Chapitre 2.2.2
Formalisation des données

Laure NUNINGER(1), Xavier RODIER(2), Hélène MATHIAN(3)

(1) Chrono-Environnement UMR 6249, MSHE C.N. Ledoux UAR 3124, CNRS, Université Marie et Louis Pasteur, Besançon, (2) CITERES-LAT UMR 7324, MSH Val de Loire UAR 3501, CNRS, Université de Tours, Université d'Orléans, Tours, (3) Environnement Ville Société - EVS UMR 5600, CNRS, Université de Lyon, Lyon

Le recueil des données en archéologie suit des protocoles spécifiques souvent adaptés aux conditions d'investigations sur le terrain (prospections, sondages, fouilles) ou de télédétection dans des jeux de données numériques (altimétrie, imagerie). Certaines techniques sont invasives et de fait, destructrices, comme la fouille, ce qui contraint l'archéologue à enregistrer le maximum de données possibles pour qu'elles puissent être utilisées par le plus grand nombre de personnes en vue de répondre à des objectifs parfois extrêmement variés. Quand les techniques sont dites non-invasives, elles impliquent néanmoins un temps de travail conséquent et sont parfois difficilement reproductibles car les objets observés eux-mêmes sont soumis à des conditions de dégradation potentielle liées au climat ou aux activités humaines.

La collecte des données archéologiques (fouille ou prospection) est effectuée avec un spectre large, pour minimiser la perte d'information liée à l'incapacité de définir *a priori* la totalité des données mises au jour, mais avec une grille de lecture qui n'est pas neutre car elle dépend de l'état des connaissances, de la question historique qui justifie l'opération, de la nature des traces observables laissées par les sociétés du passé, de leur conditions de conservation, des conditions d'investigation. Dans le cas de la fouille en particulier, mais c'est parfois le cas pour d'autres modes d'investigation, l'expérience archéologique n'est pas renouvelable, c'est pourquoi les données enregistrées sont qualifiées de données « primaires ». Ces données « primaires » peuvent donc être multiformes et elles participent des inventaires archéologiques qui sont le préalable nécessaire à toute étude, quelle que soit l'échelle considérée. Les programmes de recherches archéologiques sont fondés sur des inventaires détaillés généralement issus, dans un premier temps, de recensements patrimoniaux tels que la carte ar-

chéologique nationale, puis largement augmentés par des enquêtes appropriées après avoir déterminé les descripteurs* nécessaires pour aborder la question posée. Ainsi, les données archéologiques enregistrées dans des inventaires et que l'on dit « primaires » correspondent ici à ce que l'on qualifie d'observables* (*cf.* Chapitre introductif).

1 Le passage de l'entité archéologique dites « primaire » ou observable à l'unité statistique

La définition de l'unité statistique* pertinente pour l'analyse dépend essentiellement de la question posée, mais également de la nature des données disponibles. Comprendre le peuplement d'une région par exemple peut impliquer le recollement de données issues d'investigations variées et plus ou moins anciennes réalisées selon des conditions et des protocoles variables. En outre du point de vue de leur représentation géométrique, ces données peuvent être associées soit à des points, soit à des surfaces et parfois même à des lignes. Cette hétérogénéité acceptable en termes d'inventaire, y compris cartographique, est néanmoins problématique dès lors que l'on analyse les données d'un point de vue statistique et spatial. En effet, analyser les données avec des "outils" statistiques, nécessite d'avoir un corpus ou un ensemble d'entités définies de la même manière sous peine que l'éclairage ne fasse ressortir que des différences de définitions et non des différences liées à des associations de caractéristiques différentes. Sans être exhaustifs, nous aborderons quelques-uns de ces problèmes *via* des exemples ciblés.

Pour les études sur le peuplement*, le cas du « site archéologique » est emblématique. Le **site** est une zone localisée dans l'espace géographique où des artefacts et vestiges archéologiques ont été repérés, en d'autres termes il s'agit d'une localisation géographique associée à une activité humaine. Sa délimitation géographique et chronologique est généralement liée à l'expertise de l'archéologue qui l'a découvert, a collecté le matériel (des tessons, par exemple) et a enregistré l'information (localisation, attributs…). Ces sites peuvent être issus de différents types d'investigation (fouilles, sondages ou prospection au sol). À l'échelle régionale, l'analyse du peuplement repose généralement sur une grande majorité de sites détectés en surface tandis que les fouilles et les sondages constituent une part réduite (sauf exception dans le cas de grands décapages).

Pour faciliter leur analyse, ces sites sont souvent réduits à un point dans l'espace qui correspond généralement au centroïde de la zone d'emprise des vestiges et artefacts. En prospection de surface, l'emprise du site est définie par l'archéologue sur le terrain selon des critères partagés par une équipe et parfois de façon plus ou moins implicite. Dans le cas rare, car extrêmement coûteux en temps de collecte, d'un enregistrement dit « avec pointage des indices au réel » (Kotarba 1998), chaque artefact identifié au sol est précisément localisé et l'entité site est reconstruite *a posteriori* par l'agrégation des artefacts selon leur nature et leur dispersion spatiale. Quelle que soit la méthode d'enregistrement, la réduction du site à un point localisé dans l'espace n'est pas sans incidence sur le plan statistique dès lors que l'on s'intéresse aux questions de peuplement. Tout d'abord, les sites peuvent être de nature extrêmement variée, il peut s'agir d'un habitat isolé, de l'emprise d'une agglomération comme d'une structure utilitaire

(enclos pastoral, puits, cabanes à outils…) ou encore d'une nécropole. Si la question consiste à analyser la distribution spatiale* de l'habitat, un choix doit être opéré parmi les points enregistrés pour ne retenir comme unité statistique, par exemple, que ceux qui correspondent à des zones habitées par des personnes vivantes. Dans ce cas, les nécropoles ne sont pas forcément ignorées mais peuvent être associées aux unités statistiques sous la forme d'attributs en fonction de la distance qui les sépare de l'habitat.

Un autre problème fréquemment rencontré est celui de la caractéristique diachronique du site, en d'autres termes un même lieu peut être occupé à différentes périodes parfois successives mais souvent bien distinctes par exemple au Néolithique puis au Moyen Âge sans occupation intermédiaire. Toutefois quand il s'agit de comprendre la dynamique de peuplement d'une région, si on considère le site comme l'individu statistique, son caractère diachronique posera problème dès que l'on calculera la densité spatiale* du peuplement, par exemple. Ainsi, toutes choses égales par ailleurs, les zones où l'on s'établi à plusieurs reprises en un même lieu, soit un même site, auront une densité plus faible que les zones où l'on s'établi en différents lieux, donc dans plusieurs sites. En effet, sur le plan statistique, si un même lieu est réoccupé à plusieurs périodes, la densité spatiale toutes périodes confondues peut apparaître sous-évaluée car trois occupations, par exemple, seront résumées dans un seul site. La densité spatiale doit donc être appréciée en fonction d'une certaine contemporanéité ou synchronie[1] des occupations. Pour pallier ce problème du site en tant qu'individu statistique, on pourra créer une nouvelle unité appelée **« établissement »** afin de produire trois unités statistiques dont les caractéristiques sont à la fois spatiales et temporelles (Nuninger *et al.* 2012a). Ainsi pour une même paire de coordonnées (x,y) on pourra avoir une ou plusieurs unités statistiques.

Enfin, quand l'inventaire comporte des données anciennes, les localisations, parfois imprécises voire inexistantes, sont souvent rapportées au centroïde du chef-lieu communal, voire au centroïde du territoire communal actuel. Ainsi, on peut avoir plusieurs sites artificiellement localisés sur un même point. C'est un cas particulièrement fréquent pour les dépôts d'objets par exemple.

Ces quelques exemples permettent de mettre l'accent sur deux points importants :

1- Préalablement à toute analyse statistique ou spatiale, il convient de bien évaluer comment la zone d'étude a été explorée et selon quelles conditions de collecte car cela peut avoir une incidence sur le nombre d'entités considéré et leur répartition spatiale. Sur ce point on pourra se reporter à la notion de carte de confiance (Saligny *et al.* 2008 ; Nuninger et Bertoncello 2011, encadré 4.3 ; Nuninger *et al.* 2012b).

2- En fonction du corpus archéologique et du questionnement, un prétraitement est nécessaire pour fabriquer les unités statistiques qui vont être analysées à partir des données d'enregistrement dites « primaires ». Ces unités doivent être définies par des choix explicites pour homogénéiser les données analysées et éviter des biais statistiques

[1] contemporanéité et synchronie étant intimement liée aux contraintes de datation de l'ensemble chrono-culturel étudié.

liés en réalité à des conditions d'observations hétérogènes ou des modes d'enregistrement associés à des concepts qui ne sont pas neutres, comme le site par exemple.

Avec ces remarques initiales, nous invitons le lecteur à distinguer les unités d'enregistrement primaires ou observables, des unités statistiques qui doivent être construites à propos, c'est-à-dire en fonction de la question posée. Toutefois, il est tout à fait possible de considérer les unités d'enregistrement primaires comme des unités statistiques dès lors qu'elles forment un ensemble cohérent et que l'on cherche à analyser de façon exploratoire la structure du jeu de données par exemple (*cf.* 2.2.3).

2 Quelles données pour quelles analyses statistiques ?

Lorsque l'on considère un ensemble de données suffisamment structuré pour mener à bien des analyses statistiques voire spatiales, plusieurs questions se posent généralement : faut-il intégrer l'ensemble des données ou une partie seulement à l'analyse ? Dans le second cas, comment sélectionner les données ? Faut-il conserver les unités statistiques élémentaires comme telles ou faut-il les agréger dans de nouvelles unités ? Pour toutes ces questions, il n'existe pas de solution clef en main mais quelques guides possibles de réflexion pour y répondre au cas par cas.

2.1 Savoir décomposer son problème

Dans la plupart des cas, même quand la question paraît simple, par exemple « comment se diffusent les grandes haches en jade alpin à partir des principaux gîtes d'extraction ? », le moyen d'y répondre peut être compliqué (*cf.* 2.2.7.). Autrement dit, on ne peut pas forcément répondre avec une seule analyse statistique mais l'on doit en intégrer plusieurs dans un même raisonnement. Chaque analyse permettra de répondre à un élément de la question, ce qui suppose dans un premier temps de décomposer la problématique en autant de sous-questions que nécessaire. Par exemple, à quelle distance du site d'extraction avons-nous retrouvé des haches néolithiques en jade ? Est-ce que le nombre de haches diminue proportionnellement avec la distance ? Est-ce qu'il existe des centres de production et de redistribution de ces haches ? Est-ce qu'il existe une relation entre l'état de polissage de la hache et son parcours de diffusion ?…

Cette première démarche est fondamentale pour mettre en évidence les objets statistiques auxquels on s'intéresse et qu'il faudra précisément définir pour réaliser des analyses. Dans notre exemple, on voit déjà apparaître au moins quatre, voire cinq, objets : la hache en jade, le lieu de découverte de la hache (son point d'arrivée), le lieu d'extraction de jade, le lieu de production et/ou de redistribution des haches. Ces objets vont composer différents tableaux de données dont il faudra préciser les relations, d'appartenance par exemple une hache à un lieu de découverte. En fonction de la sous-question posée, on pourra également être amené à regrouper des objets dans des catégories qui pourront alors devenir des objets statistiques, par exemple des centres de redistribution (nœuds dans un réseau).

2.2 Qu'est-ce que je souhaite mettre en évidence avec mon analyse ?

À chaque niveau de question correspond des objets ou unités statistiques particuliers, mais la question demeure de savoir s'il est forcément nécessaire de mobiliser l'ensemble

de ces unités dans chacune des analyses. De nouveau, la réponse se formule ici par une question : qu'est-ce que l'analyse doit m'apporter comme connaissance que je n'ai pas déjà ?

Prenons, l'exemple des établissements de l'âge du fer dans une région donnée. Il s'agit de savoir si chacun de ces établissements a la même fonction et par conséquent la même capacité à structurer un territoire ou si, au contraire, on a les moyens d'observer des différences entre eux qui permettraient de les ordonner par ordre d'importance hiérarchique sur le plan fonctionnel. Dans ces conditions, on se propose d'utiliser une analyse des données multivariée, suivi d'une classification (*cf.* 2.2.8., 6. et 9.). Cette méthode permet de mettre en évidence une hiérarchie sur la base de la combinaison des critères les décrivant (analyse des données multivariées) puis de les mettre dans des classes ordonnées (classification). D'un point de vue archéologique, il peut paraître logique d'intégrer tous les établissements de l'étude comme autant d'unités statistiques. Toutefois, si on a deux agglomérations dont la description les distingue nettement des autres établissements, le risque est d'obtenir un résultat qui confirme ce que l'on sait déjà tandis que l'on n'apprendra rien sur les différences moins évidentes qui peuvent exister entre les autres établissements. Ainsi, si on se place d'un point de vue statistique, il est intéressant d'éliminer ces deux individus atypiques de l'analyse afin de mettre en évidence des indices de différenciation plus subtils entre des établissements plus difficiles à caractériser et à ordonner. Les agglomérations ne sont pas pour autant évacuées de l'étude puisque leur catégorie peut être réintégrée pour l'interprétation ou dans le cadre d'une autre analyse, par exemple celle qui consiste à caractériser l'évolution de l'organisation spatiale des établissements (*cf.* 2.2.4.).

Avec cet exemple, on voit que la sélection de tout ou partie des données pour l'analyse statistique est fonction du résultat que l'on souhaite obtenir. Il n'y a donc pas de règle *a priori* mais la nécessité d'expliciter son raisonnement en s'appuyant d'une part sur une bonne connaissance des principes de l'outil statistique utilisé, d'autre part sur une bonne connaissance des données mobilisées dans un cadre problématique bien défini. Avec cette dernière remarque, il nous paraît utile de rappeler qu'une bonne analyse exploratoire de ces données est un préalable indispensable à toute étude. L'analyse exploratoire des données peut être réalisée à l'aide d'outils relativement simples d'analyse statistique univariée ou bivariée (*cf.* 4. et 5.). Ce type d'analyse permettra de guider et de justifier les choix réalisés en termes de sélection de données et/ou de variables* dans le cadre d'analyses plus complexes à variables multiples par exemple.

2.3 Savoir ajuster sa focale d'observation

Dès lors que l'on s'intéresse à des dynamiques dans le temps et l'espace, il faut considérer une difficulté supplémentaire pour circonscrire le champ d'une analyse et définir

les objets ou les unités statistiques de référence : la focale d'observation qui se définit notamment par l'échelle[2] et la **granularité**[3] spatiale et/ou temporelle.

L'échelle de construction de l'information historique par l'archéologie est fortement contrainte par celle de l'acquisition des données. L'emboîtement des échelles d'analyse se traduit par la définition d'objets spécifiques à chacune d'entre elles et l'utilisation de descripteurs adaptés à chaque dimension (sémantique, spatiale et temporelle). Quelle que soit la question de recherche posée, son énoncé fixe la focale d'observation qui est déterminante pour la granularité de l'information recueillie, de ses descripteurs et la définition de l'unité statistique. L'échelle d'analyse est déterminée par la résolution du phénomène étudié. Dans l'étude des réseaux d'habitats et des terroirs, par exemple, la macro-échelle fait souvent apparaître des continuités (avec de légers déplacements ou des restructurations dans la longue durée), alors que la micro-échelle révèle des ruptures. Le phénomène de nucléation de l'habitat (*incastellamento*, encellulement) peut ainsi être lu comme une mutation brutale ou comme le résultat d'une série complexe de transformations successives et aux tendances parfois contradictoires.

En archéologie, le changement d'échelle est une opération cognitive qui ne peut être résumée ni à un effet de loupe ou de zoom, ni à une généralisation de l'information. En effet, le passage de la caractérisation des objets d'étude à l'analyse des dynamiques spatiales, par exemple, ne consiste pas uniquement en un changement de niveau géographique ou spatial. Spatialement, la transformation de points en surfaces, et inversement, fait appel à des processus de recomposition plus complexes que la généralisation ou l'agrégation d'entités identifiées dans le sol à l'échelle de la fouille ou de la prospection. L'emprise d'une *domus*, par exemple, n'est pas déterminée par l'agrégation spatiales des structures mises au jour par la fouille qui ont permis de l'identifier. L'étude d'un espace consiste en un changement d'échelle qui nécessite de transformer les objets individualisés à l'échelle des sources (matérielles, écrites, iconographiques) en unités statistiques pour les analyses et en objets historiques à l'échelle du territoire étudié. Pour les périodes historiques, l'interprétation passe, selon l'échelle, de la description (couche d'occupation) à l'identification d'un usage (sol intérieur), puis à la détermination d'une fonction (habitat domestique, construction civile). En revanche, la chaîne d'inférence* (*cf.* 8) n'est pas du même ordre pour une collection d'objets pour lesquels la relation contexte-datation n'est pas toujours établie, ou encore lorsqu'en Préhistoire, on passe de l'analyse technique des artefacts à une identité culturelle puis à une aire culturelle.

Le mode de raisonnement, qui conduit à l'interprétation des données archéologiques, consiste en une double approche, inductive d'une part, hypothético-déductive de l'autre. La première est fondée sur une chaîne d'inférences à partir des connaissances acquises principalement par la fouille et dont l'analyse conduit à l'interprétation his-

[2]Le concept d'échelle auquel on fait référence correspond au rapport de taille entre deux réalités, soit chronologique (de la seconde au millénaire par exemple) soit géographique (du local au global par exemple). Par extension on parle de niveau d'échelle quand le passage d'une réalité à l'autre n'est pas continu (zoom) mais s'effectue par sauts par exemple du niveau communal au niveau régional.

[3]La notion de granularité définit la taille du plus petit élément (indivisible) d'un ensemble de données (soit la taille du grain dans un sac de sable par exemple). Par extension, on pourra estimer le niveau de granularité qui sera élevé si on considère par exemple un élève dans une école ou bas si on considère une classe dans une école.

torique. Dans ce cas, la mise en œuvre des données archéologiques de terrain relève d'un processus inductif de regroupements successifs (fonctionnels, spatiofonctionnels, temporels et spatiotemporels) des atomes d'informations que sont les unités stratigraphiques (ou couches archéologiques). La seconde consiste à inscrire les schémas d'interprétation du site dans l'histoire spécifique de l'espace auquel il appartient (une ville, un territoire…) selon sa position supposée dans une aire chrono-culturelle. Il s'agit cette fois d'un processus hypothético-déductif qui consiste à établir une périodisation hiérarchisée et spécifique au site fouillé, fondée sur les hypothèses d'inscription des phases d'occupation du site dans un modèle historique. Les deux démarches interfèrent et se rejoignent dans leur apport à la compréhension des phénomènes auxquels elles doivent chacune permettre d'aboutir. À l'échelle de la fouille, ce processus permet, d'une part d'assurer la robustesse des données de l'analyse, d'autre part d'échapper à l'illusion que, du large ensemble des données les plus précises possibles à l'entrée, sortira l'étroite réponse à la question posée. Ce dernier cas relève de la confusion entre système d'enregistrement purement descriptif, prétendument objectif et exhaustif, et système d'information élaboré pour répondre à une problématique. En effet, de nombreux systèmes d'enregistrement sont conçus comme des banques de données regroupant l'information la plus détaillée possible au détriment de sa pertinence. Cette dérive est concomitante de l'informatisation des données mais traduit plus généralement la réduction du raisonnement à une démarche uniquement inductive.

L'articulation des approches inductive, inhérente à la fabrication des corpus de données en archéologie lors des fouilles, et hypothético-déductive, nécessaire pour hisser le discours archéologique au-delà de la stricte description des vestiges, intervient à toutes les échelles d'analyse. En effet, dès lors qu'il s'agit de raisonner à partir d'un corpus de données archéologiques, quelle que soit sa granularité (unité stratigraphique, objet, site, agglomération…), le raisonnement archéologique est fondé sur cette co-construction inductive et déductive.

Une fois encore, il n'y a pas de solution clef en main pour sélectionner les données à la bonne échelle et la bonne granularité, mais ces deux notions doivent nécessairement intervenir et être explicité dans la formalisation des données soumis à l'analyse d'une part pour utiliser les outils statistiques adéquat (variable continue* ou discrète* par exemple), d'autre part pour accompagner la lecture des résultats et leur interprétation.

Prenons le problème de la contemporanéité et de la synchronie des établissements : ce sont deux aspects différents dans l'analyse d'une dynamique du peuplement, soit des changements que l'on peut observer dans le temps. La notion de contemporanéité implique que les établissements aient réellement fonctionné au même moment, tandis que la synchronie repose sur la construction d'un idéal-type s'appuyant sur la relative stabilité d'un système d'unités fonctionnelles. Par exemple, si on admet un découpage par siècle, on sait que tous les établissements, surtout ceux qui ont laissé des traces fugaces, ont peu de chance d'avoir été contemporains. Ils ont très probablement eu une durée de vie de l'ordre d'une génération. Dans ce cas de figure, dans une même tranche d'un siècle on peut avoir deux établissements qui, en réalité, n'ont pas fonctionné en même temps. Pourtant, on peut admettre leur synchronie du

point de vue de l'analyse, si on estime que la disparition de certains et l'apparition d'autres se compensent et participent du même système de fonctionnement sans rupture dans l'équilibre de ce système. Aussi, même si on dispose d'une information de datation plus détaillée pour certains de ces établissements, il ne sera pas forcément utile de les mobiliser dans l'analyse. En revanche, c'est une connaissance qui devra être mobilisée au moment de l'interprétation pour affiner ou nuancer les résultats obtenus. La notion de période chronologique doit également être questionnée dans le cadre d'une analyse statistique. En effet, si on reprend le concept de synchronie, il est alors possible de mobiliser une connaissance *a priori* pour définir des périodes pour lesquelles on présuppose une certaine stabilité du peuplement. Cette posture, tout à fait justifiable, appelle néanmoins deux remarques en termes d'analyse statistique. La première est un avertissement sur les biais quantitatifs induits par un découpage du temps à géométrie variable. Malgré son évidence, c'est une erreur suffisamment courante pour qu'on s'y attarde le temps d'un exemple. Considérons, la Basse Vallée du Rhône (Figure 1), si l'on compare l'effectif des établissements occupés au 2ème Âge du fer (5 siècles environ) et ceux occupés au Haut Empire (2 siècles environ), on aura une distorsion dans les résultats tout simplement liée à la durée effective des périodes. Une solution consiste à normaliser[4] les effectifs pour obtenir des proportions équivalentes . Toutefois, cette solution nous amène à considérer une seconde remarque relative à la question que l'on se pose, à savoir l'analyse des changements dans la dynamique de peuplement. En effet, dans notre exemple, ce que pourra mettre en évidence l'analyse ce sont uniquement les changements de fréquence entre deux périodes et d'une certaine manière, on aura introduit en entrée dans l'analyse, ce que l'on cherche à voir en sortie (dans les résultats). Dans ces conditions, il est important d'interroger la granularité des données dont on dispose et d'opter par exemple pour un découpage régulier du temps qui va permettre d'observer des variations plus fines dans la distribution des fréquences que l'on ne percevait pas dans le découpage par période chrono-culturelle et historique (Figure 1). Dans cet exemple, on voit que les choix opérés sur la variable chronologique peuvent être déterminant pour les résultats, mais également qu'ils impliquent un changement d'individu statistique. En effet, ce n'est plus le nombre d'établissements qui sera comptabilisé dans l'effectif mais le nombre d'occupations dans une période donnée. Ainsi par exemple, à un établissement correspondra plusieurs occupations, autant qu'il y a de périodes pouvant décrire sa durée de vie.

Souvent au cœur des débats archéologiques, il reste à souligner que cette question du temps ou du découpage chronologique n'est pas forcément justifiée et que l'on peut s'intéresser à la distribution de sites ou d'artefacts sans poser spécifiquement la question du changement, mais simplement pour comparer des situations régionales ou explorer les données mobilisées dans une étude (*cf.* 2.2.4).

[4] le processus de **normalisation** en statistique est une méthode de mise à l'échelle et de translation des valeurs de telle façon qu'elles se situent dans une plage entre 0 et 1. Il existe plusieurs techniques de normalisation dont celles centrées sur la moyenne que l'on appelle aussi standardisation (*cf.* 4. et 2.1.5)

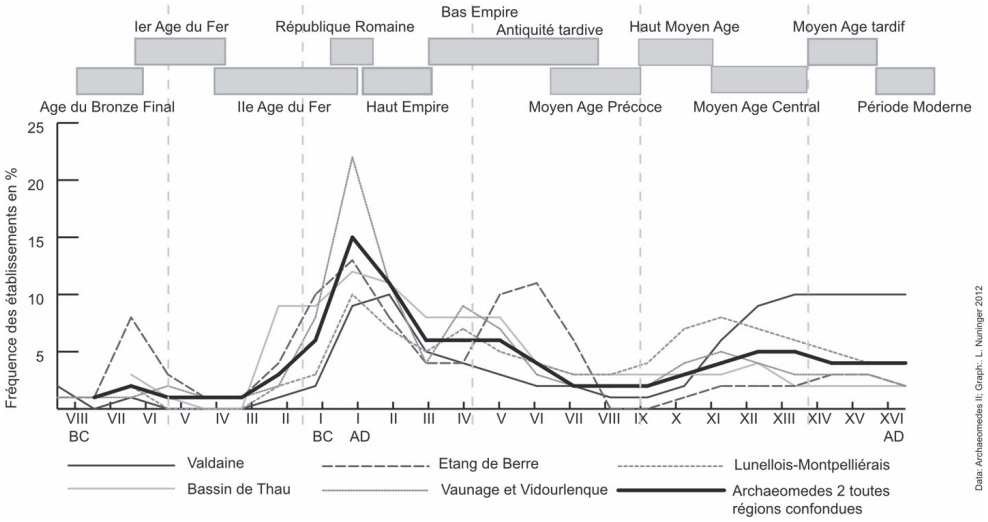


FIG. 1 : Courbes d'occupation par siècles et par micro- régions étudiées dans le programme Archaeomedes II

3 Comment définir ou sélectionner les variables ?

Les unités archéologiques, par exemple un établissement, considérées dans le cadre de l'étude peuvent être décrites par un ensemble d'attributs ou de critères. Dès lors que l'on considère l'établissement en tant qu'unité statistique d'une population (ensemble d'établissements), un attribut peut devenir un caractère ou variable, c'est-à-dire un « critère particulier » qui est propre à cette unité statistique mais que l'on peut rencontrer chez d'autres. Par exemple, si l'on considère l'établissement comme une unité statistique, sa superficie peut être considérée comme une variable. Il peut s'agir d'une variable continue si l'on s'intéresse à l'ensemble des valeurs prises par les mesures de superficie ou d'une variable discrète si l'on considère des catégories de superficie petites, moyennes, grandes (*cf.* 4.). Sélectionner cette variable de superficie pour l'analyse statistique implique qu'une majorité d'unités possède ce critère. En effet, si sur un ensemble de 100 habitants, seuls 10 d'entre eux possèdent une mesure de superficie, dans ce cas le critère ne peut pas devenir une variable car il n'aurait aucune pertinence pour l'analyse statistique même s'il en a une sur le plan thématique. Ensuite le choix du type de variable continue ou discrète va dépendre d'une part de la question posée, d'autre part de la donnée dont on dispose. Par exemple, s'il s'agit de catégoriser les établissements en quelques types que l'on souhaite ordonner du moins important au plus important, il n'est peut-être pas utile d'étudier finement la distribution des valeurs quantitatives des mesures de superficie (surtout si les biais de mesure sont importants) mais plus intéressant de considérer des seuils ou des catégories qualitatives. Cette transformation d'une variable continue en variable discrète permet notamment de prendre en compte l'inexactitude observée dans le processus de mesure. Comme on le verra à travers les exemples présentés dans les chapitres suivants, l'utilisation de variables discrètes est prépondérante car souvent mieux adaptées pour

qualifier les données étudiées et tenter de répondre à des questions qui ne nécessitent pas forcément une grande exactitude des valeurs prises par la variable (Nuninger *et al.* 2012a). Il est important de noter que quand on parle de classes dans une variable discrète, il peut s'agir de catégoriser des valeurs de description qui n'ont pas forcément de lien entre elles, par exemple la fonction productive d'un établissement : agricole, artisanal, commerciale…Mais il peut également être question de classer au sens d'ordonner, c'est-à-dire de donner un ordre, un sens à la variable (variable ordinale*), par exemple en ordonnant le type de matériaux de construction caractérisant l'établissement : pierre, pierre et tuile, plus mortier et plâtres, plus éléments de décoration (mosaïque, enduits peints…). On voit dans ce dernier exemple qu'il y a une idée de gradation pour caractériser un établissement selon sa construction de la plus simple à la plus évoluée. Adopter l'un ou l'autre classement des modalités d'une variable sera fonction de la question posée et de ce que l'on cherche à mettre en évidence avec l'analyse (*cf.* 2.2.8.).

En archéologie, on doit souvent faire face à l'incomplétude des données notamment du point de vue des attributs spatiaux et/ou temporels. Ainsi, par exemple, il est fréquent d'avoir une unité statistique (individu) bien identifiée et bien décrite par un ensemble de variables mais dont la datation ou la localisation fait défaut. Par exemple, pour les établissements dont on ne connaît pas la datation ou pas de façon suffisamment précise, on pourra les retirer de l'analyse de la distribution du nombre d'occupation par siècle afin de ne pas introduire un biais de lissage artificiel de la courbe des occupations par siècle. En revanche, si leur localisation est connue, on pourra tout à fait les faire intervenir dans certaines analyses spatiales. Dans tous les cas, on pourra mobiliser ces individus "établissements" dans la phase d'interprétation avec une représentation cartographique spécifique (en blanc ou gris par exemple) afin de nuancer les résultats par des « vides » spatio-temporel bien identifiés et sans doute liés à un problème de sources.

Dans le cas d'une analyse multivariée ou seules quelques variables ne sont pas renseignées, il est tout à fait possible d'intégrer les établissements au profil incomplet dans l'analyse statistique, dès lors que leur nombre ne pèse pas trop dans l'ensemble des données étudiées et que leur apport est jugé indispensable sur le plan thématique. Il s'agit ici de faire jouer des individus et des variables avec un juste équilibre entre les conditions requises sur le plan quantitatif, pour limiter les biais, et les contraintes thématiques nécessaires pour que l'analyse garde du sens eu égard à la question posée.

4 Pour résumer

En préalable de toute analyse statistique et avant même que se pose le choix des méthodes et des outils à utiliser, il convient de bien cerner sa problématique. Dans un second temps, il s'agit de mettre en adéquation des données en fonction des sources disponibles et de bien évaluer leur capacité à répondre en fonction de leur complétude, exactitude, précision, résolution spatiale et temporelle. Cet apprentissage des données et leur formalisation progressive pour la réalisation d'analyses statistiques solides, va de pair avec un aller-retour permanent entre la problématique et les sous-questions qu'elle implique et les résultats obtenus par analyses successives. En d'autres termes,

l'analyse statistique n'est pas une fin en soi, le processus n'est pas linéaire mais s'inscrit le plus souvent dans une démarche exploratoire qui doit nous aider à raisonner. Pour reprendre l'exemple des premières expériences menées sur la classification des établissements, il faut insister sur le fait qu'il s'agit d'un outil de réflexion et que « toute classification demeure provisoire et toute interprétation une interprétation d'attente » (Favory *et al.* 1994, p. 126).

En résumé

Avant de commencer l'analyse statistique :

- Efforcez-vous de formaliser les contours de votre problématique et les sous-questions qu'elle implique. Par exemple, pourquoi (dans une région donnée) l'habitat se regroupe-t-il dans des agglomérations à partir du second âge du fer ? Cette problématique suppose que l'on s'interroge, notamment, sur la distribution de l'habitat, à partir de quand il est dispersé ? À partir de quand on le considère regroupé ? Qu'est-ce qui fait qu'un habitat est une agglomération par rapport aux autres ? …
- Homogénéiser et estimer la qualité des jeux de données inventoriés dans le corpus étudié, selon le type d'investigation, son extension spatiale, son caractère systématique ou non…
- Pensez à adapter la transformation des variables en fonction des échelles d'analyse et du niveau de granularité des objets statistiques.

Pour éviter les erreurs (courantes) d'interprétation :

- Pensez à décrire pour chaque question précise, les unités statistiques, les variables, les choix opérés – échelles, niveau de granularité, relations entres les variables – thématique et statistique – … (dès le départ et au fur et à mesure de vos analyses). Ce travail peut prendre la forme de tableaux ou de plusieurs schémas qui pourront s'intégrer dans un schéma global précisant le type de données en entrée, le type de données en sortie, et les étapes de l'étude réalisées à l'aide de plusieurs analyses statistiques.
- Dissocier la phase de lecture des résultats statistiques, de la phase d'interprétation qui suppose l'implication de connaissances externes à l'analyse proprement dites. Par exemple, si les agglomérations ont été éliminées de l'analyse statistique visant à classer hiérarchiquement des établissements, elles ne seront pas incluses dans la présentation des résultats de l'analyse, mais réintégrées comme connaissances, dans un second temps, au moment de l'interprétation.