PACO, une fonction pour classer les données archéologiques sous R : quel avenir à cet outil ?

Présentation de projet et discussion méthodologique

Type de projet : thèse d'archéologie « Les traditions techniques céramiques de la vallée du Rhin supérieur entre le Xe et le VIIIe siècles avant J.-C. Essai d'un outil automatisé de partitionnement de chaînes opératoires (PACO) », sous la direction de S. Wirth et S. Manem, soutenue en 2018.

Porteur : **Marie Philippe**, céramologue chez ANTEA-Archéologie, membre associé à l'UMR 6298 ARTEHIS.

Thématiques concernées de l'atelier : 1 – méthodes et outils d'enregistrement et de traitement des données de terrain ; 3 – sériations et typologies des éléments matériels.

Contexte du projet

L'outil PACO est une fonction développée sous R dans le cadre d'une thèse de doctorat en technologie céramique :

"*Les traditions techniques céramiques de la vallée du Rhin supérieur entre Xe et VIIIe siècles avant J.-C. Essai d'un outil automatisé de partitionnement de chaînes opératoires (PACO)*", soutenue en 2018 par Marie Philippe.

Depuis cette date, PACO est resté en sommeil, alors que des chercheurs de plusieurs pays ont déjà témoigné leurs besoins et attentes vis-à-vis de cet outil. Le but de cette présentation est de relancer le projet de mise à la disposition de la communauté scientifique, en discutant en particulier des développements possibles et des moyens mobilisables.

Problématiques d'anthropologie des techniques, à l'origine de l'outil

Une problématique essentielle au travail doctoral était d'identifier des traditions techniques à partir des traces de fabrication observables sur les céramiques archéologiques. En anthropologie, les traditions techniques sont des manières de faire transmises entre individus, de génération en génération. **La récurrence des pratiques et leur perduration dans le temps** sont donc des caractéristiques essentielles dans cette définition. Les manières de faire sont décrites grâce au concept de la **chaîne opératoire, qui s'entend ici comme une combinaison d'opérations techniques**, la succession de ces opérations lors du processus de fabrication étant connue a priori.

Le corpus de céramiques sur et pour lequel PACO a été développé est constitué de 829 vases sélectionnés sur 19 sites de la vallée du Rhin supérieur, essentiellement des habitats occupés entre le Bronze final IIIa et le Hallstatt C (Xe-VIIIe s. av. J.-C.).

Un jeu de données à trous

L'échantillonnage a ciblé les vases les moins fragmentés et portant des macrotraces de fabrication lisibles. Au total, **34 variables techniques** ont été renseignées pour chaque vase, en plus de variables contextuelles (site, datation, forme du vase, etc.). Les modalités prennent essentiellement la forme de **valeurs qualitatives** (noms de techniques, de matériaux, etc.). Un **indice de fiabilité binaire** est renseigné pour chaque objet : il traduit la certitude avec laquelle les opérations techniques ont été identifiées à partir des tessons.

Le tableur des données est dont très grand, mais c'est aussi un véritable gruyère : il affiche **63% de données manquantes** ! En effet la théorie de la technologie céramique est calquée sur des travaux anthropologiques menés sur des populations actuelles, où les données sont directement observables, ce qui n'est pas le cas en archéologie.

Les manques d'information concernent toutes les variables, en proportions inégales. Certaines variables présentent en outre une modalité nettement majoritaire, faisant transparaître l'existence de normes techniques.

Dès lors, comment comparer des séquences à (nombreux) trous ? Le calcul d'un coefficient de similarité général est à proscrire (résultats aberrants), de même qu'une implémentation de données (trop de trous). La nature des variables (qualitative) limite de plus le choix d'une méthode. **Un constat s'impose : il n'existe pas d'outil permettant de traiter ce type de jeu de données ; il reste à créer.**

L'outil PACO

Une manière de contourner le problème est de hiérarchiser l'ordre de comparaison des variables constituant chaque séquence. **La méthode de classement des tessons préconisée par Valentine Roux (2016) est donc choisie et interprétée pour être traduite en une fonction informatisée : PACO (PArtionnement de Chaînes Opératoires).**

C'est une **fonction itérative** écrite sous R, qui subdivise étape par étape un jeu de données en groupes d'objets à partir de variables techniques. PACO réalise ainsi virtuellement le tri de tessons que ferait un céramologue sur sa table de travail. **A chaque étape de subdivision, PACO sélectionne la variable avec la modalité la plus fréquente parmi les déterminations fiables**.

Le tri est réalisé à partir d'un tableau présentant en lignes les objets et en colonnes les variables techniques. L'indice de fiabilité binaire est présenté en première variable.

Le résultat du tri est représenté par une **arborescence**, qui comprend plusieurs niveaux de nœuds correspondant à autant de variables utilisées dans la subdivision. **Les objets qui présentent une donnée manquante sur la variable sélectionnée pour la subdivision sont définitivement écartés du processus de tri**. Il en résulte des branches de différente longueur, le long desquelles l'utilisateur peut lire les caractéristiques des chaînes opératoires du corpus.

Le code de la fonction est basé sur deux boucles imbriquées :

- la "boucle des parents", qui recherche les nœuds à subdiviser (les parents) et la variable de subdivision en comptant, parmi les effectifs, chaque occurrence de chaque modalité de chaque variable technique non encore utilisée dans le tri. Si toutes les variables ont déjà été traitées, cette boucle prend fin.
- la "boucle des enfants", qui court à l'intérieur de la boucle des parents, et qui recherche les objets constituant les nœud-fils. Un parent n'aura pas d'enfant si ses membres sont strictement identiques ou s'ils présentent tous une lacune sur la variable de subdivision.

Les avantages de PACO

- Cet outil permet de trier des données qualitatives et lacunaires.
- Il peut théoriquement être utilisé pour trier n'importe quel jeu de données archéologique.
- La méthode est explicite et reproductible rapidement.
- Le tri est représenté sous forme d'arborescence.
- Sa règle de subdivision donne la priorité aux comportements récurrents.

Perspectives d'amélioration

PACO semble pouvoir répondre à un besoin de la communauté scientifique. Toutefois, en l'état actuel, plusieurs limites sont constatées :

- Nécessité de savoir utiliser R, ce qui n'est pas le cas de nombreux archéologues. PACO pourrait être **traduit dans un autre langage**, pour en démocratiser l'accès.

- Script de débutant, dont la **mise en forme pourrait être améliorée** pour être plus lisible.

- Appel à des packages plus forcément à jour, ce qui nécessiterait une **réactualisation.**

- Besoin de grandes capacités de calcul, car le code repose sur deux boucles imbriquées. La procédure pourrait être repensée pour **améliorer la vitesse des calculs.**

- Une seule règle de subdivision et un seul paramétrage. D'autres alternatives pourraient être développées. Parmi les autres règles de subdivision possibles figure celle **priorisant les variables avec le moins de données manquantes**. Julie Gravier et François Fouriaux insistent sur l'intérêt de proposer à l'utilisateur **d'imposer lui-même l'ordonnancement des variables**. Une autre idée basée sur leurs suggestions consiste à proposer **plusieurs paramétrages de gestion de la fiabilité** : soit aucune prise en compte, soit un indice de fiabilité unique par ligne, soit un indice propre à chaque variable. Pour Bruno Desachy, cette dernière solution permettrait plus de souplesse par rapport à la mise à l'écart "brutale" de certains objets entiers sur la base d'une seule information manquante.
  Ces différentes alternatives peuvent mener à des structures d'arbres différentes. Pour **évaluer la robustesse des processus de tri,** il semble intéressant à Bruno Desachy et Julie Gravier de pouvoir comparer les différentes arborescences obtenues pour vérifier la stabilité des résultats (ordonnancement des variables et classement des objets). En développant des indicateurs appropriés, on pourrait ainsi tester plusieurs configurations et ne conserver que les plus stables.

- Testé sur un seul jeu de données. Il serait idéal de pouvoir l'**essayer sur d'autres données**, éventuellement en lien avec d'autres problématiques de recherche. Anaïs Pinhède évoque les ateliers archéomatiques organisés par le Réseau ISA. Ces derniers permettent de réunir des archéologues autour d'un outil informatique pour l'appréhender et le tester.

Julie Gravier insiste sur l'intérêt de la visualisation du processus de tri. Elle suggère aussi de **différencier les feuilles par leur couleur, selon la règle qui régit l'arrêt de la subdivision** (plus de données dispo *vs* plus de variabilité).

La version repensée de PACO pourrait être développée de manière collaborative, par exemple en mettant le script en ligne sur GitHub comme le propose Julie Gravier. Lionel Tabourier suggère de faire appel à des étudiants en informatique pour réfléchir sur ce projet. Marie Philippe envisage un accueil CNRS pour disposer de temps dédié à ce travail.

Diffusion de l'outil

La diffusion de la version améliorée de PACO pourrait prendre deux formes, selon le profil d'utilisateur ciblé par ce travail. Julie Gravier envisage soit une application "clé en main" depuis un tableur, qui nécessiterait d'utiliser un autre langage que R ou un hébergement sur le web (du type application R Shiny), soit un package sur CRAN, qui serait accessible à des utilisateurs de R même débutants.

Plusieurs possibilités de publication sont évoquées :

- la revue *Journal of Open Source Software* qui accepte des articles sur le développement de package, si ce dernier est préalablement diffusé sur CRAN (qui implique une évaluation du code lui-même),
- la revue *Archéologies numériques*, suggérée par Bruno Desachy,
- *Archeologia e Calcolatori*, suggérée par Bruno Desachy,
- *Archaeological Method and Theory*, pour la méthode plus que pour le code.

Références bibliographiques

Roux 2016 : Des céramiques et des hommes. Décoder les assemblages archéologiques. Nanterre, Presses universitaires de Paris Nanterre, 415 p.