CHAPITRE 1-1
L'APPROCHE QUANTITATIVE ET LA MESURE

Plan

CHAPITRE 1-1
L'APPROCHE QUANTITATIVE ET LA MESURE

Références : Gilles (1994, Introduction et chap. 1 et 2) ; Bryman et Cramer, 1990, p. 61-74 ; Blalock (1972, chap. 2) ; Lazarsfeld (1971)

Quantitatif s'oppose à *qualitatif*. Non pas que les deux approches soient mutuellement exclusives : elles seraient plutôt complémentaires (sur les débats idéologiques et méthodologiques à propos des approches qualitative et quantitative, voir Gilles, 1994, Introduction, p. 1-9). Mais les deux termes s'opposent quant à leur définition : est quantitatif ce qui se mesure. Plus exactement, la quantité est la propriété de ce qui peut être *mesuré* ou compté, de ce qui est susceptible d'accroissement ou de diminution.

Mais que vient faire la mesure dans la démarche scientifique en sciences sociales ? Et puis, qu'est-ce que *mesurer* ?

1-1.1 L'opérationnalisation des concepts : indicateurs, mesure et variables

En sciences – et cela est aussi vrai en sciences sociales – les théories et les hypothèses sont formulées au moyen de concepts et de relations entre des concepts. Un concept est une idée, une représentation mentale abstraite et générale d'un être, d'une manière d'être ou d'un rapport : c'est, en somme, un atome de pensée. Gilles (1994, p15) met l'accent sur l'opération qui crée le concept en explicitant sa compréhension et en fixant son extension [1] : pour lui, un concept est une « construction de la pensée résultant d'une opération par laquelle on individualise des traits permettant de rapprocher des objets différents ou de distinguer des objets autrement similaires », ou, autrement dit, par laquelle on définit les critères permettant de déterminer si tel ou tel objet fait partie ou non de l'extension du concept.

---

[1] L'*extension* logique est l'ensemble des objets concrets ou abstraits auxquels s'appliquent un concept, une proposition (ensemble des cas où elle est vraie) ou une relation (ensemble des systèmes qui la vérifient). L'extension d'un concept s'oppose à la *compréhension*, qui est l'ensemble des caractères qui appartiennent à un concept. Par exemple, le concept *homme* a une moindre extension, mais une plus grande compréhension que *mammifère*.

Exemple :

- « La consommation des ménages croît avec le revenu » ;
  cette proposition contient les concepts « consommation des ménages », « revenu » et le lien entre les deux est exprimé par les mots « croît avec ».

Pour rapprocher les propositions théoriques de la réalité, ou pour confronter les hypothèses à l'observation, il faut *opérationnaliser* les concepts, c'est-à-dire établir une relation systématique entre les concepts et la réalité observable, au moyen d'*indicateurs*. On peut définir les indicateurs comme des « signes, comportements ou réactions directement observables par lesquels on repère au niveau de la réalité les dimensions d'un concept » (Gilles, 1994, p. 27). Les dimensions sont les différentes composantes d'un concept (Gilles, 1994, p. 24) : nous reviendrons plus loin sur cette notion de dimension.

Opérationnaliser un concept, c'est donc lui associer un ou plusieurs *indicateurs* qui permettront de distinguer avec exactitude les variations observées dans la réalité par rapport au concept. Distinguer les variations, cela veut dire *mesurer* : l'opérationnalisation d'un concept conduit donc à la *mesure*.

Mentionnons que ce lien entre l'opérationnalisation et la mesure existe aussi bien dans l'approche qualitative que dans l'approche quantitative. Car, même dans l'approche qualitative, il faut bien classifier et compter les sujets, ce qui constitue une opération de mesure, comme nous le verrons plus loin. À cet égard, Gilles écrit : « D'une manière générale, les méthodes dites qualitatives (histoire de vie, analyse de récit, observation participante, entrevue en profondeur, étude de cas...) font, elles aussi, usage de la statistique à des fins descriptives » (1994, p. 3) ; il distingue : « Méthodes qualitatives et données qualitatives ne procèdent pas de la même logique. Les premières reposent sur une conception humaniste, herméneutique ou interprétative des sciences sociales qui, dans cette perspective, deviennent sciences humaines. Quant aux secondes, elles peuvent être prises dans le sens de données ne permettant que l'utilisation de certaines techniques statistiques dites "robustes" » (1994, p. 3, note 5). Nous verrons d'ailleurs sous peu qu'en un certain sens, on peut mesurer des propriétés qualitatives : c'est pourquoi il n'est pas absurde de parler de l'analyse quantitative de données qualitatives (Quatrième partie de l'ouvrage).

Tout cela est résumé par Gilles (1994, p. 24), qui se réfère au schéma classique de Lazarsfeld (1971) : opérationnaliser, c'est « soumettre les concepts, par l'analyse, à un processus qui les

transforme en dimensions, puis en indicateurs permettant de les observer, de les mesurer ou de les quantifier ».

Exemple :

- Pour opérationnaliser le concept « consommation des ménages », c'est-à-dire pour mesurer la consommation d'un ménage, on peut utiliser le montant déclaré, dans une enquête auprès des ménages, en réponse à une question comme
  « La semaine dernière, combien ont dépensé l'ensemble des personnes qui composent le ménage ? »
- Mais on pourrait aussi mesurer la consommation d'un ménage par calcul, en soustrayant de ses revenus le montant d'impôt sur le revenu qu'il a payé et la somme qu'il a épargné durant une année donnée.

En général, un concept peut se traduire par plusieurs indicateurs. Le choix des indicateurs est très important en recherche. Les indicateurs retenus doivent être valides et fiables (on dit aussi fidèles).

- Un indicateur est *valide* lorsqu'il mesure bien ce que l'on veut mesurer, c'est-à-dire lorsqu'il reflète les variations relatives au concept même qu'il est censé représenter. Pour examiner la validité d'un indicateur, il faut évidemment qu'au préalable le concept ait été clairement défini.
- Un indicateur est *fiable* ou *fidèle* lorsque les variations dans la mesure correspondent à des variations véritables.

Exemple :

- Le montant dépensé la semaine dernière n'est peut-être pas une mesure valide de la consommation, parce que ce montant inclut peut-être des dépenses en capital (investissement résidentiel), alors que la définition du concept théorique « consommation » exclut le coût d'acquisition de biens durables.
- La réponse à propos du montant dépensé la semaine dernière n'est peut-être pas fiable, parce que la personne qui répond au questionnaire ne sait peut-être pas ce qu'ont dépensé les autres membres du ménage.

Le résultat de l'application d'un indicateur à un ensemble d'objets est une *variable*. Une variable est donc définie par la chose à mesurer (le concept et ses dimensions), par la façon de la mesurer (l'indicateur) et par son domaine d'application (les objets auxquels s'applique la mesure).

Soulignons enfin la distinction qu'il faut faire entre une *variable* et les différentes *valeurs* qu'elle peut prendre.

Exemple :

- Dans une enquête auprès d'un échantillon de ménages, la réponse à la question
  « La semaine dernière, combien ont dépensé l'ensemble des personnes qui composent le ménage ? »
  est une variable qui prend une valeur différente pour chacun des ménages de l'enquête.

1-1.2 Qu'est ce que la mesure ?

Mesurer, c'est comparer. Mais encore ? Dans la langue courante, on définit la mesure comme « l'évaluation d'une grandeur par comparaison avec une autre de la même espèce prise pour unité » (Dictionnaire Larousse de la langue française, cédérom, 1996). Nous verrons dans un moment que cette définition est très étroite. *Encyclopaedia Britannica* (cédérom, 1998) propose : « measurement : the process of associating numbers with physical quantities and phenomena ». Dans le même esprit, Gilles (1994, p. 34) écrit : « Mesurer, c'est établir une correspondance entre l'ensemble que constitue le phénomène à mesurer et un ensemble de nombres que l'on choisit en fonction de la nature du phénomène ». Ces deux dernières définitions, plus larges, sont cependant incomplètes tant que l'on ne spécifie pas quelles sont les conditions que doit remplir une correspondance numérique pour constituer une mesure. C'est l'objet de la *théorie de la mesure*.

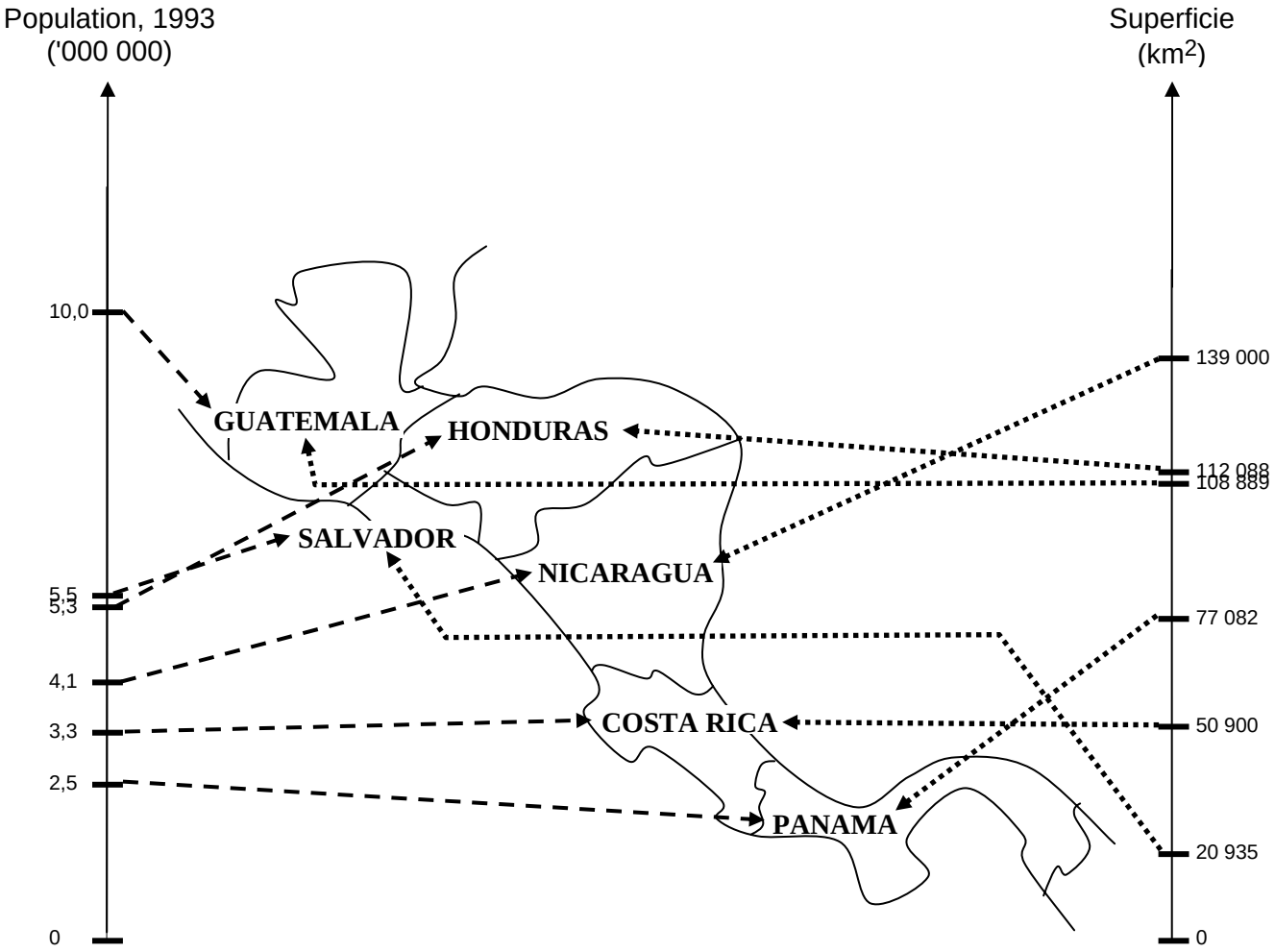
Pour nos fins, nous retiendrons ceci [2] : une correspondance constitue une mesure si elle permet de *comparer* deux *objets quelconques* par rapport à une *propriété donnée*.

Par exemple, supposons que la propriété à mesurer soit la superficie. Les pays, les chambres à coucher et les mouchoirs de poche sont des objets pour lesquels la propriété « superficie » est définie. Pour qu'elle puisse constituer une mesure de superficie, une correspondance doit permettre de comparer quant à leur superficie deux pays, ou deux mouchoirs de poche, ou même un pays et un mouchoir de poche. Mais qu'est-ce que *comparer* ? Dans le contexte de la mesure, comparer, c'est déterminer si, par rapport à une propriété donnée, les deux objets comparés sont semblables ou non, et, s'ils ne sont pas semblables, lequel possède la propriété mesurée à un degré plus grand que l'autre.

---

2 Ce qui suit est inspiré de Taylor, 1977, chap. 2, p. 38-41, mais la notation utilisée ici est différente.

UNE MESURE EST UNE CORRESPONDANCE...



Source des données : Faculdad Latino Americana de Ciencias Sociales
FLACSO, Sede Costa Rica, San José, Costa Rica

On peut formaliser ce qui précède ainsi. Désignons par *A* et *B* deux objets quelconques (deux pays, par exemple) qui possèdent la propriété à mesurer (la superficie, par exemple). Puisqu'une mesure associe un nombre à chaque objet, un peu comme une fonction mathématique, il est naturel de représenter la mesure de la même façon : convenons alors que *f(A)* est le chiffre de la superficie de *A* ; de même, *f(B)* est le chiffre de la superficie de *B*. La comparaison examine les relations suivantes :

$f(A) = f(B)$

$f(A) \neq f(B)$

$f(A) < f(B)$

$f(A) > f(B)$

Une *mesure* est une correspondance qui permet, pour au moins l'une des relations qui précèdent, de déterminer si elle est vraie ou fausse. Dans le cas de la superficie, on peut établir une correspondance entre chaque pays et le nombre de kilomètres carrés qui sont compris à l'intérieur de ses frontières, ou entre chaque mouchoir de poche et la fraction de kilomètre carré qu'il recouvre. Lorsqu'on compare les chiffres donnés par cette correspondance, on peut décider s'il est vrai que $f(A) = f(B)$ (*A* et *B* ont même superficie), ou $f(A) \neq f(B)$ (*A* et *B* n'ont pas même superficie), auquel cas, ou bien $f(A) < f(B)$ (*A* est plus petit que *B*), ou bien $f(A) > f(B)$ (*A* est plus grand que *B*).

Dans l'exemple de la superficie, la mesure permet de déteminer la valeur de vérité (vraie ou fausse) de *chacune* des quatre relations $=$, $\neq$, $<$ et $>$. Mais la définition de la mesure n'exige pas que l'on puisse déterminer la valeur de vérité des *quatre* relations. Par exemple, supposons que la propriété examinée soit la nationalité. On pourrait définir la correspondance suivante :

$f(X) = 0$ si la personne *X* est de nationalité costaricaine ;

$f(X) = 1$ si la personne *X* est d'une autre nationalité centraméricaine ;

$f(X) = 2$ dans tous les autres cas.

Alors

$f(A) = f(B)$ signifie que la personne *A* et la personne *B* sont de même nationalité (dans la classification retenue) ;

$f(A) \neq f(B)$ signifie que la personne *A* et la personne *B* ne sont pas de même nationalité.

Par contre, les relations $f(A) < f(B)$ et $f(A) > f(B)$ n'ont aucune signification. La correspondance constitue néanmoins une mesure au sens large : c'est une mesure de la nationalité. En un certain sens, donc, on peut mesurer des propriétés qualitatives.

> Note : Les valeurs numériques de la correspondance n'ont aucune signification et elles sont parfaitement arbitraires. On pourrait même définir la correspondance en termes de symboles autres que des nombres. Par exemple, on aurait pu définir
>
> $f(X) =$ 'CR' si la personne X est de nationalité costaricaine ;
>
> $f(X) =$ 'CA' si la personne est d'une autre nationalité centraméricaine ;
>
> $f(X) =$ 'OT' dans tous les autres cas.

1-1.3 Échelles de mesure et types de variables

Même si l'on admet qu'une propriété qualitative comme la nationalité d'une personne peut être mesurée, il n'en demeure pas moins que la mesure d'une telle propriété semble imparfaite, en comparaison de la mesure de propriétés comme la superficie ou le revenu. En effet, pour une propriété comme la nationalité, on ne peut pas décider s'il est vrai ou faux que $f(A) < f(B)$ ou $f(A) > f(B)$: cela n'aurait pas de signification. Par contre, pour la superficie d'un territoire ou le revenu d'un ménage, on peut déterminer s'il est vrai ou faux que $f(A) < f(B)$ ou $f(A) > f(B)$: la mesure est plus complète.

C'est pourquoi on distingue plusieurs types de variables, selon l'échelle de mesure qui leur est associée [3] :

1. Variables *catégoriques*
2. Variables *ordinales*
3. Variables *d'intervalle*
4. Variables *rationnelles*

Variables catégoriques

Les variables catégoriques (« nominal » en anglais) résultent de l'application d'une échelle de mesure qui ne permet de décider que des relations $=$ et $\neq$. La valeur que prend une variable catégorique s'appelle « modalité » : elle indique à quelle catégorie appartient l'individu auquel

---

[3] On trouve une classification similaire chez Legendre et Legendre (1998, p. 28 et suivantes).

elle se rapporte : une variable catégorique permet donc de classer les individus en groupes. On distingue

- Variables *dichotomiques* : 2 catégories possibles ;
- Variables *polytomiques* : plus de 2 catégories.

Exemples :

- Sexe (homme/femme) : variable catégorique dichotomique ;
- Nationalité : variable catégorique polytomique (lorsque l'on distingue plus de deux nationalités).

On peut remplacer une variable polytomique par plusieurs variables dichotomiques. D'ailleurs, certaines méthodes d'analyse l'exigent. Par exemple, considérons une variable polytomique de nationalité :

$NAT = 0$ si la personne X est de nationalité costaricaine ;

$NAT = 1$ si la personne est d'une autre nationalité centraméricaine ;

$NAT = 2$ dans tous les autres cas.

On peut remplacer cette variable par deux variables dichotomiques, comme

$COR = 1$ si la personne est citoyenne du Costa Rica et $COR = 0$ autrement

$CAM = 1$ si la personne est citoyenne d'un pays d'Amérique Centrale autre que le Costa Rica et $CAM = 0$ autrement.

Question : pourquoi seulement deux variables dichotomiques, alors que la variable polytomique peut prendre trois valeurs ?

Variables ordinales

Les variables ordinales résultent de l'application d'une échelle de mesure qui permet de décider que de chacune des quatre relations $=$, $\neq$, $<$ et $>$. Les valeurs que prend une variable ordinale pour différents individus permettent donc de ranger les individus en ordre croissant ou décroissant par rapport à la propriété mesurée. On distingue les ordres *faibles* ou *réduits* – incomplets, par classes d'équivalence – et les ordres *complets*.

Exemples :

- Nombre de points obtenus à un test d'aptitudes (ordre complet : si deux sujets obtiennent le même nombre de points, la mesure indique qu'ils possèdent le même degré d'aptitude selon ce test) ;

- Variable définie par : 1 si l'étudiant réussit un examen donné et 0 s'il échoue (ordre faible : si deux étudiants ont réussi, ça ne veut pas dire qu'ils sont d'égale force).

Les mesures ordinales sont définies « à une transformation monotone croissante près », c'est-à-dire que l'on ne change pas la mesure si l'on applique à la variable une transformation mathématique, pourvu que l'on ne change pas l'ordre numérique des valeurs. Par exemple, on pourrait remplacer le nombre de points obtenus par le logarithme du nombre de points, ou par le carré du nombre de points, ou on pourrait ajouter un million de points à tous les sujets.

Variables d'intervalle

Les variables d'intervalle sont similaires aux variables ordinales, mais en plus de permettre de ranger les individus en ordre croissant ou décroissant, elles permettent de comparer les *différences* entre individus.

Exemples :

- La température : s'il fait $-15°$ C à Montréal, $+24°$ C à San José (Costa Rica) et $+18°$ C à Miami, on peut dire que la différence de température est moins grande entre San José et Miami (6° C) qu'entre Miami et Montréal (33° C). Avec une variable ordinale, ce genre de comparaison n'a pas de sens.

Formellement, les variables d'intervalle résultent de l'application d'une échelle de mesure où les *différences* entre les valeurs sont aussi des mesures, ordinales : l'échelle de mesure permet de déterminer la valeur de vérité de *chacune* des relations suivantes

$$f(A) - f(B) = f(C) - f(D)$$

$$f(A) - f(B) \neq f(C) - f(D)$$

$$f(A) - f(B) < f(C) - f(D)$$

$$f(A) - f(B) > f(C) - f(D)$$

Avec une variable d'intervalle, le zéro de l'échelle de mesure est arbitraire mais les transformations de l'échelle doivent préserver la comparaison entre les écarts. C'est pourquoi les échelles d'intervalle sont définies « à une transformation linéaire près ». Par exemple, on passe de l'échelle Celsius à l'échelle Fahrenheit au moyen de la transformation linéaire

$$F = 32 + 1{,}8 \times C$$

Autre exemple, en géographie, la direction est donné en degrés, calculés dans le sens des aiguilles d'une montre à partir de la direction nord. Le zéro (franc nord) est arbitraire.

Variables rationnelles

Les variables rationnelles (aussi appelées *proportionnelles*) sont similaires aux variables d'intervalle, sauf qu'avec les variables rationnelles, il existe un zéro naturel [4]. Il en découle que le rapport entre deux valeurs a un sens (rationnel vient de « ratio », rapport).

Exemple :

- Le revenu est une variable rationnelle. Si une personne gagne 60 000 $, on peut dire qu'elle gagne deux fois plus qu'une personne qui gagne 30 000 $. Par contre, il serait absurde de prétendre qu'il fait deux fois plus chaud à 20° C qu'à 10° C (20° C = 68° F et 10° C = 50° F).

Formellement, les variables rationnelles résultent de l'application d'une échelle de mesure où les *rapports* entre les valeurs sont aussi des mesures, ordinales : l'échelle de mesure permet de déterminer la valeur de vérité de *chacune* des relations suivantes

$$\frac{f(A)}{f(B)} = \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} \neq \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} < \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} > \frac{f(C)}{f(D)}$$

Si l'on revient à la définition de la mesure selon Larousse comme « l'évaluation d'une grandeur par comparaison avec une autre de la même espèce prise pour unité », on constate que cette définition ne s'applique en vérité qu'aux échelles de mesure rationnelles. La définition du Larousse est donc restrictive.

Il arrive souvent que les valeurs observées de variables rationnelles ou d'intervalle soient regroupées en classes. Par exemple, une variable « revenu » pourrait prendre la forme suivante :

REV = 1 si revenu < 10 000 $

REV = 2 si 10 000 $ ≤ revenu < 25 000 $

---

[4] Pour cette raison, la température mesurée en degrés Kelvin est une variable rationnelle, puisqu'il existe un zéro naturel, le « zéro absolu ». Le zéro absolu, qui équivaut à environ −273,16 à l'échelle Celsius, est la température

REV = 3 si 25 000 $ ≤ revenu < 50 000 $

REV = 4 si revenu ≥ 50 000 $

Une variable de ce type est une variable ordinale qui définit un ordre faible. Le fait de regrouper les valeurs observées en classes a donc pour effet de transformer une variable rationnelle (ou d'intervalle) en variable ordinale d'ordre faible. On passe ainsi à une échelle de mesure plus « primitive » et on perd de l'information. Il est donc préférable, lorsque c'est possible, d'utiliser les données sous leur forme originale.

Échelles de mesure et méthodes quantitatives

Il existe des méthodes d'analyse quantitative adaptées à tous les types de variables. On peut donc appliquer des méthodes *quantitatives* à l'analyse de données *qualitatives*, lorsque celles-ci peuvent être mesurées au moyen de variables catégoriques ou ordinales.

1-1.4 Types de données

Cette partie du cours porte sur les méthodes quantitatives d'analyse des données. Mais la qualité de l'analyse dépend d'abord et avant tout de la qualité des données analysées. Les données ne sont jamais « parfaites » et l'analyste compétent doit adapter ses méthodes à la qualité des données qu'il doit traiter.

On peut distinguer trois types de données :

- les données primaires
- les données secondaires non publiées
- les données secondaires publiées

Il y a des problèmes de qualité spécifiques à chaque type de données.

1. Données primaires (enquêtes)

Le contrôle de la qualité doit se faire à toutes les étapes :

- préparation des instruments de cueillette des données (questionnaires)
- cueillette
- codification

---

la plus basse qui soit théoriquement possible, caractérisée par l'absence totale de chaleur : à 0° K, l'activité des molécules est inexistante.

- saisie, validation, correction et organisation
- évaluation *ex post* de la qualité

2. Données secondaires non publiées

Ces données sont souvent collectées pour des fins, administratives ou autres, différentes de la recherche (les rôles d'évaluation foncière, par exemple) : les concepts sont souvent mal définis ou différents de ceux qu'on cherche à mesurer (les variables formées à partir de ces données ne sont qu'imparfaitement valides).

Le contrôle de la qualité des données secondaires non publiées pose souvent des problèmes proches de ceux qui se posent lorsqu'il s'agit de données primaires. Cependant, dans le cas de données secondaires, l'analyste ne peut pas veiller lui-même au contrôle de la qualité à toutes les étapes.

3. Données secondaires publiées

L'utilisation judicieuse de données secondaires publiées exige que l'on tienne compte de toute l'information pertinente qui accompagne les données (métadonnées) [5] :

- définitions et concepts
- méthodes de cueillette et de compilation
- évaluation de la qualité par l'émetteur
- crédibilité des sources

En plus de dépendre de la qualité des données, la qualité de l'analyse peut être compromise par des erreurs dans le traitement préalable qu'on fait subir aux données avant de leur appliquer les méthodes d'analyse.

Exemples :

- Erreur de variable lors de l'extraction de données d'une banque de données (masse salariale au lieu du salaire horaire).
- Erreur de programmation lors de l'appariement de deux fichiers (« merge ») : dédoublement de certaines observations.
- Erreur de formule dans un tableur (adresses relatives ou absolues...) ; ces erreurs résultent souvent d'opérations de « copier-coller ».

---

5 Atkinson et Brandolini (2001) examinent les avantages et les pièges des données secondaires dans le contexte de l'analyse des inégalités de revenu dans les pays de l'OCDE.

Si vous ne trouvez pas d'erreur dans les données,
c'est parce que vous ne cherchez pas bien ...

1-1.5 Structure matricielle fondamentale des données

Pour être utilisables, les données doivent d'abord être organisées de façon à ce que l'on sache à quoi réfère chaque nombre. Il y a plusieurs manières d'organiser les données, mais toutes découlent de la structure fondamentale des données. Cette structure fondamentale est celle d'une matrice, ou d'un tableau où, par convention,

- les *colonnes* correspondent habituellement à différentes *variables* (caractéristiques, propriétés, attributs, indicateurs, descripteurs, ...) ;
- les *lignes* correspondent habituellement à différentes *observations* (cas, individus, objets).

Il arrive que les observations se rapportent à des moments ou à des périodes successives : on est alors en présence de *séries chronologiques* ou *temporelles*. On a une situation analogue lorsque les observations se rapportent aux différents lieux d'un ensemble géographique donné (pays d'un continent, villes ou régions d'un pays, quartiers d'une ville, zones...) : on pourrait parler alors de *séries spatiales*. Les données spatiales ne sont pas toujours des *séries complètes*, qui comportent une observation, et une seule, pour chacun des lieux que l'on distingue dans un espace donné. Qu'elles constituent ou non des séries complètes, on dit que des données sont *géoréférencées* lorsqu'elles contiennent une ou plusieurs variables qui permettent de situer chaque observation dans l'espace géographique.

La structure matricielle fondamentale se généralise à plus de deux dimensions[6] quand certaines des variables sont catégoriques (variables de classification). Par exemple, supposons que l'on ait réalisé une enquête auprès d'un échantillon de personnes et que, parmi les variables pour lesquelles on a recueilli des données se trouve la profession du répondant. Dans ce cas, le reste des données peut être organisé en plusieurs tableaux à deux dimensions, un par profession. Si l'on superpose ces tableaux, on peut concevoir l'organisation des données comme un cube dont les couches successives correspondent aux différentes professions, les lignes à différents répondants, et les colonnes aux différentes autres variables. Naturellement, avec plus d'une variable catégorique, on peut imaginer un « hypercube » de données à quatre

6 Attention : le mot « dimensions » est utilisé dans un contexte différent pour désigner les dimensions d'un concept.

dimensions ou plus. La représentation mentale la plus appropriée dépend des analyses que l'on veut faire.

Ajoutons que la distinction entre observations et variables n'est pas étanche. Il arrive que les observations et les variables soient interchangeables, notamment lorsque les observations correspondent aux différentes catégories d'une variable de classification, tandis que les variables sont des attributs se rapportant aux différentes catégories d'une autre variable de classification [7]. Tel est le cas, par exemple, d'un tableau du nombre d'emplois par branche d'activité et par ville dans une région donnée. Dans ces conditions, on peut considérer

- soit que chaque observation correspond à une ville, et que les variables sont les nombres d'emplois des différentes branches d'activité dans cette ville ;
- soit que chaque observation correspond à une branche d'activité, et que les variables sont les nombres d'emplois de cette branche dans les différentes villes.

Là encore, la représentation mentale que l'on privilégie dépend des analyses que l'on veut faire.

Revenons au modèle d'organisation élémentaire, celui d'un tableau à deux dimensions. On distingue deux points de vue, ou modes d'analyse, selon que l'on s'attache aux relations entre les observations ou aux relations entre les variables (Jayet, 1993, p. 1-2 ; Legendre et Legendre, 1998, p. 248, reprennent une terminologie de Cattell, 1952, et distinguent l'analyse « en mode R », qui est l'analyse des relations entre les descripteurs, et l'analyse « en mode Q », qui est l'analyse des relations entre les objets). Cette distinction permet de classifier les types d'analyses et les méthodes qui leur sont associées (voir le tableau qui suit).

---

7 Comme le montre l'exemple donné dans les lignes qui suivent, une telle ambivalence est généralement le fait de données qui ont fait l'objet d'un premier traitement et qui sont constituées en tableau de contingence ou en tableau d'analyse de variance.

Point de vue « horizontal » : entre les variables

- Combiner plusieurs variables en une seule, qui les résume : construction de nombres indices
- Comparer deux variables : mesure de la similarité/dissimilarité
- Étudier les relations de dépendance
  - entre deux variables : corrélation, régression simple
  - entre une variable dépendante et plusieurs variables indépendantes : régression multiple et autres méthodes multivariées comportant une variable dépendante
  - entre plusieurs variables parmi lesquelles on ne distingue pas de variable dépendante : méthodes multivariées

Point de vue « vertical » : entre les observations ou objets

- Caractériser la distribution d'une variable : mesure de l'inégalité ou de la concentration, méthodes statistiques univariées
- Lorsqu'il existe un ordre naturel entre les observations, étudier les relations entre les différentes observations d'une même variable : mesure et modélisation de l'évolution des séries temporelles, analyse de l'autocorrélation (temporelle, spatiale)
- Comparer deux objets : mesure de la similarité/dissimilarité