Dialogues entre concepts archéologiques, objets empiriques et méthodes statistiques

Lena SANDERS
Géographie-cités UMR 8504, CNRS, Université Paris 1

1 Introduction

Ce livre fait dialoguer les sciences archéologiques avec les sciences statistiques au travers des trois termes mis en avant dans le sous-titre, les questions, les données, les méthodes. Le point de départ est donné par les questions que se posent les archéologues et la première partie de l'ouvrage, consacrée à la présentation d'études de cas divers, fournit un large panorama de thématiques. Pour chacun de ces cas, les archéologues ont rassemblé des données de nature diverse reposant sur des observations, des dénombrements et des mesures (morphométriques, minéralogiques, chimiques, radiométriques etc.). L'enjeu consiste à traduire les questions archéologiques en questions statistiques pour lesquelles on dispose d'une méthode pour « faire parler » les données et ouvrir des pistes de réflexion pour avancer sur les questions posées. Les questions archéologiques peuvent être très générales avec des interrogations sur l'organisation spatiale du peuplement, les formes de circulations tant locales que de grande portée (des productions, des techniques, des populations), les modes de rapport aux ressources, les pratiques alimentaires, les logiques d'implantation de l'habitat, l'évolution démographique etc. Ces grandes questions portant sur les sociétés du passé nécessitent en général d'être décomposées en plusieurs étapes et font appel à des données et des méthodes statistiques variées. À l'autre bout de l'échiquier, les chercheurs sont également amenés à répondre à des questions précises et pragmatiques : – dater et/ou identifier la provenance d'un objet ; – estimer une propriété (l'âge au décès d'un individu, la quantité de protéines dans son alimentation) ; – identifier des zones géographiques qui offrent un potentiel élevé de contenir des vestiges archéologiques. Même de telles questions très précises nécessitent souvent d'être décomposées en étapes et de combiner différentes sources de données et de méthodes statistiques

comme le montrent plusieurs des cas évoqués dans cet ouvrage. Les cas d'étude présentés en première partie de l'ouvrage sont répertoriés au travers de trois entrées génériques, « quand », « où », et « quoi », suivant la triade de Peuquet (Peuquet 1994), alors que la deuxième partie du livre propose une entrée par les méthodes statistiques, des plus simples aux plus avancées.

Les approches statistiques prennent appui sur une formalisation mathématique et ont été largement appliquées dans le domaine des sciences humaines et sociales à partir des années 1960, chaque discipline se distinguant par la nature de ses objets et la manière de les décrire. Afin de discuter des objets propres à être traités par des méthodes statistiques en archéologie, je prendrai appui sur une démarche ontologique. Suivant Pierre Livet (Livet *et al.* 2010), « l'ontologie consiste à analyser un domaine, en identifiant les entités pertinentes (objets, propriétés, relations, événements et processus), et les opérations qui peuvent être opérées sur ces entités ». Cette définition met l'accent sur deux aspects centraux dans cet ouvrage : – d'une part l'explicitation de l'ensemble des entités intervenant dans un questionnement archéologique ; – d'autre part l'identification des « opérations » qui sont ici les diverses méthodes statistiques développées pour répondre à la problématique posée. Face à une question, il s'agit ainsi d'identifier et de réfléchir au sens des entités pertinentes (2), puis de passer en revue les grandes familles de méthodes statistiques pouvant être appliquées à ces entités (3) et enfin, de montrer à partir des études de cas, la diversité des choix statistiques à la disposition de l'archéologue pour répondre à ses questions (4).

2 Point de vue ontologique : des objets aux individus statistiques et des propriétés aux variables

L'information archéologique provient de sources multiples et repose sur des objets de niveaux différents, du plus élémentaire (tessons de céramique, fragments d'os, scories) au plus élaboré (établissements, agglomérations, pôles). Pour mener à bien une réflexion ontologique sur les « entités » et les « opérations » en jeu, il est utile de différencier de manière explicite trois domaines (Livet et Sanders 2014a) (Figure 1) : 1- le domaine conceptuel relatif au questionnement de l'archéologue qui porte sur des faits et phénomènes du passé, les relations à l'environnement, l'organisation de l'habitat, les pratiques alimentaires, le fonctionnement économique, les mobilités, des sociétés passées ; 2- le domaine empirique qui est relatif à ce qui est *observable*, qu'il s'agisse des objets eux-mêmes (vestiges, traces de nature diverse) ou de leurs propriétés (issues de dénombrements et de mesures) ; 3- le domaine de la modélisation statistique qui implique de formaliser les données sous forme d'individus statistiques et de variables que l'on peut soumettre à l'analyse statistique afin d'établir des ordres de grandeur et de mettre en évidence des relations entre différents phénomènes.

Ces trois domaines sont en interaction lors d'une recherche. Pour l'archéologue il s'agit ainsi de définir les concepts liés à son questionnement, tant les entités en jeu que les hypothèses sur les relations entre elles, et d'identifier les éléments empiriques correspondants, à partir de ce qui est observable. La pratique de la recherche montre que la proximité, en termes de sens, entre le concept étudié et ce qui est observable peut être plus ou moins grande. Il s'agit donc, dans un premier temps, d'évaluer dans

quelle mesure des objets et des propriétés observables sont bien représentatives, dans le domaine empirique, des concepts d'intérêt (2.1) et ensuite, à partir de ce qui est observable, de construire des individus statistiques propres à être analysés par des méthodes statistiques (2.2). Je discute ici séparément de ces deux étapes, afin de mettre en avant leurs aspects propres, mais dans la pratique elles sont indissociables.

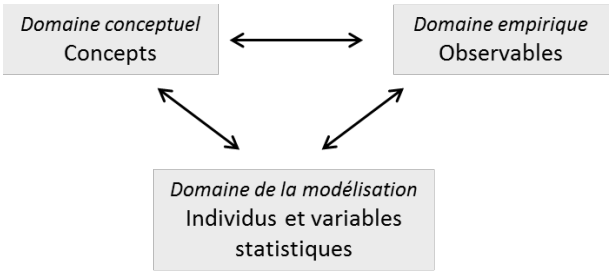


FIG. 1 : Un dialogue entre trois domaines pour répondre au questionnement de l'archéologue (d'après Livet *et al.*, 2014b)

2.1 De la question au concept, du concept à l'observable

Face à un questionnement, le premier objectif est de définir le (ou les) concept(s) d'intérêt, puis d'identifier les **observables**[1] qui vont permettre d'approcher ces concepts dans le domaine empirique, et sur lesquels on va pouvoir opérer des mesures et des traitements statistiques. Lorsque la question de recherche concerne l'époque actuelle, la relation entre le concept et l'observable, même si elle nécessite une explicitation, peut être relativement directe. Tel est par exemple le cas lorsque Berroir *et al.* (2017) cherchent à identifier l'existence de sous-systèmes de villes en France, c'est-à-dire des ensembles de villes fortement interdépendantes dans leur fonctionnement et évolution. Les concepts d'intérêt sont alors d'une part les villes, entités élémentaires des systèmes de villes, et d'autre part les interactions spatiales qui réfèrent aux relations entre les villes qui « font système ». Dans le domaine des observables, les auteurs ont choisi de mobiliser les « aires urbaines » pour représenter les villes. Dans la définition INSEE celles-ci sont des agrégats de communes polarisées par un centre. Il s'agit d'un *choix* et dans un autre contexte, s'ils avaient eu accès à l'information nécessaire, ils auraient pu préférer une définition reposant par exemple sur la continuité du bâti plutôt que sur un découpage administratif. Pour ce qui est des interactions spatiales entre villes, plusieurs observables sont disponibles pour l'époque actuelle. Les migrations résidentielles ou les flux de marchandises par exemple sont souvent utilisés pour en rendre compte. Ils en représentent effectivement chacun un aspect, mais un aspect seulement, différent de ce que reflèterait par exemple l'importance des transports à grande vitesse ou des partenariats scientifiques entre entreprises. Chacun de ces observables apporte une information particulière sur la nature et l'intensité des liens unissant entre elles les villes. Afin d'obtenir une vue plus globale, Berroir *et al.* (2017) ont choisi de combiner sept observables différents pour mesurer les relations existantes entre les villes dans le domaine empirique. Cet exemple illustre que même dans un cas

[1] Le terme d'observable sera utilisé de façon générique pour désigner tout ce qui peut être observer, qu'il s'agisse des objets eux-mêmes ou de leurs propriétés.

où des données existent et où le lien entre l'observable et le concept d'intérêt semble assez direct, il y a un décalage qu'il est essentiel d'assumer. Plus le questionnement porte sur un phénomène éloigné dans le temps, plus le lien entre les deux a de chances d'être indirect. L'explicitation d'un tel décalage a une véritable dimension heuristique. Elle amène à réfléchir sur la capacité d'un observable donné à représenter le concept d'intérêt, et à construire le chemin, fondé sur un travail d'interprétation, qui permet de les relier l'un à l'autre, en précisant avec quels apports et quelles limites.

Les cas d'étude présentés dans les premiers chapitres de cet ouvrage offrent un panorama très large de la variété des relations entre les concepts d'intérêt correspondant au questionnement archéologique et les observables qui sont mobilisés pour y répondre. Plusieurs des interrogations portent sur des mouvements, que ce soit de personnes, de biens ou de marchandises, qui témoignent de l'existence d'échanges ou de mobilités entre différents lieux géographiques dans des périodes du passé. Aucune donnée directe n'est dans ce cas disponible et les observables mobilisés sont souvent centrés sur des propriétés portées par des objets divers, artefacts ou restes d'humains. Dans le cas 2.1.3 l'observable est ainsi l'obsidienne, c'est-à-dire un matériau, sans qu'il y ait nécessairement un intérêt propre pour l'artefact associé, qu'il s'agisse d'un objet entier ou d'un simple fragment. Ce sont les propriétés du matériau, que ce soit la forme des stigmates de sa taille, ou sa composition géochimique qui vont permettre d'établir une signature, une « empreinte digitale », qui va fournir une information sur le lieu de provenance des artefacts analysés. On pourra en déduire qu'un déplacement s'est produit entre ce lieu et le lieu de découverte, et obtenir ainsi un indice sur la circulation des objets et/ou des populations passées.

Dans le cas 2.2.7, l'objectif est également de mettre en évidence des mobilités, à travers cette fois la circulation d'objets de prestige. La hache de jade est choisie comme représentante d'un tel objet sur le plan conceptuel. Dans le domaine empirique, les observables sont composés de lames entières ou de bouts de lames trouvés dans divers sites archéologiques. Leurs différentes propriétés, comme le niveau de finition du polissage par exemple, permet aux chercheurs de différencier l'objet de prestige de l'objet commun, celui d'un simple outil. L'enjeu pour les chercheurs est de suivre la « biographie » de ces objets entre lieu de production, lieux de transit et de transformation et lieu de découverte, souvent dans des tombes. D'un point de vue ontologique, cette « biographie », en tant que trajectoire dans l'espace-temps, représente un processus, c'est-à-dire une entité qui a une étendue dans le temps et qui est caractérisée par une succession de phases temporelles. Par ailleurs, deux autres types d'objets, correspondant à un niveau plus agrégé, sont mobilisés dans cette étude de cas : il s'agit des « pôles », lieux de forte concentration de haches, et des « tronçons », porteurs de la circulation des haches d'un pôle à un autre. Sur un plan conceptuel, ces objets sont définis à partir de la théorie des graphes en termes de nœuds et d'arêtes. Sur le plan empirique, les observables sont des quantités de haches trouvées en différents lieux (définissant les nœuds-pôles) ou ayant transité entre eux (définissant les arêtes-tronçons). L'ensemble de ces objets et de leurs propriétés vont permettre aux auteurs de construire des trajectoires et d'identifier un modèle de diffusion des haches qui opère par l'intermédiaires de pôles.

Le cas 1.3 illustre une autre forme de relation indirecte entre le concept d'intérêt et les observables. L'enjeu consiste à mettre en évidence un faciès de céramique et à le dater. Les observables mobilisés sont des tessons de céramiques dont il sera possible, grâce à l'analyse statistique, de caractériser l'assemblage. Pour le dater, il sera en revanche nécessaire de mobiliser des pièces de monnaie trouvées à proximité de ces tessons et porteuses de leur date d'émission. Les observables sont ainsi de deux types, les premiers (les tessons) permettant de construire un assemblage céramique dont la composition va permettre de dégager un profil (une forme de signature) qui va caractériser le lieu de découverte (atelier, site), alors que les seconds (les pièces de monnaie) apportent une information supplémentaire sur les premiers, de par la date de leur émission, qui constitue une propriété que l'on va pouvoir associer à l'assemblage.

Dans certaines études les observables concernent une partie seulement de l'objet d'intérêt et leur analyse va permettre de tirer des conclusions sur l'objet pris dans sa globalité. Tel est par exemple le cas présenté dans le chapitre 3.6 portant sur les bisons du pléistocène. Les observables sont des métacarpes (os des pattes antérieures) à partir desquels on a pu, en se fondant sur l'analyse de leurs caractéristiques morphométriques, mettre en évidence une opposition entre les petits et graciles d'un côté et les grands et massifs de l'autre. L'analyse faite sur les métacarpes reflète ainsi les propriétés des bisons femelles de cette population auxquelles on s'intéresse.

Dans d'autres cas il y a un véritable décalage entre les concepts d'intérêt et les objets observables qui vont servir à construire les individus soumis à l'analyse statistique. Le même cas 3.6 en présente un exemple à propos d'un questionnement portant sur les principaux prédateurs à une période du passé. Le concept d'intérêt désigne ces prédateurs mais les observables sont les traces de dents laissées sur des restes osseux. Les formes de ces traces, appréhendées par des mesures morphométriques, vont permettre d'identifier les prédateurs concernés.

Dans plusieurs cas d'étude enfin, le questionnement porte sur les conditions de vie et de décès des populations à certaines périodes du passé (*cf.* 3.2, 3.4 et 3.6). Les observables sont des squelettes humains que l'on peut d'une part caractériser à partir de données anthropologiques et d'autre part en procédant à des analyses physico-chimiques élaborées des tissus osseux et dentaires. La composition isotopique permet par exemple d'avoir une idée de l'alimentation des individus (*cf.* 3.3 et 3.4).

Dans ces études de cas, l'analyse va ainsi porter sur des objets matériels relativement simples (bouts d'obsidienne, haches, tessons, pièces de monnaie, os, dents), qui ont parfois un lien *a priori* assez indirect avec le concept d'intérêt, et c'est l'interprétation des résultats d'analyse qui va permettre de construire ce lien. Les hypothèses sous-jacentes à une telle interprétation sont au cœur du chapitre 3.2 qui illustre parfaitement le nombre d'étapes qui peut séparer l'observable et l'estimation que l'on en tire sur le fait auquel on s'intéresse. L'auteure explicite ainsi les hypothèses qu'il est nécessaire d'accepter pour remonter de l'observation d'un échantillon anthropologique de 87 individus à une estimation de la population vivante, en passant par la construction successive de la population exhumée, puis inhumée, puis décédée, chaque étape impliquant des conditions précises pour que les résultats soient fiables.

L'exemple de l'établissement illustre à l'inverse un cas où l'analyse va porter sur un objet matériel lui-même complexe et résultant d'une construction (*cf.* 2.2.2, 2.2.4 et 2.2.8). Le concept d'établissement désigne une entité cohérente en termes de fonctionnement (qu'il s'agisse d'un habitat ou d'une annexe agricole), c'est-à-dire formant un tout dans un contexte spatial et temporel donné. En tant que tel il représente un objet élémentaire dans le système de peuplement. Du côté des observables, l'archéologue dispose de vestiges et traces variés, sous la forme de fragments de matériaux ou de mobiliers, de traces au sol, de vestiges de murs etc. recueillis au niveau d'un site archéologique (*cf.* 2.2.2) et il s'agit, à partir de l'observation et de l'analyse de l'ensemble de ces éléments, de construire un objet « établissement » pertinent, qui soit, de plus, compatible en termes de sens, avec les établissements construits en d'autres lieux et/ou pour d'autres périodes. L'ensemble des établissements peuvent alors constituer des individus statistiques pertinents que l'on peut caractériser par des variables rendant compte de leurs caractéristiques propres (niveau hiérarchique), temporelles (siècle d'apparition, de disparition, durabilité) et spatiales (localisation, distance au plus proche voisin) comme cela l'est fait dans les chapitres 2.2.4 et 2.2.8.

2.2 Des observables aux individus statistiques

Les observables mobilisés dans les chapitres de cet ouvrage sont ainsi d'une grande variété, à la fois dans leur nature (os, graines, céramique, minerais, traces, parcelles etc.) et dans leur degré d'élaboration en tant qu'objets. Ils constituent, soit directement, soit par les objets que l'on a construits à partir d'eux, les individus et les variables statistiques que l'on pourra représenter et analyser avec des méthodes statistiques. Il est commode de répartir ces objets en quatre catégories : 1- des propriétés existant hors du support d'objets matériels repérable dans l'espace ; 2- des fragments d'objets ; 3- des objets élémentaires ; 4- des objets composites. Le degré de correspondance entre le concept correspondant au questionnement et les individus statistiques soumis à l'analyse diffère suivant ces catégories. Alors qu'il est généralement faible pour les fragments d'objets, il est plus direct pour les objets composites du fait même qu'ils représentent des construits. Un bref exposé du contenu de ces quatre catégories permet d'illustrer la variété des objets formalisés comme individus et variables statistiques dans cet ouvrage.

1- *des propriétés existant sans le support d'objets matériels repérables dans l'espace* : tel est le cas lorsqu'on élabore des mesures en des points de l'espace géographique, en y relevant par exemple l'altitude ou la température. La vision de l'espace correspondant à cette première catégorie est celle d'un champ continu plutôt que d'un espace composé d'objets (Mathian *et al.* 2014). Les individus statistiques sont alors de simples points dans l'espace pour lesquels on a effectué des mesures (altitude ou épaisseur sédimentaire par exemple). Celles-ci peuvent ensuite être utilisées pour construire, à partir d'une interpolation spatiale, une carte représentant la distribution de la variable à laquelle on s'intéresse sur l'ensemble de l'espace considéré. Dans le chapitre 2.2.6 la topographie de la surface des alluvions de la plaine alluviale de Tours a ainsi pu être reconstituée de manière continue. Lorsque l'objectif est de mettre en évidence une densité d'objets (par exemple de vestiges archéologiques), l'espace d'intérêt peut être découpé en mailles régulières afin d'effectuer les comptages appropriés (*cf.* 2.2.3).

Chaque maille est alors un construit qui va servir de support pour évaluer la quantité d'objets.

2- *des fragments d'objets* : les observables de ce type, très courants en archéologie, sont mobilisés dans de nombreux chapitres de cet ouvrage. Il s'agit de vestiges matériels, artéfacts et écofacts, qui se présentent sous la forme de bouts d'objets (tessons de céramique, ossements, coquilles, fragments de matières colorantes, bouts de minerai, morceaux d'obsidienne par exemple). Ces fragments d'objets sont surtout porteurs de sens de par les propriétés physiques, chimiques, pétrologiques ou radiométriques de la matière première utilisée qui vont permettre de repérer leur origine, de les dater ou encore d'en déduire d'autres propriétés particulières. Ces propriétés vont représenter les variables de l'analyse statistique.

Un assemblage ou une composition particulière permettent par exemple d'identifier l'origine géographique (un site ou même un atelier précis) d'une production. Connaissant ainsi la provenance d'un objet, il sera possible d'établir des hypothèses sur les échanges et les mobilités auxquels il a été associé. Les fragments mobilisés comme individus statistiques et dont les analyses physiques et/ou chimiques fournissent les variables, sont de nature très variée dans cet ouvrage. Dans le chapitre 2.1.2 sont ainsi évoqués tessons de céramique, minerais, échantillons géologiques dont l'analyse chimique permet d'identifier l'origine géographique. Dans le cas 2.1.5 ce sont des minerais, scories, inclusions, qui sont analysés. La comparaison de leur signature chimique avec celles obtenues dans de potentiels lieux de gisement ou de production permettra d'en établir l'origine. La provenance des matériaux ferreux utilisés dans la construction du Palais des Papes a ainsi pu être identifiée. Dans le chapitre 2.1.6 il s'agit de fragments d'objets en argile, de déchets de fabrication de potiers, dont la composition chimique permettra d'évaluer si des céramiques sont issues d'un même site de production, celui-ci étant caractérisé par un type de matériau associé à un type de technique. Dans le chapitre 2.1.4 il s'agit de déterminer si des fragments de matières colorantes rouges proviennent de sources géologiques situées à proximité ou sont issus de sites plus lointains, témoignant alors de la possible existence de réseaux d'échanges. Dans le chapitre 2.1.7, enfin, il s'agit de comparer la signature d'artefacts cuivreux avec celle de minéraux extraits en différents lieux afin de tester des hypothèses sur leur provenance.

Dans d'autres cas l'enjeu consiste, à partir de l'analyse de fragments d'objets, à dater un phénomène[2], un événement[3], et aussi à construire des chronologies. Dans le cas du chapitre 1.5, les individus statistiques sont constitués d'échantillons organiques comme des graines déposées en offrande et des fragments de textile recueillis dans une tombe qui ont pu être datés par carbone 14 et insérés dans un modèle statistique permettant de reconstruire la chronologie des règnes en Egypte ancienne. Les tissus dentaires et osseux permettent quant à eux d'apporter des informations sur les conditions de vie des individus (état de santé, modes alimentaires). Dans le cas

[2]Le terme de phénomène est utilisé dans son sens scientifique tel que le définit le CNRTL : « Ce que l'on observe ou constate par l'expérience et qui est susceptible de se répéter ou d'être reproduit et d'acquérir une valeur objective, universelle ».

[3]Le CNRTL donne comme définition : « Tout ce qui se produit, tout fait qui s'insère dans la durée ». Dans le domaine statistique, l'événement fait référence au résultat d'une expérience aléatoire, c'est-à-dire à quelque chose qui se produit (pour reprendre les termes du CNRTL).

3.4, la composition isotopique mesurée sur des fragments d'os et de dents permet ainsi d'évaluer la consommation de protéines des individus concernés. Dans d'autres cas, ces fragments d'objets sont des objets-supports qui présentent l'intérêt de comporter des traces. Dans le cas 3.6, les fragments concernent ainsi des restes osseux consommés : les individus statistiques sont alors des traces de dents et les variables les mesures morphométriques effectuées dessus. Dans 3.5 les individus statistiques sont des squelettes trouvés dans un ensemble de nécropoles de la France septentrionale. Les variables d'intérêt sont l'âge et le sexe dont il s'agit d'analyser la distribution statistique afin de mettre en évidence d'éventuelles surmortalités traduisant l'impact d'événements dramatiques de l'époque mérovingienne, dont on connait l'existence par des textes écrits.

3- *des objets simples* : il s'agit d'objets élémentaires comme des outils, des céramiques, des humains, des tombes, des maisons, que l'on peut observer et caractériser directement. Tel est par exemple le cas des pièces de monnaie (*cf.* 1.3 évoqué plus haut) qui portent une information essentielle, la date d'émission, qui est une variable qui va caractériser aussi bien la pièce elle-même que les éléments qui l'entourent. Le cas des haches en jade, également évoqué plus haut (*cf.* 2.2.7), relève aussi de cette catégorie : pour chaque objet on connait l'origine (même gîte dans les Alpes, identifié grâce à des analyses de matières premières) et le lieu d'utilisation finale qui est le lieu de découverte, souvent dans des tombes (il s'agit d'un « objet-signe »). Les autres variables caractérisant la hache sont de nature variée. Il s'agit de sa morphologie, de son niveau de polissage, du type de forme. Par ailleurs, les coordonnées géographiques du gîte et du lieu de découverte permettent de calculer la variable « distance à la source ». Chaque individu-hache en tant que tel présente un intérêt pour les chercheurs qui cherchent à reconstituer sa « biographie », en identifiant les étapes intermédiaires qu'elle a réalisées entre son lieu de production et le lieu où elle a été trouvée, afin de comprendre les mécanismes sous-jacents à la diffusion de ces haches.

4- *des objets composites* : certains objets mobilisés dans les travaux ne sont pas directement observables en tant que tels mais représentent le fruit d'une construction. Tel est le cas des établissements (*cf.* 2.2.2, 2.2.4 et 2.2.8). Sur un plan conceptuel, leurs propriétés d'intérêt sont la fonction socio-économique, le niveau hiérarchique, le statut de tête de réseau ou non (qui est associé à leur capacité de polarisation). Les observables qui serviront à reconstituer ces informations sont issues des vestiges matériels (superficie, qualité des matériaux, qualité du mobilier) relevés sur le site. Le statut de l'établissement par exemple pourra être déterminé en fonction des traces de sépulture trouvées dans des sites funéraires associés par l'archéologue à l'établissement étudié. La combinaison de ces critères permet, à l'aide d'une analyse en composante principales (ACP), de construire une variable synthétique mesurant le niveau hiérarchique de l'établissement (*cf.* 2.2.8). D'autres variables caractérisent la localisation des établissements : coordonnées géographiques (X,Y) qui permettent de calculer la distance à l'établissement le plus proche existant à une époque contemporaine, ou encore la distance aux n établissements les plus proches qui permet de caractériser la position relative de l'établissement dans le réseau d'habitat, d'avoir une idée de la nature des interactions que chaque établissement pourrait entretenir avec les autres et des formes d'organisation de l'espace (*cf.* 2.2.4). Enfin, sont considérées des variables

temporelles (siècle d'apparition, de disparition) qui permettent d'estimer la durée de vie des établissements.

Les « pôles », lieux identifiés par le nombre importants de haches qui y ont été trouvées dans le cas d'étude 2.2.7, sont également un exemple d'objets composites. Ces objets, relevant d'un niveau d'agrégation supérieur (en tant que lieux desservis), sont ensuite formalisés comme les nœuds d'un graphe. Cette étude de cas repose ainsi sur des objets relevant de différents niveaux d'agrégation, chacun caractérisé par un ensemble propre de variables. Alors que les variables décrivant la hache relèvent d'observations relativement directes de leurs caractéristiques morphométriques et composition physico-chimique, celles décrivant les nœuds comprennent d'un côté un dénombrement de haches et de l'autre des variables décrivant la localisation de ces nœuds. Celles-ci sont représentées par des indicateurs plus sophistiqués calculés à l'aide de l'analyse des graphes : accessibilité, centralité, connectivité et intermédiarité. Ceux-ci donnent une image complète de la localisation relative d'un nœud dans un réseau de pôles.

Dans le cas des fragments d'objets, on ne s'intéresse guère à chaque individu statistique pris individuellement. C'est la composition d'ensemble qui importe, c'est elle qui permet de cerner une origine ou une date, d'identifier des pratiques ou de caractériser des ressources. Les objets composites en revanche, s'ils sont eux aussi intéressants dans leur ensemble, renseignant par exemple sur le caractère plus ou moins concentré ou dispersé du semis d'habitat, sont également pertinents à considérer dans leur individualité. Le fait par exemple qu'un établissement soit plus important que les autres ou localisé en périphérie au sein d'un réseau d'échanges, permet de comprendre la dynamique propre à cet établissement.

3 Les grandes familles de méthodes statistiques

On distingue généralement les statistiques descriptives et les statistiques mathématiques prenant appui sur la théorie des probabilités. Les premières sont ancrées dans les données, dans leur description, analyse, représentation, dont il s'agit d'extraire des informations, des connaissances, sur le phénomène auquel on s'intéresse. Dans ces approches les traitements portent exclusivement sur les observations dont on dispose et les résultats de ces traitements sont donc associés au lieu et à la période concernés par ces données, même s'il est possible de monter en généralité, notamment à partir de la comparaison de plusieurs cas empiriques. Dans cet ouvrage cinq chapitres leur sont consacrés, les chapitres 4, 5 et 6 portant sur les « Outils statistiques », avec une complexification croissante en fonction du nombre de variables considérées et les chapitres 9 et 10 consacrés à des méthodes de classification. Dans le chapitre 11, dédié aux statistiques spatiales, des méthodes descriptives sont également présentées, adaptées aux cas où les variables concernent des localisations et des distances entre objets.

Les méthodes de statistique mathématique se placent d'emblée sur un plan général et prennent appui sur des lois de probabilités. Il est utile ici de distinguer deux familles de méthodes. La première est composée des statistiques inférentielles dont l'objectif est d'inférer (chapitre 8) à l'ensemble de la population statistique à laquelle on s'intéresse (qu'il s'agisse d'une population d'établissements, de squelettes, de céramiques,

de pièces de monnaie etc) des résultats obtenus à partir de calculs réalisés sur un échantillon, qu'il s'agisse de la mise en évidence de relations entre des phénomènes différents ou de la comparaison entre des ordres de grandeur mesurés sur des populations différentes par exemple. Une telle inférence des résultats obtenus sur des échantillons à des populations globales entraîne un risque d'erreur qu'il est indispensable de préciser. Il permet en effet d'associer aux résultats leur significativité statistique. Les chapitres 7 et 8 sont consacrés respectivement aux notions indispensables sur les lois de probabilité et aux méthodes d'inférence statistique. Le chapitre 11, dédié aux statistiques spatiales, présente l'intérêt d'une approche inférentielle pour évaluer si la distribution spatiale d'un phénomène ou l'organisation d'un semis de points dans l'espace géographique (par exemple des établissements) s'écartent ou non significativement d'une répartition aléatoire. Les statistiques bayésiennes enfin, représentent une deuxième famille de méthodes reposant sur des calculs de probabilité. Elles permettent d'introduire des connaissances *a priori* sur le phénomène étudié (la prise en compte de tables de mortalité lorsqu'on cherche à dater un squelette, par exemple). Le chapitre 12 leur est consacré.

Dans chacune de ces approches, les observables dont on dispose ont un statut différent. Dans le premier cas l'enjeu consiste à décrire le mieux possible les tendances générales et les relations que révèlent les données analysées. La démarche est souvent exploratoire, on cherche à identifier des régularités, des organisations, concernant le phénomène auquel on s'intéresse. Dans le deuxième cas on a une interrogation précise que l'on formule sous forme d'une hypothèse[4] statistique (sur des différences entre diverses populations statistiques, sur des relations entre différents phénomènes) et les méthodes statistiques permettent de tester ces hypothèses à partir des données, en attachant un risque d'erreur aux résultats. Dans le troisième cas, on dispose d'informations *a priori* sur le phénomène étudié, émanant de travaux précédents ou d'expertises, permettant d'établir des lois de probabilités dites *a priori*, qui sont ensuite combinées avec les données observées afin de corriger ces lois et produire des lois dites *a posteriori* qui permettent notamment de faire des estimations.

Ces trois familles de méthodes statistiques sont toutes trois mobilisées dans les cas d'étude évoqués dans les trois premiers chapitres de cet ouvrage, souvent de manière combinée.

3.1 Les statistiques descriptives : entrer par les données

Les statistiques descriptives permettent d'abord, dans un cadre dit univarié, de mettre en évidence les grandes tendances du phénomène auquel on s'intéresse, notamment en calculant les ordres de grandeur de chacune des variables qui le caractérise (dont rendent compte la moyenne, la médiane, le mode) et leur degré de variabilité d'un individu statistique à une autre (mesuré par l'écart-type ou le coefficient de variation).

Dans le cadre univarié, les outils de représentation de base sont les histogrammes, les courbes, la cartographie. Ces outils relativement simples sont très utiles, d'une part

[4] L'expression « hypothèse statistique » est utilisée afin d'éviter toute confusion : le chercheur qui mobilise des méthodes descriptives a bien évidemment des hypothèses de travail au départ de sa recherche. L'hypothèse au sens statistique s'exprime sous forme d'une hypothèse dite nulle (H0) et d'une hypothèse dite alternative (H1).

en amont pour obtenir une première vision de la distribution statistique des observables analysés, et d'autre part comme aboutissement lorsqu'il s'agit de représenter des résultats pour des objets et/ou des variables résultant d'une construction élaborée. Tel est le cas dans le chapitre 2.2.4 où le travail a consisté à créer un indicateur adéquat, qui permet de mesurer et de comparer l'organisation spatiale de l'habitat, pour des régions et des périodes différentes, à partir de modes de représentation relativement simples (courbes etc). Les individus statistiques (des établissements) et les variables (par exemple la distance aux n établissements les plus proches) sont ici des construits relativement élaborés auxquels on applique, pour visualiser l'évolution de l'organisation spatiale du peuplement, des méthodes statistiques élémentaires. Un autre exemple de mesures élaborées est donné par le cas 2.2.7 : des centres principaux et des centres secondaires sont identifiés à partir du degré de concentration de haches en jade trouvées en différents lieux d'Europe et considérés comme les nœuds d'un réseau pour lequel on peut calculer divers indicateurs de centralité. Un résultat important sur la relation entre le niveau des nœuds (principaux ou secondaires) et leur niveau de centralité est représenté sous simple forme d'histogramme alors que leur identification et caractérisation en tant qu'objets sont le fruit de méthodes avancées d'analyse des graphes.

Dans un cadre bivarié, on va de plus s'intéresser à la relation entre les différentes variables considérées prises deux par deux. On peut par exemple évaluer, au niveau des établissements, s'il existe une relation entre l'importance des vestiges matériels et les caractéristiques de l'environnement dans lequel ils sont situés (*cf.* 2.2.8), ou encore entre les différents niveaux de polissage des haches en jade trouvées en différents lieux d'Europe et leurs distances moyennes aux gîtes de production (*cf.* 2.2.7).

Parmi les variables mobilisées, une joue un rôle essentiel en analyse spatiale, il s'agit de la distance : distance à un établissement ou à un point de mesure voisin, distance à un lieu structurant, ou encore l'accessibilité d'un nœud au sein d'un graphe. Tout un champ de recherche, la géostatistique, a été développé pour rendre compte des effets des localisations respectives des objets étudiés dans l'espace (*cf.* 11). L'approche nécessite de travailler sur des objets géoréférencés. Le variogramme (qui prend appui sur la notion de corrélation entre deux points de mesure en fonction de la distance qui les sépare) est par exemple mobilisé dans le chapitre 2.2.6, ainsi que différentes méthodes d'interpolation spatiale permettant d'estimer les valeurs d'une variable, par exemple l'épaisseur sédimentaire, en des lieux où il n'y a pas eu de mesure.

Dans un cadre multidimensionnel enfin, où l'on dispose de nombreuses variables pour caractériser le ou les phénomènes étudiés, des traitements plus élaborés (analyse graphique, analyses factorielles, classifications) peuvent être réalisés. La démarche consiste à analyser les données (issues des observations, des mesures, des comptages) et à rechercher des tendances, des organisations, des régularités, qui apportent des connaissances sur le phénomène étudié.

La qualité des résultats des analyses statistiques dépend beaucoup de celle des données. D'abord se posent les questions sur l'incertitude et l'incomplétude des données (Batton-Hubert et al, 2017), que celles-ci soient manquantes, irrégulières ou encore mal mesurées. Les risques de biais associés sont abordés dans de nombreuses études

de cas (*cf.* 1.4, 3.2, 2.2.5 et 2.1.7). Dans le cadre d'une analyse statistique multidimensionnelle, se posent ensuite les questions d'homogénéité et d'exhaustivité statistiques, critères largement mis en avant dans les travaux de Benzécri (1973a, 1973b). L'exhaustivité consiste à prendre en compte toutes les dimensions caractérisant le phénomène étudié. Tel est par exemple le cas dans le chapitre 2.2.8 quand les auteurs combinent des variables portant sur les vestiges matériels, les caractéristiques environnementales, situationnelles et chronologiques des établissements pour analyser le système de peuplement. Tous les aspects (ou du moins un nombre important) permettant de caractériser les établissements et d'évaluer leur position relative dans le système de peuplement, sont ainsi pris en compte. Quant au principe d'homogénéité, il consiste d'une part à considérer des individus statistiques de même nature, ce qui est ici le cas des établissements, tous construits suivant les mêmes principes, et d'autre part de donner un poids similaire aux différentes dimensions les caractérisant. Un contre-exemple aurait par exemple consisté à mobiliser une dizaine de variables pour décrire le niveau hiérarchique de l'établissement et une seule pour décrire sa situation environnementale. Dans le cas d'étude 2.2.8 les auteures ont au contraire procédé, lors d'une première phase d'exploration des données, à une analyse d'indépendance entre les variables afin d'éliminer des indicateurs qui seraient redondants et n'apporteraient pas d'information nouvelle lors de la construction de la variable caractérisant le niveau hiérarchique de chaque établissement. En effet si on garde dans l'analyse factorielle deux critères reflétant les mêmes différenciations entre les établissements, cela revient à doter le phénomène sous-jacent d'une pondération de 2 relativement aux autres. Si l'on considère la propriété « taille de l'établissement » par exemple, il ne serait ainsi pas opportun de mobiliser comme variables à la fois la superficie, le périmètre et la longueur maximale. Un tel travail préalable est indispensable pour donner toute leur puissance aux résultats d'analyses factorielles ou de classification, méthodes largement utilisées dans les études de cas (*cf.* 2.1.2, 2.1.3, 2.1.4, 2.1.5, 2.1.6, 2.1.7, 2.2.8, 3.6 et 1.3).

Dans le cas 2.2.8, les individus statistiques, des établissements, sont des objets géographiques dans la mesure où leurs localisations respectives jouent un rôle dans ce qu'ils sont, notamment de par leurs capacités d'interactions avec les autres objets géographiques. Le chapitre 11, qui propose une entrée dans la démarche statistique par la géographie, souligne la spécificité de tels individus statistiques, notamment de par leur non indépendance, des lieux proches dans l'espace tendant à entretenir des relations d'interdépendance. Quand les données sont soumises à des analyses factorielles, c'est en général *a posteriori* que les proximités spatiales entre les individus statistiques sont interprétées, par exemple par la mise en relation entre leurs positions respectives dans les plans d'axes factoriels et dans l'espace géographique (Lebart *et al.* 2006, Sanders, 1989).

3.2 Les statistiques inférentielles : identifier des effets de structure

Les statistiques inférentielles (chapitre 8) prennent appui sur les tests d'hypothèses. L'inférence des résultats de traitements effectués sur un échantillon à la population entière concernée par le questionnement et dont est extrait cet échantillon, implique un risque d'erreur que l'on peut contrôler en prenant appui sur des lois de probabilités.

Supposons, à titre d'exemple, que l'on s'intéresse à la relation entre deux variables V1 et V2 (par exemple la superficie des établissements dans le Languedoc et leur spécialisation fonctionnelle). Le test d'hypothèse sera formulé de la façon suivante :

H0 (hypothèse nulle) : absence de relation, les variables V1 et V2 sont indépendantes

H1 (hypothèse alternative) : existence d'une relation entre les deux variables V1 et V2.

Si les variables en jeu sont quantitatives et suivent des lois à peu près normales, alors le statisticien a les moyens, à partir de calculs empiriques, de déterminer laquelle de ces deux hypothèses est la plus juste. En revanche, la qualité de sa conclusion dépendra de la nature même de cette conclusion. En effet, si les calculs empiriques conduisent à rejeter H0 (et donc à conclure que H1 est vrai), cette conclusion sera associée à un risque d'erreur contrôlé, qui peut être de l'ordre de 5%, 1% ou même moins (cette erreur, qualifiée de « erreur de 1ère espèce » indique la probabilité de se tromper en affirmant que H0 est faux). Si les calculs empiriques conduisent au contraire à accepter H0, la probabilité que l'on a de se tromper en affirmant que H0 est vrai, qualifiée de « erreur de 2e espèce », a une valeur en général bien plus élevée et est parfois difficile à estimer. Il y a ainsi une dissymétrie dans le rendu des résultats auquel l'utilisateur ne peut évidemment rien puisque ce sont les données qui déterminent dans quel cas on se trouve. Cette dissymétrie amène Otarola-Castillo et Torquato (2018) à écarter les statistiques inférentielles classiques reposant sur les tests d'hypothèse pour leur préférer les statistiques bayésiennes. Si ces dernières se sont beaucoup développées depuis les années 2000, notamment en archéologie, il n'en reste pas moins intéressant de maintenir dans la panoplie des outils statistiques à disposition, les statistiques inférentielles à base de tests d'hypothèse.

Les statistiques inférentielles sont en fait très satisfaisantes lorsqu'il s'agit de mettre en évidence quelque chose de structuré, qui s'écarte de l'aléatoire. Dans l'exemple ci-dessus, il s'agissait de la relation entre deux phénomènes. Différents tests peuvent être appliqués dans ce cas, suivant la nature des variables en jeu. Si les deux variables sont quantitatives, le test porte sur le coefficient de corrélation, quand les deux variables sont qualitatives, il s'agit du test du chi2 (utilisé par exemple dans le chapitre 2.2.7 pour étudier le lien entre le degré de finition des haches de jade et la distance au Mont Viso – le lieu d'origine – suivant 15 classes) et quand on s'intéresse à la relation entre une variable qualitative et une variable quantitative, on mobilise l'analyse de la variance. Dans chaque cas, H0 correspond à l'hypothèse d'absence de relation, c'est-à-dire à une corrélation nulle dans le cas quantitatif et à l'indépendance entre les variables dans le cas qualitatif.

De façon générale, dans un test d'hypothèses, l'hypothèse nulle correspond à une situation aléatoire ou non structurée. Quand on compare les moyennes ou les variances d'une variable dans deux sites ou deux régions géographiques différents, H0 correspond ainsi à l'égalité des moyennes ou des variances. Quand on s'intéresse à la forme d'une distribution spatiale, H0 correspond à un semis aléatoire (*cf.* chapitre 11). Dans un modèle statistique (de type régression multiple ou régression logistique) où il s'agit d'expliquer (au sens statistique) une variable par la combinaison de plusieurs autres,

H0 correspond à la situation où les coefficients associés à ces variables sont nuls. Dans le chapitre 1.3.3 par exemple, une régression est réalisée avec pour variable « à expliquer » la date (donnée obtenue à partir de pièces de monnaie pour un certain nombre de sites) et comme « variables explicatives » les axes d'une analyse factorielle faite au préalable pour caractériser le facies céramique du site. Les coefficients de la régression s'avèrent tous significativement différents de 0 et la qualité de cette régression est telle qu'elle permet d'en utiliser les coefficients pour dater des assemblages d'après leur facies dans des lieux où aucune donnée de date n'est disponible.

Les tests dits « paramétriques » sont appliqués dans le cas où les variables en jeu suivent des lois normales. Pour le vérifier, il est nécessaire de disposer d'échantillons d'une certaine taille mais il existe également des tests dits « non paramétriques », appropriés pour le cas des petits échantillons. Un tel test est utilisé dans le chapitre 3.3 pour comparer les isotopes correspondant aux ressources alimentaires (à partir de vestiges bioarchéologiques) dans deux milieux différents (montagne et vallée) dans le sud du Caucase.

Dans chacun de ces cas, le rejet de H0 conduit donc à conclure à l'existence d'un effet de structure, c'est-à-dire des conclusions du type suivant : les moyennes entre les populations sont différentes, la répartition spatiale du semis étudié est non aléatoire, il y a une relation significative entre les variables, les coefficients de régression sont différents de 0 etc. Une telle conclusion correspond à une erreur contrôlée (l'erreur de 1re espèce), et plus cette erreur est faible, plus on a mis en évidence un effet important. Une limite est cependant que cette erreur dépend du nombre d'observations en jeu, plus celui-ci est grand, plus facile il sera de mettre en évidence des effets de structure. En revanche, si les calculs conduisent à accepter H0, le chercheur reste sur sa faim, le risque de 2^{e} espèce étant rarement connu.

3.3 Les statistiques bayésiennes : valoriser des connaissances *a priori* et raisonner de manière itérative

L'approche bayésienne repose sur la prise en compte de connaissances *a priori*, qu'il s'agisse d'informations ou de croyances. Celles-ci peuvent en effet s'exprimer sous des formes très variées, il peut s'agir de connaissances expertes, de travaux empiriques effectués précédemment dans un lieu ou une période similaire, ou encore, dans le cas de phénomènes bien connus comme les comportements démographiques, de tables de probabilités rendant compte par exemple de la mortalité par classe d'âge en un lieu et une période donnés. Dans chacun de ces cas il sera possible d'établir des probabilités *a priori*, tenant compte des connaissances antérieures sur le phénomène étudié (qu'il s'agisse d'estimer une valeur comme une date ou d'évaluer une différence dans des pratiques alimentaires, par exemple), avant même de s'intéresser aux données empiriques dont on dispose. L'analyse des données permet ensuite de corriger cette loi de proba *a priori* et d'obtenir une distribution de probabilité dite *a posteriori* qui donne une estimation du phénomène étudié. Cette distribution de probabilités est à la fois un résultat et l'éventuel point de départ de nouveaux travaux disposant d'autres données, dont elle représente alors la probabilité *a priori*. L'inférence bayésienne (chapitre 12) repose ainsi sur des distributions de probabilité que l'on peut faire évoluer de manière itérative à partir des observations empiriques pour se rapprocher

progressivement d'une meilleure estimation du phénomène étudié. Si on ne dispose d'aucune connaissance ni croyance *a priori*, l'outil reste pertinent, il suffit alors à l'utilisateur de supposer des probabilités équivalentes pour les événements en jeu. Alors que les données ont un statut de vérification d'une hypothèse dans le cas des méthodes inférentielles classiques, elles ont un statut d'enrichissement des connaissances et croyances dans l'approche bayésienne.

Cette approche prend appui sur la notion de probabilité conditionnelle. Un exemple simple permet de comprendre cette notion. Supposons que l'on s'interroge sur la pertinence de fouiller une parcelle donnée P de 1 ha dans une aire géographique dont on connait globalement le potentiel archéologique sous la forme d'une densité de sites par ha. Celui qui ne dispose d'aucune autre information estimera la probabilité de trouver du matériel archéologique sur cette parcelle P à partir de cette densité. Il existe par ailleurs des travaux montrant que le potentiel archéologique d'une zone dépend de la nature géologique du sol. Les sols limoneux, par exemple, peuvent s'avérer plus riches en matériau archéologique. La question devient alors : sachant que (pour prendre un vocabulaire statistique) le sol de la parcelle P est de type limoneux, quelle est la probabilité d'y trouver du matériel archéologique ? La probabilité de trouver des vestiges dans la parcelle P sachant que le sous-sol est de type limoneux sera ainsi différente de celle d'en trouver sans cette information supplémentaire. Le chapitre 2.2.5 passe en revue trois approches, inductive, déductive et bayésienne pour évaluer dans quelle mesure un espace donné contient un potentiel archéologique.

Si on note A l'événement « il y a du matériau archéologique dans la parcelle P » et B l'événement « la parcelle P a un sol limoneux », le théorème de Bayes s'exprime de la façon suivante :

$$P(A/B) = P(B/A).P(A)/P(B)$$

L'événement A est celui dont on recherche la probabilité et l'événement B correspond à une information que l'on peut observer. P(A/B), la probabilité que l'événement A se produit sachant que l'événement B est réalisé, représente alors la probabilité *a posteriori* de trouver du matériau archéologique dans la parcelle P.

L'approche bayésienne a connu un développement important dans le domaine de l'archéologie, beaucoup plus qu'en géographie par exemple. Le fait que ces méthodes sont particulièrement bien adaptées lorsque les données sont incertaines et en faible quantité l'explique probablement. Dans cet ouvrage, l'approche est adoptée dans plusieurs études de cas relevant de différentes catégories de questions, illustrant ainsi la variété des applications possibles et notamment la variété des connaissances mobilisées pour construire les probabilités *a priori*.

Pour les cas d'étude répertoriées dans la catégorie « Quand » (chapitre 1), l'objectif est en général d'estimer la date d'un événement ou d'établir une chronologie. Les probabilités *a priori* sont alors déterminées à partir d'informations archéologiques sur l'intervalle de temps dans lequel un événement a pu se produire ou sur l'ordre de succession de différents événements. Les données observées sont quant à elles composées de mesures physiques (par radiocarbone, luminescence, archéomagnétisme), opérées sur des artefacts de nature diverse, afin d'estimer leurs âges. La combinaison de telles

informations chronologiques issues de l'expertise archéologique et des datations par mesures physiques, permettent de construire les probabilités *a posteriori* servant à reconstruire les chronologies étudiées. Telle est la démarche mobilisée dans les cas exposés dans les chapitres 1.2 et 1.5, dans le premier cas pour identifier la succession de différentes phases culturelles dans le cadre du paléolithique supérieur en France et dans l'autre pour établir la chronologie des pouvoirs en Egypte ancienne.

Dans le chapitre 3.5, l'objectif est de détecter d'éventuelles traces sur le plan démographique d'événements dramatiques dont on connait, par des textes écrits, l'existence en basse Normandie pendant la période mérovingienne. La variable d'intérêt est l'âge au décès et la probabilité *a priori* de décès pour chaque classe d'âge est déterminée par une loi centrée sur le « standard préindustriel » pour lequel on dispose de tables de mortalité. Les données observées portent sur les squelettes, trouvés dans différentes nécropoles, dont on a pu déterminer le sexe et l'âge au décès. La loi de probabilité *a posteriori* a pu être déduite de la combinaison de ces deux éléments. Un travail de comparaison avec d'une part d'autres régions (Ile de France et Champagne-Ardenne) à la même période et d'autre part la basse Normandie à la période gallo-romaine, a permis de mettre en évidence une surmortalité des hommes jeunes en basse Normandie lors de cette période de crises.

Dans le cas 2.2.5, l'objectif est de prédire la localisation de sites archéologiques et l'étude de cas porte sur la région de Rijssen-Wierden au Pays-Bas. Dans cette étude, les probabilités *a priori* sont déterminées à partir des avis d'experts sur les chances de trouver un site en fonction du type de texture des sols. Les données sont représentées par les résultats d'un modèle de prédiction réalisé antérieurement et prenant appui sur 80 sites observés. Des itérations entre ces deux sources de connaissance permettent de construire une carte des densités de sites et d'estimer les effets d'incertitude (des experts et des données) sur ces estimations.

Plutôt que d'opposer ces différentes familles de méthodes, méthodes descriptives d'un côté, méthodes d'inférence de l'autre, classiques ou bayésiennes, il est fructueux de prendre appui sur leurs complémentarités. Plusieurs des cas d'étude illustrent d'ailleurs l'intérêt de combiner des méthodes issues de ces différentes famille. Dans le cas 1.3., une régression multiple, relevant d'une approche inférentielle classique, est ainsi réalisée sur les axes factoriels obtenus à partir de l'analyse factorielle des correspondances (AFC) d'un tableau de contingence portant sur des types de céramiques. Dans le chapitre 2.2.5, des méthodes descriptives comme la corrélation et des tests du chi2 sont utilisées dans un premier temps pour identifier des catégories environnementales propices pour contenir des sites archéologiques, résultats ensuite utilisés par des experts pour évaluer les attractivités relatives de différents profils environnementaux, ce qui a permis de construire des probabilités *a priori* et donner lieu à l'application d'une démarche bayésienne. Pour étudier la diffusion des haches de jade (2.2.7), sont combinées techniques descriptives, test du chi2, classifications, régressions et analyse des graphes. Les différentes études de cas montrent ainsi combien est fructueuse la combinaison de méthodes différentes ainsi qu'une démarche itérative.

4 Questions génériques et méthodes statistiques : diversité des choix à la disposition de la recherche archéologique

« Le point de départ de toute analyse est une problématique » comme le soulignent les auteurs du chapitre 2.1.2 où ils rendent compte, pour les recherches de provenance, de l'ensemble de la démarche méthodologique menant de la problématique aux résultats, en passant par la formalisation des données. Si le point de départ est effectivement fondamental et implique souvent des démarches bien précises, on peut aussi noter que face à une même question, les approches statistiques peuvent être différentes suivant le choix des chercheurs, choix influencé par leurs habitudes et compétences propres. Afin de rendre compte de ce double aspect, réponses statistiques convergentes ou différentes pour répondre à un même questionnement, il est intéressant de parcourir d'abord les grandes catégories de questions que l'on se pose en archéologie (4.1), puis les catégories de « tâches » que permet l'approche statistique et dont on trouve des exemples dans cet ouvrage (4.2). Enfin, seront abordées les précautions et réflexions nécessaires pour que les résultats des traitements statistiques donnent tous leurs fruits (4.3).

4.1 Diversité des choix face aux questions « Quand », « Où » et « Quoi/Qui »

Dans les premiers chapitres de l'ouvrage, les études de cas sont organisées suivant la nature de la question archéologique appréhendée à un niveau très général. La triade élaborée par Peuquet (1994) est en effet utilisée pour formuler les questionnements en des termes très génériques suivant trois dimensions articulées autour des questions « Quand », « Où » et « Qui/Quoi ». La question « Quand » renvoie à la dimension temporelle du phénomène étudié et les études présentées portent sur l'estimation de l'âge de divers artefacts, sur la datation d'un événement ou la reconstitution d'une chronologie. Ces travaux permettent d'identifier la co-occurrence ou le décalage temporel entre différents processus. La question « Où », référant à la dimension spatiale du phénomène, comporte deux parties. La première s'intéresse à l'identification de la provenance de différents objets et matériaux, ce qui permet d'avoir des informations sur les lieux de production de ces objets et sur leur circulation. Ces informations permettent d'émettre des hypothèses sur les formes de complémentarités entre des lieux différents et éventuellement sur les mobilités des personnes et des biens. La seconde partie du « Où » porte sur la localisation d'objets dans l'espace, ce qui permet d'étudier leurs répartitions géographiques et de mettre en évidence des formes différentes d'occupation du territoire et de modéliser des interactions spatiales. Enfin, la partie portant sur la question « Qui/Quoi » réfère à la dimension thématique d'une recherche et présente des exemples d'estimation de l'existence, de l'importance ou de la nature de différents faits et phénomènes. Ces catégories sont ainsi très génériques, ce qui permet de rapprocher et comparer des travaux concernant des objets, des périodes et des lieux très différents, ce qui est un des intérêts de cet ouvrage.

Pour la question de « Où – provenance » les différentes études de cas exposées mobilisent assez systématiquement une même chaîne de méthodes d'analyse des données, d'abord analyse en composantes principales (ACP) et classification ascendante hiérarchique (CAH) pour mettre en évidence des profils (de faciès, de composition physico-

chimique) et ensuite l'analyse discriminante (chapitre 10) pour identifier si un profil donné peut être rapproché d'un profil de référence. Pour l'ensemble des autres cas d'étude, on ne note en revanche pas de correspondance systématique entre les grandes catégories de questions et les méthodes mobilisées. La question du « Quand » est ainsi abordée tant par les statistiques bayésiennes que par la régression multiple associée à une AFC, et dans un cas par la matrice de Bertin (2005, *cf.* 1.6). Les études de cas rangées sous la question « Où – localisation » et « Qui/Quoi » mobilisent quant à elles l'ensemble des famille de méthodes énoncées plus haut, en les combinant de différentes façons : méthodes de statistique descriptive d'abord, qu'il s'agisse de simples statistiques univariées, appliquées à des indicateurs parfois très élaborés, ou d'analyses factorielles et de classification, ou encore de statistiques spatiales (variogramme, interpolation), de statistiques inférentielles ensuite, à base de tests d'hypothèse (régression spatiale, chi2, tests de comparaison de moyennes) et enfin de méthodes bayésiennes.

4.2 Principales « tâches » réalisées dans les études de cas grâce à des outils statistiques

L'objectif d'une recherche archéologique est de rendre compte de ce qui a été, faits et processus du passé, en utilisant traces et vestiges pour reconstruire des objets (des établissements, des pôles) et des chronologies, pour estimer des paramètres (des dates, des flux, des stocks), ou faire des prévisions (sur les potentialités archéologiques d'un espace par exemple). L'approche statistique contribue à un tel projet par un certain nombre de tâches relativement génériques que ses méthodes permettent de réaliser : identifier et décrire les phénomènes étudiés, effectuer des comparaisons entre des périodes ou des lieux différents, mettre en relation des phénomènes divers. Les études de cas illustrent, à nouveau, la diversité des approches statistiques qui peuvent être mobilisées pour accomplir ces tâches.

4.2.1 Décrire ou prédire, estimer, construire et identifier

Face à un phénomène donné, l'objectif est de le *décrire* de manière pertinente. Afin d'éviter toute confusion, il est utile de distinguer deux phases dans cette tâche essentielle. La première se situe en amont de la recherche : il s'agit de faire le point sur les observables en jeu, de caractériser leur nature, leur qualité (homogénéité, complétude, degré d'incertitude des mesures), d'évaluer les ordres de grandeur des mesures effectuées (qu'elles résultent de dénombrement ou de mesures physico-chimiques). La deuxième phase porte sur la description des objets construits au cours de la recherche et les cas d'étude exposés dans ce livre développent plutôt cet aspect : il s'agit de décrire, non plus les observables, mais les objets caractérisant le phénomène étudié. Les méthodes de description, prenant appui sur les statistiques descriptives, les représentations graphiques ou la cartographie, peuvent alors être très élémentaires, toute la complexité de la démarche résidant dans la construction des objets et de leurs propriétés à partir des observables à disposition. Tel est le cas, par exemple, quand il s'agit de *décrire* l'évolution spatiale du système d'habitat (*cf.* 2.2.4) en termes de plus ou moins dense, plus ou moins régulier, en voie de concentration ou de dispersion. La description prend appui sur de simples calculs de fréquence, de densité et des repré-

sentations cartographiques, mais les objets (les établissements) comme les variables spatiales mobilisées sont des construits très élaborés.

Dans d'autres cas où l'étape de description finale des résultats est également simple, c'est la méthode d'estimation ou de construction des indicateurs qui est élaborée. Cette situation se présente notamment lorsque l'enjeu consiste à *estimer* des paramètres, par exemple la date d'un événement, ou de reconstituer une chronologie, que ce soit en mobilisant un modèle de régression multiple (*cf.* 1.3) ou l'approche bayésienne (*cf.* 1.2 et 1.5). La situation est similaire quand il s'agit de *construire* des indicateurs, par exemple un faciès de céramique (*cf.* 1.3) ou une signature chimique (*cf.* 2.1.2, 2.1.3 et 2.1.6), afin de caractériser un lieu de production. Dans le chapitre 3.5 l'objectif est d'*identifier* des événements démographiques et l'âge moyen au décès, estimé pour différentes régions à l'aide d'une approche bayésienne, est représenté sous forme de courbes de densité de probabilité. Dans le chapitre 2.2.3 un Système d'Information Géographique (SIG) est mobilisé pour développer une approche par mailles aboutissant à la cartographie des lieux de concentration des vestiges à différentes périodes et à la mise en évidence des changements dans le temps.

Dans plusieurs cas les résultats sont présentés sous forme de carte. Cela est par exemple le cas lorsque le but est de *construire* des données homogènes pour un territoire alors que l'on dispose soit d'un petit nombre d'observables sur le phénomène étudié, soit d'observables indirects sur des phénomènes qui lui sont liés. Dans le chapitre 2.2.5 par exemple, le but est ainsi de *prédire* la densité ou l'existence de sites archéologiques dans une zone à partir de ses caractéristiques environnementales (topographique, taphonomique, géologique). Différentes approches statistiques sont mobilisées, du test du chi2 et de la régression logistique à l'approche bayésienne et le résultat est présenté sous la forme d'une carte de densité relative de sites archéologiques potentiels. Dans un même esprit, l'objectif du chapitre 2.2.6 est de *construire,* à partir de la connaissance de données ponctuelles, une surface située en sous-sol à l'aide d'un modèle d'interpolation. Les données, issues de sources diverses (carottages géologiques, sondages géotechniques etc), sont harmonisées sous forme de logs stratigraphiques indiquant les caractéristiques géologiques des différentes couches du sous-sol. Le résultat est une carte représentant de manière continue la variable d'intérêt, par exemple l'altitude des alluvions (*cf.* 2.2.6).

Enfin, dans le chapitre 2.2.8, un des objectifs est de *construire* les liens potentiels entre les établissements alors que l'on ne dispose d'aucun observable pour les mesurer. Les auteures proposent d'utiliser un modèle d'interaction spatiale pour *estimer* l'intensité potentielle de ces liens, modèle fondé sur l'hypothèse qu'ils sont proportionnels aux tailles respectives des établissements et inversement proportionnels à la distance les séparant. Un tel modèle peut être développé quand il existe des observables permettant d'évaluer le niveau d'importance des différents établissements ainsi que leurs localisations respectives.

4.2.2 Comparer des ordres de grandeur ou des signatures

La comparaison est un outil essentiel dans le raisonnement scientifique et les méthodes statistiques offrent de nombreuses possibilités pour comparer les caractéristiques de

différentes populations d'objets. Un premier ensemble d'applications concerne la comparaison à travers le temps. Dans le chapitre 3.4 par exemple, deux phases d'occupation funéraire, correspondant à des pratiques mortuaires différentes, ont pu être identifiées dans un site précolombien du Pérou. Ces populations aux comportements funéraires différents avaient-ils également des pratiques alimentaires différentes ? Telle était la question posée dans ce cas d'étude et la comparaison, reposant sur une analyse isotopique des os et dents retrouvés, a été effectuée à l'aide d'une approche bayésienne.

Un deuxième ensemble concerne des comparaisons dans l'espace. Dans le chapitre 3.5 évoqué plus haut, il s'agissait de comparer l'âge moyen au décès dans trois régions françaises à l'époque mérovingienne afin de mettre en évidence les effets sur la démographie d'événements violents survenus dans la Basse Normandie. Dans le cas 2.2.4 l'objectif est de comparer la densité et l'organisation spatiale de l'habitat de deux micro-régions de la Gaule. La comparaison repose sur des courbes représentant la distribution d'une part des distances de chaque établissement à chacun des autres pour différentes dates (1er s. av. J.-C. à 4^{e} s. apr.) et d'autre part celle des espacements entre établissements pérennes au cours d'une période donnée ainsi que celle des espacements entre anciens et nouveaux établissements. L'ensemble de ces mesures permettent d'identifier des processus de « remplissage » du territoire et d'en comparer le rythme et l'intensité. Dans le chapitre 3.3, il s'agit de comparer des mesures isotopiques alimentaires issues de vestiges animaux et botaniques recueillis dans des sites situés dans deux milieux différents, montagne et vallée, du territoire actuel de la Géorgie. L'enjeu y est d'identifier des différences éventuelles de ressources alimentaires à l'aide de tests non paramétriques.

Un troisième ensemble porte sur la comparaison de populations relevant de différentes catégories. Tel est par exemple le cas dans le chapitre 3.5 lorsqu'il s'agit de comparer l'âge au décès des hommes et des femmes pour quatre nécropoles en Normandie. Les auteurs identifient une surmortalité des hommes par rapport aux femmes. La méthode bayésienne permet même aux auteurs de préciser que la probabilité d'un âge moyen au décès plus faible pour les hommes est de 0.65, ce qui correspond clairement à un signe mais pas à une évidence forte. Dans le cas 2.2.7, il s'agit de comparer le degré moyen de centralité des nœuds correspondants à deux catégories de pôles définis en fonction des densités de haches qui ont pu y être répertoriées. Les auteurs mettent en évidence une densité supérieure dans les centres principaux relativement aux centres secondaires.

Dans un quatrième ensemble d'applications, il s'agit de comparer les caractéristiques d'une population donnée avec celles d'une population de référence. Dans le chapitre 3.6 il s'agit ainsi de comparer un jeu de données fossiles avec un échantillon de traces de dents de prédateurs modernes clairement identifiés (ours, lion, hyène) afin de déterminer quels sont les prédateurs qui ont opéré à des époques passées. Dans plusieurs cas il s'agit de comparer la signature (physico-chimique ou en termes de faciès) d'un ensemble de vestiges trouvé en un lieu donné avec celle correspondant à un lieu de référence. C'est le cas pour le chapitre 2.1.5 où est comparée la signature d'un métal utilisé dans la construction du Palais des Papes d'Avignon avec les signatures correspondant à différentes régions d'approvisionnement possibles (des sites de pro-

duction ou des sites miniers). Une combinaison de différentes techniques d'analyse des données est utilisée pour réaliser cette comparaison. Une ACP permet d'étudier comment se structurent les interrelations entre les éléments chimiques composant les métaux analysés et une CAH permet ensuite, avec certaines précautions, de définir des classes, chacune caractérisant une signature. Enfin est réalisée une analyse discriminante afin de déterminer si le fer utilisé à Avignon provient d'une des zones testées. Dans le chapitre 2.1.3 il s'agit de comparer la composition géochimique (notamment les teneurs en rubidium, strontium, zirconium, niobium et césium) d'un artefact en obsidienne avec celles d'échantillons géologiques correspondant à de possibles sources. Une méthode reposant sur une élimination étape par étape, à partir d'ACP et de comparaisons de log-ratios de différents éléments chimiques, permet d'identifier l'origine de l'obsidienne utilisée. Dans l'étude 2.1.4 portant sur la provenance de matières colorantes en Afrique du sud et en Namibie, la méthode consiste à réaliser une ACP sur des échantillons géologiques afin de construire un modèle type pour les sources possibles, et ensuite de projeter les échantillons archéologiques, en tant qu'individus supplémentaires, sur les axes de cette ACP. Cette démarche permet de repérer si les échantillons archéologiques ont la même signature géochimique que les échantillons géologiques ayant servi à structurer ces axes. De telles comparaisons à des référentiels permettent ainsi d'identifier des origines, qu'il s'agisse d'un type d'atelier de production, d'un type géologique, ou encore d'un atelier, d'une mine, d'un site géographique précis. Appliquant une démarche similaire pour rechercher la provenance de lingots de type peau de bœuf, les auteurs mettent en avant les précautions à prendre et soulignent l'intérêt de croiser les analyses (*cf.* 2.1.7).

4.2.3 Mettre en évidence, synthétiser, « expliquer »[5] les relations entre différents phénomènes

L'approche statistique permet d'étudier la distribution des variables auxquelles s'intéresse le chercheur, c'est-à-dire la manière dont se répartissent les valeurs faibles ou élevées des variables mobilisées. Les mesures de dispersion permettent d'avoir une première idée du degré de différenciation entre les différents individus statistiques composant l'échantillon soumis à l'analyse. Pour aller plus loin dans l'étude de la distribution statistique d'une variable, on peut émettre des hypothèses sur les facteurs accompagnant des valeurs plutôt plus faibles ou plus fortes, et tester l'existence et l'intensité d'éventuelles relations entre différentes variables, que celles-ci soient quantitatives ou qualitatives, que l'approche soit descriptive, explicative ou prédictive.

Dans le travail sur la diffusion de la hache de jade (*cf.* 2.2.7), le rôle de la dimension spatiale est ainsi étudié à deux niveaux. D'abord au niveau de la hache, où les auteurs s'intéressent à la relation entre le niveau de finition des haches et la distance entre le lieu de découverte (*i.e.* le lieu de consommation) et le lieu d'origine. Un test du chi2 montre que ces deux variables ne sont pas indépendantes, le niveau de finition est d'autant plus élevé que la hache a été trouvée loin de son lieu d'origine. Ensuite, au

[5] Les guillemets sont utilisés pour attirer l'attention sur le fait qu'il s'agit d'explication au sens statistique. Celle-ci repose sur l'expression d'une variable dite « à expliquer » en fonction d'un ensemble d'autres variables dites « explicatives ». Cette relation repose sur des covariations des variables qui n'impliquent pas nécessairement des causalités.

niveau des pôles, la quantité de haches trouvées est mise en relation avec la distance au Mont Viso, point fondateur pour les haches en jade. Enfin, à ce même niveau des pôles, formalisés cette fois comme nœuds dans un réseau, est interrogée la relation entre le score d'intermédiarité (nombre de fois où un nœud est un point de passage dans le plus court chemin entre deux autres nœuds) et le volume global des flux cumulés ayant transité par chaque nœud. Le coefficient de corrélation très positif permet de conclure à la forte relation entre cet indicateur de la théorie des graphes et ce volume de flux.

Dans le chapitre 2.2.5 l'enjeu est, dans un but prédictif, d'identifier la forme de la relation entre les caractéristiques physiques de l'espace et la densité de sites archéologiques qui s'y trouvent. Le test du chi2 a ainsi été utilisé pour mettre en évidence la relation entre les catégories taphonomiques (types de dépôts alluviaux) et les densités respectives de sites visibles et non visibles. Un modèle logistique a ensuite été développé pour exprimer la probabilité de trouver un site en fonction du contexte topographique (pentes combinées à l'orientation, texture du sol, par exemple) des différentes zones.

Dans le cas 1.2 l'enjeu consiste à mettre en relation des phases typo-technologiques qui illustrent différentes cultures paléolithiques et des variables paléoclimatiques qui rendent compte des conditions de l'environnement au cours de ces phases. La dimension chronologique étant essentielle dans ce questionnement, les auteurs utilisent l'approche bayésienne pour cette mise en relation.

Dans plusieurs des cas d'étude, des données multi-dimensionnelles sont mobilisées afin de construire, à partir de leur combinaison, des typologies. Celles-ci synthétisent un ensemble de relations caractéristiques entre les variables qui permettent de constituer des « signatures », ou des « empreintes » (« empreintes digitales » dans le cas 2.1.3), ou encore des « faciès » (*cf.* 1.3), ou simplement des « compositions » (*cf.* 2.1.6) suivant le vocabulaire utilisé dans les différents chapitres. Elles rendent effectivement compte de différentes compositions, qui résultent chacune d'un ensemble précis de relations entre un ensemble de variables (chimiques et/ou physiques dans la plupart des applications, ou bien de forme pour les faciès céramique). Comme exposé plus haut, les méthodes d'analyse des données sont idéales pour identifier de telles combinaisons. Dans le chapitre 2.1.5 par exemple, les individus statistiques sont ainsi des minerais et scories de réduction caractérisés par 18 variables correspondant à des teneurs chimiques (uranium, silicium etc). L'ACP montre quels composants sont corrélés entre eux, *i.e.* se trouvent réunis (combinés) au sein des mêmes individus statistiques. Une typologie en 7 classes a été obtenue par CAH, un type de minerai ayant pu être identifié pour chacune d'entre elles. La démarche est de même nature dans le chapitre 2.1.3 à propos des artefacts en obsidienne.

On peut également mettre en évidence les interrelations entre plusieurs variables à l'aide d'une analyse graphique matricielle, méthode graphique manuelle mise au point par le cartographe Jacques Bertin et consistant à permuter progressivement les lignes (individus) et les colonnes (variables) d'un tableau de données afin de rapprocher les individus ayant un profil similaire et en fournir une image interprétable visuellement. La sériation, permettant de mettre en évidence une succession sans ruptures,

particulièrement bien adaptée pour des interprétations chronologiques, fait l'objet du chapitre 1.6 où la complémentarité entre cette approche et l'analyse des données est mise en avant.

4.3 Les précautions nécessaires à l'application des méthodes statistiques

Les outils statistiques permettent ainsi de répondre à une multitude de questions que se pose l'archéologue en prenant appui sur les données empiriques dont il dispose. Chacune des méthodes mobilisées a cependant ses limites propres et si des calculs peuvent toujours être effectués (il suffit d'appliquer des formules), les résultats de ceux-ci ne prennent sens que si les conditions d'application sont respectées. Les précautions à prendre sont ainsi très régulièrement soulignées dans les études de cas, quelles que soient les questions traitées, et les limites des méthodes statistiques sont discutées. Ces limites concernent d'une part les individus statistiques qui peuvent être trop peu nombreux ou pas assez représentatifs, et d'autre part les variables, si les mesures sont incertaines, ou encore si leur distribution statistique ne correspond pas aux conditions d'application de la méthode statistique envisagée.

Le chapitre 3.2 traite ainsi de la représentativité d'un échantillon et donne une définition claire de ce qu'est la population mère, indispensable pour réaliser une inférence statistique correcte. L'auteure passe en revue l'ensemble des biais possibles en prenant appui sur le domaine des échantillons anthropologiques où il s'agit, à partir de l'analyse d'ossements, de déduire des informations sur les populations passées. Plus généralement, quels que soient les objets considérés, il est extrêmement rare que les vestiges obtenus lors des fouilles permettent de constituer un échantillon tiré au hasard dans la population mère considérée. Or c'est une condition essentielle pour que les résultats trouvés sur l'échantillon puissent être inférés de manière satisfaisante à la population mère. Des biais peuvent provenir du fait que le phénomène étudié lui-même est hétérogène dans l'espace, mais aussi du fait que les fouilles ne sont pas toujours exhaustives et peuvent passer à côté d'une portion d'espace renfermant justement un type d'objets particulier. Le risque existe donc toujours de sur- ou sous-évaluer certaines catégories d'objets.

La taille de l'échantillon représente également une contrainte. Si le nombre d'individus est trop faible, de nombreuses méthodes statistiques parmi les plus usuelles ne sont pas adaptées et il est alors essentiel de choisir une approche et une méthode qui le soient. Tel est par exemple le cas lorsque l'on souhaite procéder à des comparaisons. Les tests usuels de comparaison des moyennes par exemple nécessitent des distributions normales. Or, dans les cas où les effectifs sont faibles on ne peut tester la normalité correctement. Dans le chapitre 3.3, où l'enjeu consiste à comparer des échantillons situés dans deux milieux physiques différents, des tests non paramétriques adaptés à la comparaison de moyennes d'échantillons de faibles effectifs sont donc mobilisés. Dans le cas 3.6 où se pose également le problème de très petits échantillons, une méthode adaptée est proposée et testée. Elle consiste à appliquer une ACP et une CAH sur des indicateurs VSI* (*Variability Size Index*) mesurant l'écart entre des mesures morphométriques effectuées sur des échantillons fossiles avec celles que l'on obtient dans une population de référence. Dans le chapitre 3.4 enfin, les auteurs discutent de la manière d'intégrer des données isotopiques sur de faibles effectifs dans une approche

bayésienne. Dans le cas 3.5 est évoqué le double problème que représente la faiblesse à la fois des échantillons étudiés et des collections de référence. Les auteurs montrent cependant que la nature même de ce qu'ils cherchent à estimer, une structure par âge, est parfaitement adaptée à une approche bayésienne dans ce cas. Ces quelques exemples illustrent que la question des petits effectifs peut être résolue dans des contextes méthodologiques divers.

Une autre source possible de biais réside dans les mesures qui sont effectuées sur les objets. Les différentes méthodes chimiques et physiques mobilisées pour établir la signature géochimique d'un objet ou d'un gisement peuvent ainsi être source d'erreur. Dans le chapitre 1.4 est ainsi développée une démarche pour déterminer les incertitudes lors de la datation par luminescence des matériaux utilisés dans le bâti. Deux types d'incertitude y sont distingués, l'une statistique portant sur les imprécisions des mesures faites sur un échantillon donné pour déterminer l'âge de l'objet étudié, et l'autre qualifiée de systématique, liée au système même de mesure et touchant donc toute une série d'échantillons. Une fois les mesures effectuées il y a plusieurs étapes essentielles de nettoyage et de transformation des données à réaliser (*cf.* 2.1.2 qui les passe en revue). Le cas 2.1.4 donne lieu à deux étapes de normalisation des variables représentées par des teneurs en différents éléments chimiques présents dans les gîtes étudiés : 1- une normalisation classique qui consiste à ramener les mesures dans un même ordre de grandeur ; et 2- une normalisation par la teneur en oxyde de fer afin de tenir compte de l'hétérogénéité de la distribution de ce dernier dans l'espace étudié. Dans d'autres cas enfin, les données peuvent simplement être incomplètes, les informations manquant pour certains individus. La possibilité d'utiliser alors le krigeage est évoquée dans le cas 2.2.6.

Les auteur.e.s des différents cas d'étude reviennent régulièrement sur l'importance d'adopter une lecture critique des résultats des analyses statistiques. Au-delà des divers biais évoqués ci-dessus, liés à la constitution de l'échantillon d'individus statistiques ou à la mesure des variables, il peut en effet y avoir un décalage entre la logique statistique et la nature des données dont on dispose. Analysant des céramiques relevant de différentes périodes et catégories provenant des fouilles d'Ephèse, l'auteure du chapitre 2.1.6 préconise ainsi que « les options statistiques choisies permettent de rester suffisamment proche des données initiales ». Dans le cas étudié, les différents sous-groupes de céramiques ont en effet leur « logique géochimique » propre qui n'est pas rendu visible par le découpage en classes fourni par la CAH. En revanche, une interprétation des coupures en classes de la CAH au niveau de chacun des sous-groupes fait sens. Un aller-retour constant entre les résultats des analyses statistiques, les données initiales et le cadre conceptuel permet en général d'éviter les problèmes dus à un léger écart relativement aux conditions d'application des outils statistiques.

4.4 Dialoguer, circuler entre observables, concepts et modélisation

Pour terminer cette introduction, revenons à la figure 1 qui illustre justement le dialogue entre les domaines conceptuel, empirique et de la modélisation. Ce dialogue se fait pas à pas, au fur et à mesure que l'on avance dans la recherche. Il se renoue à chaque étape des traitements statistiques : l'interprétation des résultats conduit à un retour sur les données, qui conduit à son tour à un enrichissement des concepts

et ouvre vers la construction de nouveaux objets. Ces derniers, construits à partir des données empiriques et du cadre conceptuel mis au point, sont de plus en plus chargés de sens : la combinaison de diverses informations et traitements permet ainsi de construire des établissements, des pôles, des logs stratigraphiques, des courbes de densité, qui peuvent à leur tour être soumis à de nouveaux traitements statistiques dont les résultats permettent d'enrichir le domaine conceptuel et de mettre en place de nouvelles hypothèses à explorer, et ainsi de suite. L'exposé des cas d'étude présentés dans les trois premiers chapitres et les divers appels à précautions qui les accompagnent, laissent imaginer que ce dialogue ne se fait pas à plat et qu'à chaque « tour », les chercheurs le posent à un niveau plus élevé. La circulation entre les trois domaines de la figure 1 correspondrait ainsi à une spirale qui s'élève progressivement (Mathian *et al.* 2014). Un des intérêts de cet ouvrage est de présenter une grande diversité d'objets, non seulement sur le plan thématique, mais aussi sur un plan ontologique, tant au niveau des objets eux-mêmes (fragments, traces, objets construits) que des propriétés appréhendées à partir de mesures diverses. Le choix d'entrer dans cet ouvrage de statistiques à partir d'études de cas sur des thématiques aussi diverses contribue à mettre en évidence le caractère universel de l'approche statistique. Ce parti pris est intéressant d'un point de vue didactique car il permet aux chercheurs de découvrir de nouvelles voies méthodologiques à travers le transfert de chaînes de traitement statistique mis au point pour étudier une question précise relative à une période et un lieu donnés à des questions portant sur d'autres objets correspondant à des échelles de temps et d'espace très différents.