemble quelconque par un mélange de gaussiennes.

On obtient ainsi, pour chaque ensemble, une courbe plurimodale, les assemblages céramiques n'étant jamais constitués d'un seul groupe technique. Cette courbe représente l'estimation de la loi du temps de l'accumulation fondée sur ce mélange de densités unimodales gaussiennes correspondant à la datation de chaque groupe technique (*cf.* ANNEXE).

Le tracé de cette loi correspond à la **courbe (de densité) noire** sur les figures à suivre ; par définition, l'aire sous cette courbe de densité vaut 1 (donc 100%).

Exemple

Pour le contexte archéologique D1, daté par une monnaie 1341, on obtient un intervalle à 95% pour *dateAc* de [1206 ;1499] et la courbe de densité plurimodale suivante (Figure 2).

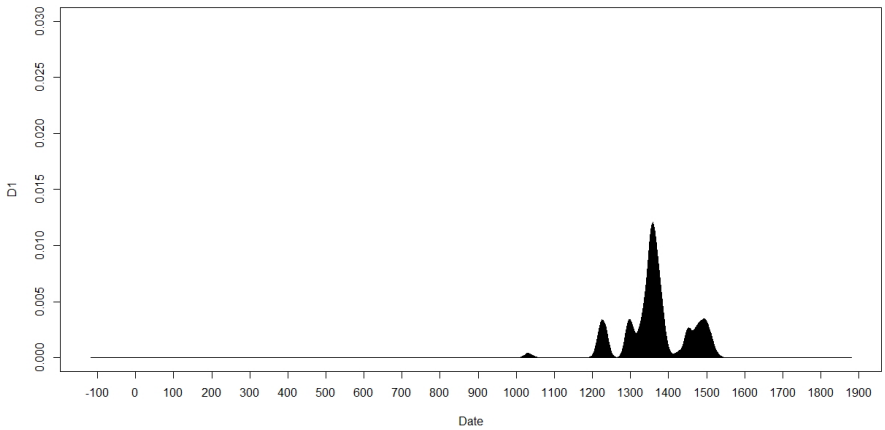


FIG. 2 : Courbe de densité *(dateAc)* du dépotoir D1 (Tours)

La dernière étape consiste à confronter les deux courbes de densité. Il est ainsi possible d'analyser plus finement les datations proposées d'un point de vue chronologique en explorant la « durée » (ou l'intensité de l'occupation) de chaque contexte. L'interprétation des courbes de densité orange et noire se fait visuellement à l'aide d'un graphe combinant les deux courbes (Figure 3).

Exemple

Pour le contexte archéologique D1, on observe une superposition de l'unique pic (ou mode) de la courbe orange (estimation de *dateEv*) et du pic principal de la courbe noire (estimation de *dateAc*). Les ondulations mineures visibles de part et d'autre du pic principal de la densité associée à *dateAc* s'interprètent en fonction de la nature de l'ensemble, ici comme une possible représentation d'un matériel redéposé ou intrusif et/ou comme l'existence d'une occupation/utilisation plus longue du dépotoir (inscrite dans la matière céramique). On peut interpréter le moment où les deux pics se superposent comme celui qui caractérise le moment d'activité ou d'utilisation principale du contexte. En admettant cette hypothèse, on peut imaginer une occupation qui s'inscrit dans la durée, plus que dans l'action ponctuelle, sans pour autant avoir la preuve que la largeur du pic principal de la courbe noire représente la totalité de la durée d'occupation. C'est toute l'ambiguïté entre les notions d'accumulation d'événements et de durée d'occupation !

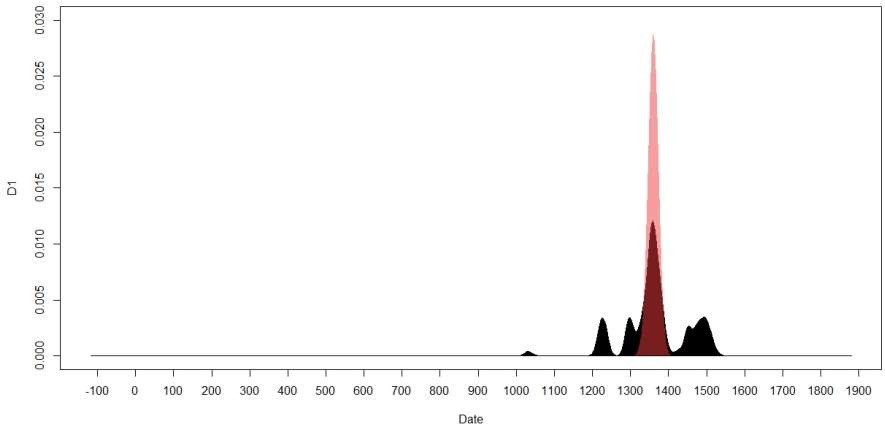


Fig. 3 : Superposition des courbes de densité *(dateEv et dateAc)* du dépotoir D1 (Tours)

Certains contextes archéologiques peuvent s'avérer être des "cas d'école", avec deux courbes de densité présentant un profil quasi identique. C'est le cas par exemple pour une occupation domestique d'un site de Tours (ensemble J1). La courbe orange (qui représente la date d'une monnaie) et la courbe noire (qui reflète le faciès chronologique de l'ensemble) se superposent de manière quasi parfaite (Figure 4). On peut ici en conclure que le matériel redéposé (ou intrusif polluant) est marginal et que les estimations des dates des groupes techniques se concentrent en majorité entre la fin du 15^{e} et le début du 16^{e} s. (*cf.* courbe noire). En outre, s'il existait une monnaie dans cet ensemble, il y aurait 95 % de chances qu'elle appartienne à cette même fourchette chronologique (*cf.* courbe orange). C'est en comparant ces résultats à la stratigraphie et à la réalité de la composition de cet assemblage céramique que l'on peut proposer ici une fourchette chronologique resserrée, avec une accumulation du matériel dans un laps de temps probablement assez court. Cet exemple reste un cas d'école et de telles superpositions presque parfaites sont rares. Il est important de ne pas perdre vue que

cet outil de datation statistique doit être utilisé comme une aide à l'interprétation chronologique des ensembles et non comme un outil de datation automatique.

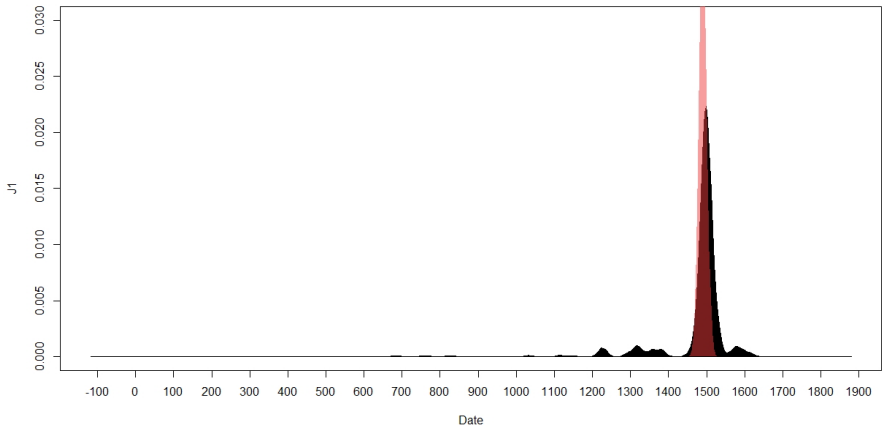


FIG. 4 : Courbes de densité *(dateEv* et *dateAc)* du dépotoir J1 (Tours)

4 Limites du modèle

Les aires socio-économiques et culturelles fondées sur les faciès de la céramique ne dépassent que rarement une quarantaine de kilomètres autour des centres de consommation principaux, souvent constitués par les villes, révélant la dimension locale d'un produit sans grande valeur ajoutée (Bellanger and Husi, 2012 ; Husi, 2013). L'application du modèle *CeraRdate*, construit à partir du référentiel de Tours, à un contexte archéologique d'un site localisé à Poitiers, ville située à environ 120 km au Sud de Tours, permet de montrer les limites de la méthode.

Exemple

L'application du modèle *CeraRdate* à une fosse dépotoir poitevine, choisie pour la qualité de son assemblage céramique et pour la présence d'une monnaie datable de la fin du 10^{e} s. permet de construire deux courbes de densité qui montrent clairement l'incohérence des résultats. La courbe orange conduit à une estimation pour *dateEv* début 11^{e} s., en totale contradiction avec le pic principal de la courbe noire, situé début 13^{e} s. (Figure 5). La distance entre Poitiers et le site de référence, Tours, mais aussi l'appartenance de ces deux villes à des entités culturelles différentes, sont les principales causes de cette distorsion chronologique. Certaines productions peuvent présenter des caractéristiques techniques et esthétiques proches d'un espace économique à l'autre, mais correspondre à des périodes chronologiques différentes.

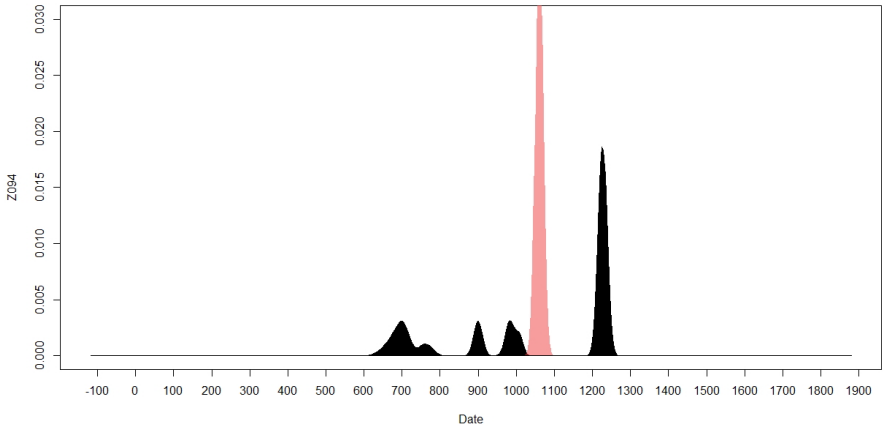


FIG. 5 : Courbes de densité *(dateEv et dateAc)* d'un dépotoir d'un site de Poitiers.

5 Annexe

5.1 Présentation théorique de la démarche statistique pour la datation

5.1.1 1re étape : modélisation du temps événementiel (dateEv)

On dispose pour la ville de Tours (qui définit ici le référentiel), d'une matrice de contingence N à 95 lignes et 200 colonnes résultant du croisement des deux variables qualitatives "ensemble" (ou contexte) et "groupe technique" :

$$\mathbf{N} = \begin{bmatrix} n_{1;1} & \cdots & n_{1;200} \\ \vdots & \ddots & \vdots \\ n_{95;1} & \cdots & n_{95;200} \end{bmatrix} = \begin{bmatrix} \mathbf{N^{(1)}} \\ \mathbf{N^{(2)}} \end{bmatrix} \quad (1)$$

Ainsi, $n_{i;j} (i = 1, \ldots 95$ et $j = 1, \ldots, 200)$ représente le nombre de tessons du groupe technique j présents dans l'ensemble i.

On effectue tout d'abord une Analyse factorielle des Correspondances (AFC) sur la matrice $\mathbf{N}$ de façon à mieux appréhender (par des représentations graphiques) les relations entre variables. Cette première étape permet de construire de nouvelles variables (artificielles) moins nombreuses. L'information contenue dans la matrice $\mathbf{N}$ peut être résumée dans un espace de dimension réduite (ici, un espace de dimension 10). Le profil d'un contexte archéologique peut alors être résumé par ses coordonnées dans les 10 directions obtenues par l'AFC notées $\boldsymbol{F}^k (k = 1, \ldots, 10)$ (au lieu d'avoir ses effectifs dans les 200 groupes techniques retenus). Nous connaissons alors les coordonnées des i points lignes (les ensembles) notées F_i^k qui sont reliées à celles des points colonnes (les GT) notées G_j^k par les formules dites de transition (Saporta, 2011).

Puis, comme nous disposons de monnaies en contexte pour 25 ensembles formant la matrice (notée $\mathbf{N^{(1)}}$) dans l'équation (1) parmi les 95 ensembles actifs retenus, nous

ajustons un modèle de régression linéaire multiple gaussien (*cf.* équation (2)) sur les 25 ensembles de Tours datés en fonction des composantes factorielles significatives de l'AFC (ici les 8 premières uniquement notée $\boldsymbol{F}^k (k = 1, \ldots, 8)$ sur les 10 conservées). Il est alors possible d'estimer la datation d'un contexte notée *dateEv* ; mais aussi de prévoir celle d'un contexte dans lequel aucune monnaie n'a été retrouvée (donc non daté à partir des composantes factorielles significatives de l'AFC). Ce modèle de régression s'écrit sous la forme :

$$dateEv_i = \beta_0 + \sum_{k=1}^{8} \beta_k F_i^k + \varepsilon_i \quad \forall i = 1, \cdots, 25 \tag{2}$$

où les ε_i sont indépendantes et Normalement distribuées suivant une loi $\mathcal{N}\left(0; \sigma^2\right)$ et F_i^k est la coordonnée du contexte i sur la composante k.

Ce qui s'écrit sous forme matricielle :

$$dateEv = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

où $\boldsymbol{\varepsilon} \sim \mathcal{N}_{25}\left(\mathbf{0}; \sigma^2 \mathbf{I}_{25}\right), \mathbf{X} = \left[\begin{array}{cccc} \mathbf{1} & \boldsymbol{F}^1 & \ldots & \boldsymbol{F}^8 \end{array}\right]$ est une matrice 25×9,
$\boldsymbol{\beta} = \left[\begin{array}{cccc} \beta_0 & \beta_1 \ldots & \beta_8 \end{array}\right]' \in \mathbb{R}^9$.

Ainsi, pour tout contexte i_0 dont on connaît les $k = 8$ coordonnées factorielles, $\boldsymbol{x}_{i0}{}' = \left[\begin{array}{cccc} 1 & F_{i0}^1 \ldots & F_{i0}^8 \end{array}\right]$. Les résultats liés au modèle linéaire gaussien nous permettent de prévoir ponctuellement la date $\widehat{dateEv}_{\mathrm{i0}}$:

$$\widehat{dateEv}_{\mathrm{i0}} = \boldsymbol{x}_{i0}{}'\widehat{\boldsymbol{\beta}} \text{ où } \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}} \mathbf{X}'(dateEv)$$

On obtient aussi un intervalle de confiance de prévision de niveau 95% pour $\widehat{dateEv}_{\mathrm{i0}}$:

$$\left[\boldsymbol{x}_{i0}{}'\widehat{\boldsymbol{\beta}} \pm \mathrm{t}_{25-9;97.5\%}\widehat{\sigma}\sqrt{\boldsymbol{x}_{i0}{}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{x}_{i0} + 1}\right]$$

où $\widehat{\sigma}^2 = \frac{\|dateEv - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{25-9}$ et $\mathrm{t}_{25-9;97.5\%} \cong 2.12$ est le fractile 97.5% d'une loi de Student à $25 - 9 = 16$ degrés de liberté[5].

On peut alors approcher la loi de $\widehat{dateEv}_{\mathrm{i0}}$ par une loi Normale $\mathcal{N}\left(\widehat{dateEv}_{\mathrm{i0}}; \widehat{\sigma}^2\left(\boldsymbol{x}_{i0}{}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{x}_{i0} + 1\right)\right)$. Le tracé de cette densité correspond à la **courbe orange** sur les différentes figures de l'article.

5.1.2 2^e étape : du temps événementiel au temps de l'accumulation (dateAc)

En utilisant la formule de transition reliant les coordonnées des points lignes (les ensembles) notées F_i^k à celles des points colonnes (les GT) notées G_j^k, on peut aussi

[5] On rappelle que $\mathrm{u}_{97.5\%} \cong 1.96$, fractile 97.5% d'une loi $\mathcal{N}(0; 1)$.

prévoir la datation des GT notée $\widehat{dateEv}^{j0}$ et obtenir, pour cette date estimée, un intervalle de confiance à 95%.

On définit alors le temps de l'accumulation d'un ensemble, noté *dateAc*, par la somme pondérée des dates des groupes techniques où les poids sont les proportions de NMI de chaque groupe technique présent dans l'ensemble :

$dateAc_i = \sum_{j=1}^{200} \pi_{j/i} \left(dateEv^j \right)$ où $\pi_{j/i} = \frac{n_{ij}}{n_{i+}} = \frac{n_{ij}}{\sum_{j=1}^{200} n_{ij}}$ fréquence conditionnelle.

On peut alors estimer $dateAc_i$, le temps de l'accumulation d'un ensemble i, par :

$$\widehat{dateAc}_i = \sum_{j=1}^{200} \pi_{j/i} \left(\widehat{dateEv}^j \right)$$

En supposant que les variables aléatoires $dateEv^j$ sont indépendantes[6], on peut alors approcher la loi du temps de l'accumulation de tout contexte i_0 par le mélange de gaussiennes suivant :

$$dateAc_{i0} \approx \sum_{j=1}^{200} \pi_{j/i0} \mathcal{N} \left(\widehat{dateEv}^j ; s^2 \left(\widehat{dateEv}^j \right) \right)$$

Le tracé de cette densité correspond à la **courbe noire** sur les différentes figures de l'article. On peut aussi déterminer un intervalle de confiance à 95% pour $\widehat{dateAc}_{i0}$ et estimer ponctuellement $dateAc_{i0}$ en calculant la médiane*.

[6] Par construction des GT, hypothèse réaliste.