Chapitre 3.6
Statistiques au "Passé"
exemples en Paléontologie et Taphonomie Quaternaire

Jean-Philip BRUGAL(1), Antigone UZUNIDIS(2),
Jean-Baptiste FOURVEL(1), Nicolas FREREBEAU(3)

(1) Aix-Marseille Université, CNRS, Minist. Culture,
UMR 7269 LAMPEA, MMSH, Aix-en-Provence,
(2) ArScAn UMR 7041, Paris I Panthéon Sorbonne, Ministère de la Culture/ASM UMR 5140, Montpellier
(3) Archéosciences Bordeaux UMR 6034, CNRS, Université Bordeaux Montaigne, Université de Bordeaux, EPHE-PSL, Pessac

→ Nous invitons le lecteur à lire au préalable le chapitre 9. Classification non supervisée et la partie 4.1. Notions fondamentales : population, individu, variable statistique, modalité.

Les méthodes statistiques sont depuis longtemps utilisées dans de nombreux domaines scientifiques dont les disciplines s'intéressant au Passé. Elles concernent tant la Paléontologie ou la Paléoanthropologie, que l'Archéologie, la Paléoécologie ou la Paléogénétique par exemple. Les deux courts exemples qui suivent – la phylogénie des grands Bovidés (genre *Bison*) en Paléontologie et, la reconnaissance de marques dentaires de grands prédateurs en Taphonomie – illustreront ces applications.

1 Paléontologie : VSI* et Classifications Ascendantes Hiérarchiques

L'un des objectifs des études paléontologiques est de décrire et comparer les caractéristiques morphométriques de formes fossiles. Les outils statistiques permettent de rapprocher ou différencier certains groupes d'individus* en mettant en valeur leur proximité morphométrique apportant des informations à n'importe quelle échelle de la systématique (genre, espèce, sous-espèce). Parmi ces outils, l'utilisation conjointe du *Variability size index* modifié (VSI*) et des classifications ascendantes hiérarchiques[1]

[1] *Cf* 9.3. Classification hiérarchique : classification ascendante hiérarchique (CAH).

(CAH) autorise des analyses sur des échantillons* de faibles effectifs*. Ils peuvent être appliqués à des cas très variés et seront illustrés ici à plusieurs séries (ou 'populations')[2] de bisons du Pléistocène inférieur à l'Holocène, sur la base des métacarpiens[3] III-IV de sujets femelles.

Le VSI* (Escarguel, 2008), dérivé du VSI d'Uerpmann (Uerpmann, 1982), compare les données* quantitatives (mesures* linéaires) de plusieurs séries fossiles par rapport à une population de référence (ici, *Syncerus caffer*, le buffle d'Afrique) à l'aide du programme multi-SIST[4]. Pour une variable* donnée (ici, i), la différence entre les valeurs moyennes* de la série* analysée (Xi) et de la population de référence (Ri) est rapportée à l'écart-type conjoint de ces deux populations et non au seul écart-type de la population de référence, selon la formule :

$$\mathrm{VSI}_i^* = \frac{X_i - R_i}{\sqrt{\frac{\left(\left(n_{\mathrm{Xi}} - 1\right) \times s_{\mathrm{Xi}}^2\right) + \left(\left(n_{\mathrm{Ri}} - 1\right) \times s_{\mathrm{Ri}}^2\right)}{n_{\mathrm{Xi}} + n_{\mathrm{Ri}} - 2}}}$$

Une fois le VSI* calculé pour chaque variable décrivant chaque population, les données sont visualisées dans un plan factoriel, *via* une analyse en composantes principales[5] (ACP). Il regroupe les séries étudiées ayant les mêmes caractéristiques* en fonction de la contribution de chaque variable initiale à chaque composante principale. La corrélation* des variables aux axes et leurs contributions relatives sont nécessaires pour interpréter morphométriquement les axes et la distribution* des populations. De plus, un réseau traversant bootstrap, calculé par le programme SIST est superposé à l'ACP. Il représente les relations de ressemblances* entre populations (ou groupe d'individus) sous la forme d'un réseau non-hiérarchique et statistiquement robuste (Escarguel, 2008).

La CAH classe les populations les unes par rapport aux autres (Drennan, 2010). Dans le cas présent, nous avons choisi de nous servir des coordonnées des populations sur les deux premiers axes de l'ACP afin de ne considérer que les liens les plus significatifs. L'ajout de trop de données dotées d'un poids équivalent aux autres mais d'une force explicative moindre selon l'ACP risquerait de brouiller l'information finale. L'éloignement des individus les uns par rapport aux autres est défini par la distance euclidienne* et celui entre les classes, par la distance de Ward. Le nombre de classes conservées est déterminé *via* le R^2 semi-partiel en accord avec le critère de Ward qui permet de conserver la plus grande inertie interclasse[6]. La position des

[2] Nous entendons par « série » un ensemble de restes ou de spécimens analysés qui concernant un taxon (genre ou espèce animale) est mis en synonymie avec « population ». Dans ce contexte, il s'agit de fossiles correspondant à un ensemble homogène de restes osseux issu d'un unique ensemble sédimentaire. Ce terme doit donc être distingué des populations au sens biologique.

[3] Le métacarpien (membre antérieur), comme le métatarsien (membre postérieur), est l'os compris entre le poignet, ou la cheville, et les phalanges. Il est parfois appelé os canon (en particulier pour les chevaux).

[4] SIST est l'abréviation pour « Size Index Scaling Technique ». Ce programme est disponible sur demande auprès de G. Escarguel, UMR CNRS 5023, Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, Univ. Lyon 1, France.

[5] *Cf* 6.1. Analyses en Composantes Principales (ACP).

[6] En d'autres termes, le choix du critère ou saut de Ward permet d'obtenir les classes les plus dissemblables les unes par rapport aux autres. Ce choix est fait au léger détriment de l'homogénéité à l'intérieur des classes.

groupes, définie par la CAH, les uns par rapport aux autres ne traduit pas de réalité paléontologique à l'inverse de leur composition.

L'ACP sur les données du VSI* concerne 11 gisements/séries pléistocènes et 2 populations modernes (n° 12 et 13) concernant 6 variables, dont un indice* (notées V1 à V6) de métacarpiens femelles (Figure 1A). Le plan factoriel IxII représente 91.27% de la variabilité totale ce qui traduit de façon très satisfaisante les données morphométriques des ossements étudiés. Le premier axe est construit sur des variables relatives à la taille des os (V1 et indice) et le second reflète la morphologie de ceux-ci (V3 et V4) (Figure 1B). Ainsi, les populations se répartissent schématiquement selon un gradient morphométrique (Figure 1A) : en haut à droite, les spécimens graciles et élancées ; en bas à droite, les grands bisons massifs ; en bas à gauche, les petits sujets massifs et en haut à gauche les petits individus graciles. Dans notre exemple, cette répartition correspond également à la succession chronologique des séries examinées.

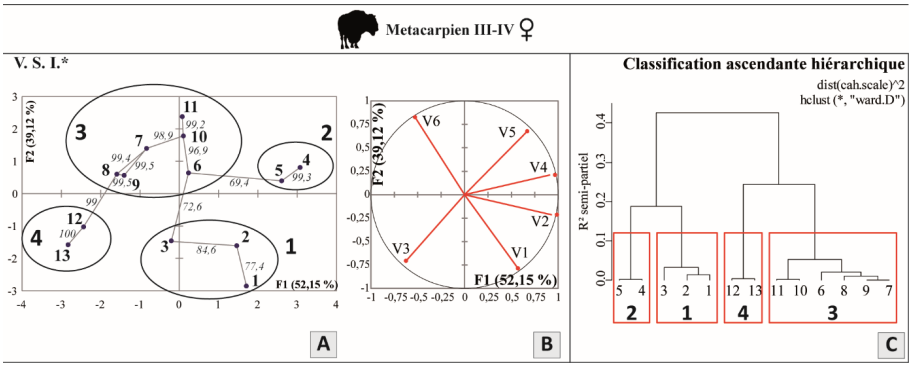


Fig. 1 : (A) ACP sur VSI* de métacarpiens de femelles de Bisons fossiles et modernes ; (B) Cercle des corrélations relatif à l'ACP ; (C) CAH sur les métacarpiens de femelles de Bisons fossiles et modernes à partir des deux premiers axes de l'ACP. 1 : Untermessfeld ; 2 : Vallonnet ; 3 : Durfort ; 4 : Mosbach ; 5 : Châtillon-Saint-Jean ; 6 : Arquet (niveaux inférieurs) ; 7 : Willment's Pit ; 8 : Jaurens ; 9 : Siréjol ; 10 : Gral ; 11 : Arquet (niveaux supérieurs) ; 12 : *B. bonasus* (Empel & Roskosz, 1963) ; 13 : *B. bonasus* (Schertz, 1936). V1 : longueur ; V2 : largeur de l'extrémité proximale ; V3 : épaisseur de l'extrémité proximale ; V4 : largeur de la diaphyse ; V5 : largeur de l'extrémité distale ; V6 : indice : largeur de l'extrémité distale*100/longueur.

Dans un second temps, la CAH regroupe statistiquement les séries en groupes homogènes (Figure 1C) ; quatre regroupements sont visibles. Les caractéristiques de ces groupes sont obtenues grâce à la lecture préalable de l'ACP : 1) les taxons *B. menneri* et *B. shoetensacki* du Pléistocène inférieur et moyen ancien de Durfort, Vallonnet et Untermessfeld désignent des métacarpiens relativement allongés et graciles ; 2) les *B. priscus priscus* du Pléistocène moyen de Chatillon-Saint-Jean et de Mosbach possèdent des os encore assez longs et larges (en particulier l'extrémité proximale) ; 3) les *B. p. mediator* et *B. p. minor* du Pléistocène supérieur de l'Arquet (niveaux infé-

rieur et supérieur), du Gral, Jaurens, Siréjol et Willment's Pit dotés de métacarpiens robustes et courts ; 4) les *B. bonasus* holocènes (actuel bison européen) montrant les métacarpiens les plus petits et très graciles.

Appliquée au métacarpiens III-IV de bisons femelles du Pléistocène et de l'Holocène, l'utilisation conjointe du VSI* et de la CAH permet donc d'une part de caractériser morphométriquement les différentes populations et d'autre part, de les réunir en groupes statistiques. Ce type d'analyse a déjà été appliquée dans diverses études (*e.g.* Grange *et al.*, 2018) et permet une étude statistique sur de petits échantillons avec un minimum d'informations (effectif, moyenne et écart-type).

2 Taphonomie : Statistiques des Marques

Depuis plus de trente ans, l'analyse d'ensembles archéo-paléontologiques s'accompagne d'un examen taphonomique[7] (*e.g.* Behrensmeyer & Hill, 1980 ; Brugal, 2017). Si la taphonomie s'est pendant longtemps développée sur la base d'observations* naturalistes (*e.g.*, Behrensmeyer & Hill, 1980 ; Brain, 1981 ; Sutcliffe, 1970), l'approche statistique des données taphonomiques reste indispensable, à plus forte raison si l'on considère que les ensembles osseux mis au jour ne représentent qu'une vision partielle du dépôt originel. La taphonomie statistique s'est essentiellement développée au cours des années 1990 et trouve son essor à partir des années 2000 (*e.g.* Dominguez-Rodrigo *et al.*, 2015 ; Fourvel & Fosse, 2015). Qu'il s'agisse, entre autres, de méthodes d'analyses multivariées ou de la mise en place d'indices statistiques, cette approche quantitative affine notre vision des ensembles fossiles et tend à limiter les biais interprétatifs dès lors qu'elle est mise en œuvre avec parcimonie.

Le cas de la taphonomie appliquée aux modes de prédation et de consommation par les grands carnivores est un bon exemple de l'apport de l'analyse numérique. Il existe, dans le registre fossile, toute une diversité de prédateur (renard, loup, dhole, hyène, léopard, lion, ours, etc.) ayant la capacité de produire des traces de dents sur les restes osseux qu'ils ont consommés. Depuis près d'une quinzaine d'années, les chercheurs tentent d'identifier les prédateurs ou accumulateurs sur la base de la morphologie, de la taille ou de la localisation des traces de dents (*e.g.* Domınguez-Rodrigo & Piqueras, 2003 ; Parkinson *et al.*, 2014, 2015). Cependant, si la consommation des proies par les carnivores est relativement reconnaissable au moyen d'observations macro- ou microscopiques (marques dentaires, os digérés, fracturations), il reste délicat de préciser (au niveau générique ou spécifique) le ou les agents à l'origine de cette consommation et, par extension, de restituer une chaîne trophique.

Plus récemment, à partir de jeux de données morphométriques toujours plus importants, plusieurs auteurs ont exploré différentes approches statistiques en vue de reconnaître l'origine de ces traces : identifier plus finement les différents agents à l'origine de la prédation/consommation et, par extension, l'implication des différents prédateurs dans la mise en place (ou la remobilisation) des dépôts osseux (voir par

[7] La taphonomie (*taphos-* enfouissement, *nomos-* lois) est l'étude de l'ensemble des processus altérant des assemblages* fossiles, déterminant leur conservation différentielle (il s'agit d'une démarche cherchant à déterminer l'intégrité physique et biologique des séries fossiles).

exemple les traitements 3D et l'usage de la morphométrie géométrique dans Yravedra *et al.*, 2018).

Afin d'apprécier l'apport du traitement statistique dans l'analyse taphonomique, considérons le cas suivant : disposant de deux échantillons de traces de dents, l'un résultant de l'action de prédateurs modernes clairement identifiés (ours, lion, hyène) et l'autre, un jeu de données fossiles pour lesquelles nous émettons l'hypothèse de l'agent (ours, jaguar, loup, hyène) à partir des données contextuelles. Il s'agit ici : (i) de décrire la variabilité de taille des traces (ii) d'envisager une discrimination par espèce et (iii) d'éprouver nos hypothèses sur le registre fossile à partir de comparaison avec les données modernes.

Cette analyse repose sur l'examen d'un échantillon statistiquement représentatif de plus de 2000 observations (*i.e.* traces de dents) dont 1755 pour les prédateurs modernes et 919 pour les faunes fossiles. Chaque individu statistique correspond ici à une trace de dents : il s'agit d'une empreinte dentaire sub-circulaire produite lors de la consommation d'un os (perforation) et pour laquelle nous disposons systématiquement de deux variables, le grand et le petit diamètre (en millimètre), perpendiculaires l'un à l'autre.

On considère ici qu'une telle marque est une ellipse, ce qui permet de calculer différentes grandeurs liées aux deux diamètres. Parmi ces grandeurs, l'excentricité (notée e et définie entre 0 et 1) quantifie l'écart de forme avec un cercle parfait (pour lequel l'excentricité est nulle) :

$$e = \frac{\sqrt{(a^2-b^2)}}{a} \text{ où } a \text{ est le grand rayon et } b \text{ le petit rayon de l'ellipse.}$$

Au sein des taxons modernes étudiés (Figure 2), on constate que les hyénidés produisent une grande quantité de traces circulaires, au contraire des lions (qui en produisent très peu) et des ours (qui n'en produisent pas). Au sein du corpus fossile (Figure 3), le même constat peut être réalisé pour les genres *Crocuta* et *Ursus*, mais les lions sont cette fois associés à des traces circulaires. Enfin, les canidés fossiles produisent eux aussi une majorité de traces circulaires. La distribution de l'excentricité des traces permet ainsi de mettre en évidence des premiers traits distinctifs entre les agents de prédation.

L'excentricité ne renseigne cependant pas sur la dimension de la trace. On peut ainsi s'intéresser à la distance (notée d et exprimée en mm) qui sépare le centre de l'ellipse et un de ses foyers :

$$d = \sqrt{(a^2 - b^2)} \text{ où } a \text{ est le grand rayon et } b \text{ le petit rayon de l'ellipse.}$$

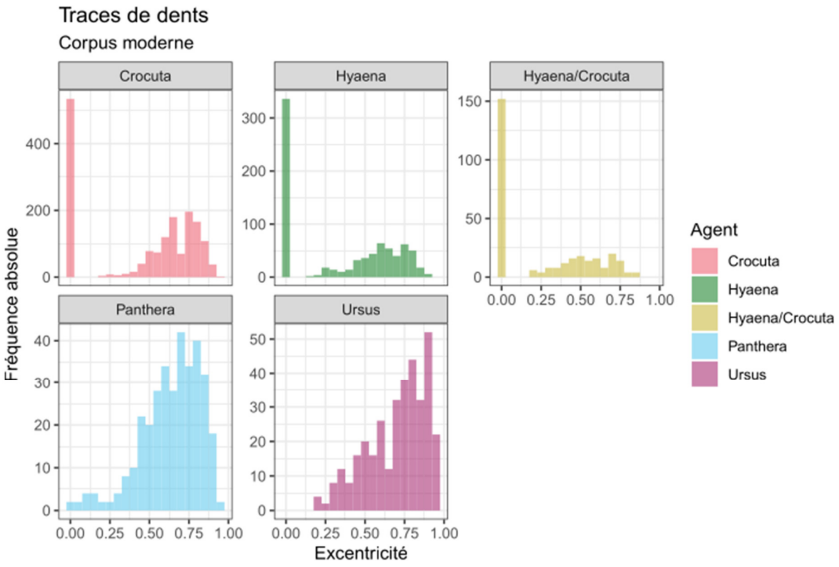


FIG. 2 : Distributions de l'excentricité des traces de dents par agent d'accumulation dans le registre moderne.

On constate ainsi que dans les registres fossiles et modernes, le genre *Crocuta* produit les traces les plus grandes (Figure 4). Au sein du corpus fossile, il est possible de distinguer en moyenne le lion des autres taxons. En l'état, le recouvrement important des valeurs [8]ne permet cependant pas une discrimination fine au cas par cas. L'analyse des tailles de traces de dents n'est pas suffisante pour obtenir l'identification de l'agent (i.e., prédateur) au niveau de l'espèce.

L'étude des données du registre moderne permet également un retour sur celles du registre fossile. Ici, nous avons émis l'hypothèse de l'agent d'accumulation fossile (selon le cas ici, jaguar, loup, hyène ou ours) en fonction du contexte des dépôts osseux. La confrontation des données modernes et fossiles et les différences observées pour un même genre (Figure 3) peut permettre de réinterroger les hypothèses initiales et les déterminations spécifiques.

Cette première analyse doit être complétée, par le croisement avec d'autres informations, notamment qualitatives, relatives au contexte (géo-biologique), à l'âge et à la taille de la proie, au type d'élément osseux, localisation des traces, etc. Il s'agit, à terme, de produire des modèles toujours plus fins caractérisant au mieux les processus de consommation et d'accumulation des restes osseux propres à chaque espèce prédatrice, afin de mieux préciser leur paléo-éco-éthologie.

[8]Ce recouvrement est notamment matérialisé sur la figure 4 grâce au corps des boîtes à moustaches* dont la largeur correspond à l'écart interquartile (la médiane* est figurée par le trait central).

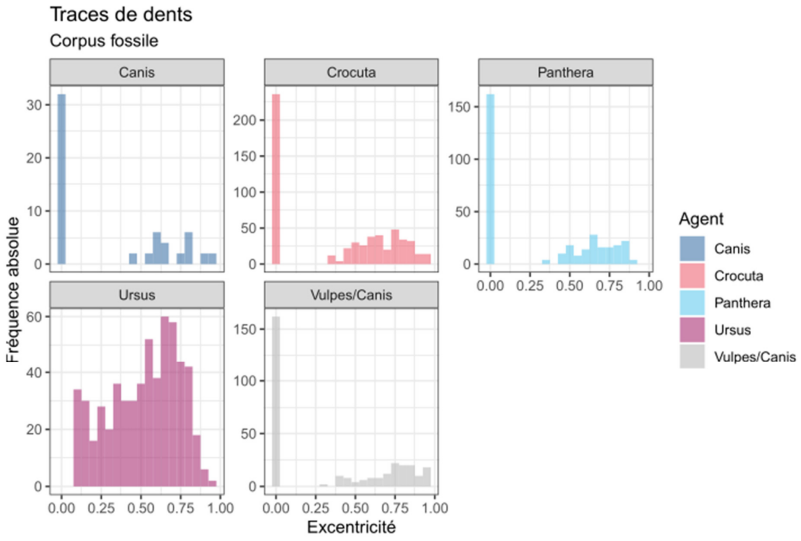


FIG. 3 : Distributions de l'excentricité des traces de dents par agent d'accumulation dans le registre fossile.

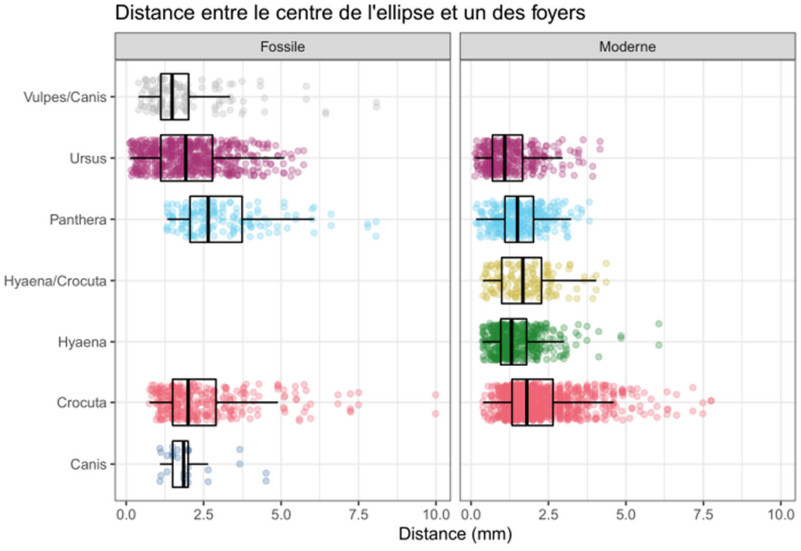


FIG. 4 : Distributions de la distance entre le centre de l'ellipse et un de ses foyers par agent d'accumulation et par registre. Chaque point représente une trace de dent. Les boites à moustaches sont également représentées pour chaque ensemble de valeurs.

3 Remarques *in fine*

Les statistiques sont omniprésentes, voire incontournables, dans les études et recherches sur le 'Passé' s'intéressant à des problématiques très diverses. Les biostatistiques, avec le développement des méthodes en écologie numérique, sont largement utilisées, associées à des démarches à la fois de modèles mathématiques mais aussi d'ingénierie numérique. Le problème des données manquantes ou de leur validité (biais taphonomiques et/ou de documentation) et la mise en œuvre de logiciels, la spatialisation des données et leur 'exploration' (*data mining*), les reconnaissances automatiques de formes par du '*machine-learning*', les calculs matriciels et le développement d'algorithmes spécifiques… (Clavel *et al.*, 2014, 2015 ; Courtenay *et al.*, 2019 ; Crégut-Bonnoure & Fernandez, 2018 ; Fernández & Monchot, 2007 ; Fernandez *et al.*, 2017 ; Galbrun *et al.*, 2018 ; Rodríguez *et al.*, 2004, 2016 ; Travouillon *et al.*, 2011, …, pour n'en citer qu'une 'poignée') constituent autant d'approches analytiques (éco-, bio-, environnnement ou archéo-métriques) qui se combinent et s'appuient sur des bases statistiques multiples dans l'élaboration et validation de modèles (de conceptuels à formels) ou de référentiels*. Ceux-ci sont précieux car ils permettent à terme des reconstitutions, et parfois des prédictions, plus fines et dynamiques de systèmes *s.l.* par définition complexes, et ce dans un cadre et des objectifs d'ordre évolutif, chronologique, biogéographique, paléobiologique (âge, sexe), anthropologique, culturel, ou éco-éthologique pour ce qui concerne la dimension du Passé.

En résumé :

- Dans nos domaines, les fossiles sont souvent peu abondants, voire fragmentaires, et donc choisir des tests statistiques les plus simples et les plus adaptés aux petits échantillons.
- Les bisons ne sont pas uniquement en Amérique du Nord ! : l'analyse des bisons européens fossiles et modernes révèle des processus évolutifs et adaptatifs.
- Ne pas négliger l'importance du dimorphisme sexuel : voir l'exemple des os femelles chez bison.
- Jouer au détective : identifier les traces de dents d'un carnassier sur un os !
- Si le quantitatif est important, ne pas oublier le qualitatif (qui *via* des codages* peut aussi se traiter statistiquement).