De Filemaker à R, ou comment l'Allemagne m'a donné des ailes

Boîte à outils, discussion méthodologique

- thèse : Les *Parisii*, un « petit » peuple entre Gaule Celtique et Gaule Belgique. Dynamiques territoriales et identité culturelle (IIIe s. av. J.-C. - Ier s. ap. J.-C.)

- thématiques de l'atelier concernées : 1 - méthodes et outils d'enregistrement et de traitement des données de terrain ; 4 - exploration multidimensionnelle de sources croisées pour l'archéologie ;

résumé de présentation

Quatrelivre Carole

La séance proposée ici constitue la suite de la présentation de ma base de données (la célébrissime Carobase) en janvier 2019. La structuration et l'acquisition des données désormais terminées, il s'agit maintenant de passer à leur exploitation. Pour ce faire, je suis partie étudier auprès d'Oliver Nakoinz, priv.-doc. à l'Université de Kiel, en Allemagne. Malheureusement, les mesures de confinement tombent deux semaines après mon arrivée, et je suis contrainte de rentrer en France. Ainsi, seules les toutes premières étapes du traitement statistique ont été abordées.

Pour compenser cet allègement de programme, je présenterai un tutoriel sur les *Cluster Analyses*, créé dans le cadre d'une *Winter School* internationale à Kiel du 16 au 19 mars 2020, et qui est en cours de perfectionnement.

Enjeux du corpus francilien

D'abord, une présentation du corpus à son état définitif permettra de faire le point sur l'hétérogénéité des données et d'évoquer les solutions adoptées pour son exploitation statistique.

Présentation du protocole

L'ensemble du protocole sera élaboré sur le logiciel de statistique R, avec pour objectif de :

1. Créer des faciès de consommation à partir des assemblages mobiliers
2. Hiérarchiser les sites franciliens à partir de l'ensemble des données (mobilier et immobilier)
3. Mesurer les interactions

En réalité, seuls les deux premiers aspects ont été discutés, et les premiers essais ont été menés uniquement sur les données de mobilier, hors céramique. J'avais en effet sous-estimé le temps qu'il me faudrait pour extraire et préparer le tableau à importer dans R.

La normalisation des données

Les effectifs de types d'objets extraits, la normalisation des valeurs a été abordée avec Oliver Nakoinz, afin d'éviter un effet d'écrasement des valeurs trop faibles par les valeurs fortes.

Elle s'effectue en divisant les valeurs par la valeur la plus haute, afin d'utiliser la distance euclidienne. Ainsi, les valeurs sont ramenées entre 0 et 1. Les variables deviennent comparables entre elles, surtout lorsqu'il y a des écarts numériques importants. L'exemple le plus courant reprend des données archéométriques (longueur, largeur), mais Oliver Nakoinz a tout même proposé d'appliquer la normalisation à un tableau de comptages.

La fonction utilisée sur R est la suivante :

```
normalize <- function(x) {return ((x - min(x)) / (max(x) – min(x)))}
```

L'utilisation de la fonction **apply** du package **base** permet d'appliquer la formule à l'ensemble des colonnes du dataframe. En revanche, le résultat est une matrice numérique ; il y a donc un changement de la nature de l'objet.

Mesures de distances et cartes thermiques

Les distances entre les valeurs normalisées sont ensuite mesurées, pour créer une matrice de similarité, indispensable aux étapes de partitionnement des données.

La fonction utilisée est **dist** du package **stats**.

```
dist_mat <- dist(LTA_BNORM1)
```

Avant d'aller plus avant, il faut traiter le problème des cellules vides, « NA ». Elles sont remplacées par la valeur 1, qui est la valeur d'indépendance totale. Cela permet d'éviter par la suite des regroupements artificiels d'individus.

```
dist_mat[is.na(dist_mat)] <- 1
```

Avant de réaliser la carte thermique, il faut également supprimer la ligne des totaux, qui n'est pas signifiante dans le cadre des mesures de distance.

```
dist_mat <- dist_mat[1:47,1:47]
```

Enfin, il est possible de réaliser la carte thermique à partir de la matrice de similarité. Il s'agit d'une visualisation synthétique de la similarité des individus entre eux. Une diagonale doit être clairement perceptible ; elle indique que chaque individu est identique à lui-même.

```
heatmap(dist_mat)
```

Des cartes thermiques ont été réalisées à partir des données de mobilier pour La Tène ancienne (fig. 1), pour La Tène moyenne (fig. 2) et pour La Tène finale (fig. 3). Les résultats pour la dernière période sont les plus probants parce qu'elle comptabilise le plus d'objets et parce que les types de mobilier identifiés sont plus variés.

Il reste encore à analyser en détail les regroupements d'individus (c'est-à-dire les sites archéologiques), manifestes par des blocs de couleur différente sur la carte thermique, et d'identifier les variables discriminantes.

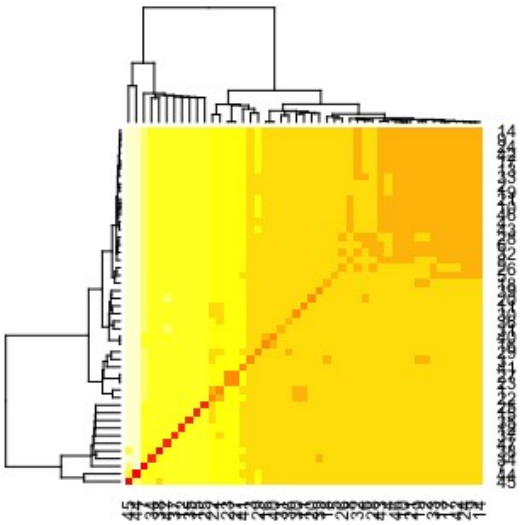

figure 1 : Similarité des sites de La Tène ancienne d'après le mobilier non céramique

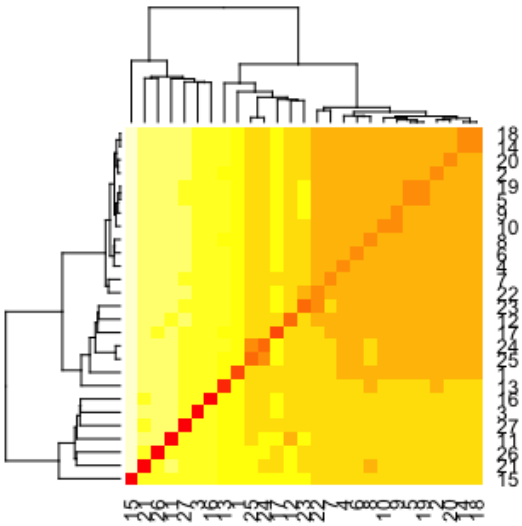

figure 2 : Similarité des sites de La Tène moyenne d'après le mobilier non céramique

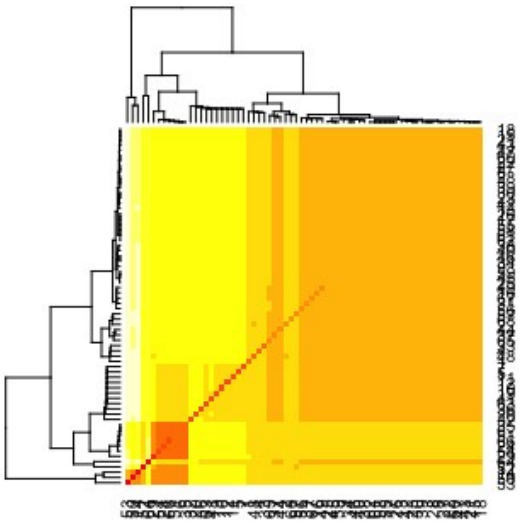

figure 3 : Similarité des sites de La Tène finale d'après le mobilier non céramique

Perspectives concernant le partitionnement des données

À partir de la matrice de similarité, il est possible de se lancer dans divers types d'analyses multivariées. L'analyse factorielle des correspondances est envisagée puisqu'elle est adaptée aux variables qualitatives. Oliver Nakoinz a également proposé une analyse en composantes principales, dont il faudrait discuter les modalités.

Tutoriel sur les *Cluster Analyses*

Intitulé « Introduction to Classification, Distance matrices and hDBscan Tutorial », ce tutoriel a été réalisé sur Rmarkdown avec Sophie Schmidt (Université de Cologne), Robert Staniuk (Université de Kiel), Sarah Martini (Université de Yale) et moi-même.

La création de ce tutoriel avait pour double objectif de nous former sur la classification des données archéologiques sur R, tout en produisant un document utile à un maximum de personnes. C'est pourquoi il sera traduit en français, en allemand et en polonais.

La réalisation des matrices de similarité constitue la première étape vers la classification. Ont été abordées plusieurs méthodes de calcul de similarité sur des jeux de données binaires et des jeux de données qualitatives : indice de Rand (*simple matching*), indice de Jaccard, indice d'Ochiai et la distance de Hamming. (Un autre groupe s'occupe des données quantitatives.)

Enfin, la classification des données est expliquée en pratique. L'algorithme utilisé dans le cadre de ce tutoriel, *Hierarchical Density-Based Clustering of Applications with Noise* (hDBscan), permet de traiter le problème du « bruit », c'est-à-dire des individus qui tombent entre les groupes. Cet algorithme tente également de déterminer le nombre idéal de classes pour le jeu de données – habituellement, le nombre de groupes est déterminé *a priori* par l'utilisateur. La démonstration s'appuie sur des jeux de données spatiales, binaires et qualitatives.

Une dernière partie encore en cours d'élaboration sera consacrée au coefficient de silhouette. Il s'agirait de la méthode la mieux adaptée pour déterminer le nombre optimal de classes pour chaque jeu de données.