

**Analyse de données d'enquêtes sur ordinateur :  
Organisation d'un fichier pour des données à niveaux multiples**

*par Chris Wolf, Service Informatique, Département d'Agro-Économie  
Michigan State University*

Les enquêtes socio-économiques de grande envergure demandent beaucoup de temps pour être mises en place, menées et analysées. L'effort demandé pour analyser de tels ensembles complexes de données est souvent plus grand qu'il ne devait être à cause de l'inexpérience du chercheur pour ce genre de travail. Peut être l'aspect le moins compris est la gestion des données à niveaux multiples dans la même enquête. Les chercheurs seront plus productifs si les différents niveaux de données sont identifiés à temps dans le projet, et s'ils sont gérés proprement à travers les différents stades de l'analyse.

Cette discussion mettra premièrement l'accent sur l'organisation des fichiers de données dans l'ordinateur et comment elle est liée à l'élaboration et l'analyse de l'enquête. Le principal point est l'identification des différents niveaux de données qui sont collectées et analysées dans l'enquête. Un sujet secondaire qui est fortement lié au premier est l'usage de variables clés dans les fichiers de données.

**LE CONCEPT DE FICHER DE DONNÉES**

Un fichier de données pour l'analyse d'une enquête est composé d'observations, de variables, et de valeurs. Ceci est facilement perçu comme une matrice, où chaque rangée contient toutes les données pour une observation, et chaque colonne contient toutes les données pour une variable particulière (ou question dans l'enquête) sous forme de tableau. Le fichier ressemble à ceci :

Observation	Nom de la variable									
	VILLAGE	HH	Q1A	Q1B	Q2	Q3	Q4A	Q4B	Q4C	
1	1	1	10	23	2	7	1	3	1	
2	1	2	3	34	2	3	2	2	4	
3	1	3	12	40	1	2	1	2	1	
4	1	4	15	32	2	4	1	1	1	
5	2	1	2	26	2	2	1	3	2	

Chaque cas ou observation dans un groupe de données contient des informations sur une entité physique ou logique spécifique. Des exemples d'entités qui pourraient correspondre à un cas sont le ménage, une personne, une ville, une culture, une vente, ou une parcelle de terre. Chacun de ceux-ci est d'une manière, une chose qui peut faire l'objet d'investigation dans une enquête.

Ces exemples semblent être simples, mais dans une enquête complexe comme nous allons le voir, il peut être difficile de faire en sorte qu'un cas représente ce que vous voulez qu'il représente ou ce que vous pensez qu'il représente. L'un des objectifs de cette discussion est de vous permettre de définir clairement et sans ambiguïté la signification des cas dans chacun de vos fichiers de données, et de vérifier pour vous même que vos définitions sont correctes.

Les variables dans un fichier de données contiennent des informations sur les attributs ou caractéristiques de chaque cas. Des variables qui pourraient décrire un cas sont l'âge, la quantité récoltée, le nombre d'animaux possédés, le montant emprunté, etc. Évidemment, ce n'est pas n'importe quelle variable qui est adaptée à chaque type d'entité; ceci est un point important dans la détermination d'une meilleure organisation de fichier.

Pour l'analyse sur ordinateur, les variables doivent avoir des noms de variables usuellement composés de lettres et de nombres comme ceux montrés en haut du tableau ci-dessus. Les cas n'ont pas de noms, mais peuvent être référencés de plusieurs manières dont la plus simple est la numérotation séquentielle comme celle montrée ci-haut. Si nous référons à une variable particulière, nous sommes en train de nous référer à une colonne entière de la matrice avec tous les cas inclus. D'une façon similaire, si nous référons à un cas particulier, nous sommes en train de nous référer à une rangée entière avec toutes les variables incluses.

### **EXEMPLES D'ORGANISATION DE DONNÉES : LA MEILLEURE VERSUS LA MAUVAISE FAÇON**

Il y a souvent une seule meilleure façon d'organiser un groupe particulier de données en variables et cas pendant qu'il existe plusieurs mauvaises façons. Pour que vous sachiez où nous allons, voyons un exemple qui démontre un problème commun dans l'organisation de fichiers en montrant à la fois la mauvaise façon ainsi que la meilleure façon de résoudre le problème.

Supposez que vous aviez une enquête qui ressemble en partie à ce qui suit (ceci n'a pas l'intention de montrer un format achevé d'enquête, mais représente uniquement un modèle pour des buts de démonstration) :

Code du village..... \_\_\_\_\_  
Code du ménage..... \_\_\_\_\_

1. Champ cultivé en 1987

1987

Champ #	Distance Champ/Maison	Grandeur	Partie cultivée	Parcelle?	Méthode de culture	...	Comment L'obtenir	Durée Possession
1	_____	_____	_____	_____	_____	...	_____	_____
2	_____	_____	_____	_____	_____	...	_____	_____
3	_____	_____	_____	_____	_____	...	_____	_____
4	_____	_____	_____	_____	_____	...	_____	_____
5	_____	_____	_____	_____	_____	...	_____	_____

2. Source primaire de  
revenue..... \_\_\_\_\_

.

.

13. Depuis quand vous vivez  
ici?..... \_\_\_\_\_

Tableau de données pour "Fields Planted in 1987" permet jusqu'à 5 champs par ménage. Il y a 11 questions au sujet de chaque champ, commençant avec "Distance from House" et finissant avec "Length of Ownership" (pour des raisons d'espace, toutes ces questions ne sont pas montrées ci-dessus).

### La mauvaise organisation

Une façon d'organiser les données de cette enquête serait de les mettre toutes dans un seul fichier avec des variables définies comme suit (pour des raisons d'espaces, certains noms de variables sont écrits dans des endroits où les valeurs seraient dans le questionnaire actuel) :

1. Champs cultivé en 1987									
# du champ	Distance entre la maison	Dimension	1987 partie cultivée	Parcelle?	Méthode de plantation	...	Comment obtenir	Durée possession	
1	H1	H2	H3	H4	H5	...	H10	H11	
2	H12	H13	H14	H15	H16	...	H21	H22	
3	H23	H24	H25	H26	H27	...	H32	H33	
4	H34	H35	H36	H37	H38	...	H43	H44	
5	H45	H46	H47	H48	H49	...	H54	H55	
2. Source primaire de revenue.....								_____	H56
.									
.									
13. Depuis combien de temps vous vivez ici.....								_____	H67

Ce type d'organisation nous donne un total de 69 variables avec un cas par ménage. Parce que plusieurs ménages auront moins de 5 champs, plusieurs variables auront des valeurs manquantes pour certains cas.

Ceci est probablement la manière la plus simple et la plus évidente pour organiser les données, mais malheureusement, c'est la mauvaise façon. Le principal inconvénient est que l'analyse sera plus tordue et plus susceptible d'erreurs qu'elle ne mérite comme on le verra bientôt.

**La meilleure organisation : Données de "Fields-Planted"**

Une meilleure structure à utiliser pour cette enquête serait de la diviser en deux fichiers, avec des parties différentes de données dans chacun des fichiers. La première partie, les informations concernant les champs semés, doit être entrée dans un fichier propre avec 14 variables comme montré ci-dessus (cette division des données pourrait être faciliter par la division du questionnaire en deux parties). Cependant, pour simplifier cet exemple, on utilisera la structure et la numérotation du questionnaire original en montrant la portion qui va dans le fichier sous discussion.

								Code de la ville.....	VILL
								Code du ménage.....	HH
1.	Champs cultivés en 1987								
#	Distance		1987						
	entre		Partie		Méthode	...	Comment	Durée	
<u>champ</u>	<u>champ/maison</u>	<u>Dimension</u>	<u>cultivée</u>	<u>Parcelle?</u>	<u>plantation</u>	...	<u>Obtenir</u>	<u>Possession</u>	
<b>CHAMP</b>	<b>F1</b>	<b>F2</b>	<b>F3</b>	<b>F4</b>	<b>F5</b>	...	<b>F10</b>	<b>F11</b>	
1	---	---	---	---	---	...	---	---	
2	---	---	---	---	---	...	---	---	
3	---	---	---	---	---	...	---	---	
4	---	---	---	---	---	...	---	---	
5	---	---	---	---	---	...	---	---	

Remarquez premièrement que nous avons enlevé toutes les questions qui ne s'appliquent pas aux champs semés. Aussi, les noms des variables sont maintenant placés en haut du tableau de "fields-planted" pour indiquer que chaque colonne entière a maintenant besoin d'une seule variable. Cette réduction du nombre des variables provient du fait que chaque rangée dans le tableau deviendra maintenant un cas séparé dans le fichier de données au lieu d'être combinée avec d'autres rangées dans un seul cas. Ainsi, chaque ménage qui a cultivé plus d'un champ sera représenté par plus d'un cas dans le fichier de données "fields-planted".

Ceci est un changement important, et qui doit être compris entièrement dans le but de pouvoir s'occuper proprement des données de niveaux multiples. Pour le dire d'une autre manière, différents ménages peuvent maintenant avoir différents nombres de cas, dépendant du nombre de champs semés. Par exemple, un ménage avec un seul champ aura un seul cas pendant qu'un ménage avec 4 champs aura 4 cas ou observations. Le changement dans la structure des cas a nécessité des changements dans la forme, la conversion de la colonne de Field (champ) qui va d'une juste notation utile sur la forme à une variable actuelle appelée FIELD. Comme vous le verrez plus tard, ceci nous permet de distinguer un cas (un champ) d'un autre.

Vous pouvez voir qu'en effet, ce que nous avons fait est d'échanger des variables contre des cas. La première structure proposée avait 55 variables pour les données de "fields-planted", toutes occupant un seul cas et représentant un ménage. La structure améliorée de données a peu de variables, mais plus de cas, chacun représentant un champ (field).

Il est aussi important de comprendre que même si la structure du questionnaire ne le montre pas explicitement, chaque cas doit inclure les variables VILL et HH en plus de FIELD jusqu'à F11. Comparez la structure de la matrice suivante du fichier de données à la forme au dessus de la page.

Observation	VILL	HH	FIELD	F1	F2	F3	...	F10	F11
1	1	1	1	23	2	2	...	3	1
2	1	1	2	34	3	2	...	2	4
3	1	2	1	40	1	1	...	2	1
4	1	2	2	32	2	2	...	1	1
5	2	2	3	26	4	2	...	3	2

Remarquer que la ferme désignée comme village 1, ménage 1 apparaît deux fois avec des entrées identiques pour les variables VILL et HH, mais avec différentes valeurs pour la variable FIELD.

### **La meilleure organisation : Données de ménages**

Le reste des données serait entreposé dans un second fichier contenant aussi 14 variables (le fait qu'il y ait le même nombre est purement une coïncidence), comme montré ci-dessous. Ce fichier, à la différence du fichier de "fields-planted", contiendrait seulement un cas par ménage. Il serait principalement identique au fichier que nous avons décrit plus haut comme la mauvaise organisation, mais avec les données sur champs plantés enlevées.

	Code du Village_____	VILL
	Code du ménage _____	HH
2. Source primaire de revenue.....	_____	HI
.		
.		
13. Depuis quand vous vivez ici?.....	_____	H12

La division des données dans deux fichiers avec différentes structures de variables et de cas est désirable parce que les questions relatives aux champs plantés représentent un niveau différent des questions relatives aux ménages qui les suivent. Ne soyez pas surpris si vous ne savez pas pourquoi ceci apparaît immédiatement apparent. Le reste de ce document essaiera d'expliquer les concepts derrière tout ceci, pour montrer comment reconnaître ce genre de situation dans vos propres données, et de démontrer comment s'en occuper quand vous le faites.

### **Effort de programmation demandé : Mauvaise vis-à-vis la meilleure organisation**

Pour une démonstration vivante du pourquoi de la préférence pour le type d'organisation de fichier séparé, regardons un type particulier de calcul qui pourrait être fait avec ces données. Supposez que nous voulions calculer deux nombres additionnels pour chaque ménage, la taille moyenne de tous les champs que les ménages ont cultivé, et la taille moyenne des champs qu'ils n'ont pas cultivé.

Avec toutes les données dans un seul fichier comme meilleure forme d'organisation des données, ce calcul demanderait les commandes suivantes de SPSS/PC+ :

```
COMPUTE NUM_PLOW=0.
COMPUTE TOT_PLOW=0.
IF(H4 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H4 = 1) TOT_PLOW=TOT_PLOW+H2.
IF(H15 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H15 = 1) TOT_PLOW=TOT_PLOW+H13.
IF(H26 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H26 = 1) TOT_PLOW=TOT_PLOW+H24.
IF(H37 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H37 = 1) TOT_PLOW=TOT_PLOW+H35.
IF(H48 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H48 = 1) TOT_PLOW=TOT_PLOW+H46.
COMPUTE AV_SIZP=TOT_PLOW/NUM_PLOW.
COMPUTE NUM_NPLO=0.
COMPUTE TOT_NPLO=0.
IF(H4 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H4 = 0) TOT_NPLO=TOT_NPLO+H2.
IF(H15 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H15 = 0) TOT_NPLO=TOT_NPLO+H24.
IF(H26 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H26 = 0) TOT_NPLO=TOT_NPLO+H24.
IF(H37 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H37 = 0) TOT_NPLO=TOT_NPLO+H35.
IF(H48 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H48 = 0) TOT_NPLO=TOT_NPLO+H46.
COMPUTE AV_SIZNP=TOT_NPLO/NUM_NPLO.
```

Ceci est presque imposant, n'est ce pas? Et pire encore, si le questionnaire avait jusqu'à 10 champs au lieu de 5, ceci allait doubler le nombre des IF phrases (commandes) exigées! Non seulement c'est fatigant d'entrer toutes ces phrases, mais il existe de grandes potentialités pour des erreurs (dans les noms des variables, par exemple) qui ne seraient jamais décelées.

Par contre, le fichier séparé de Fields-Planted (champs semés) demanderait une seule phrase pour calculer les mêmes variables:

```
AGGREGATE OUTFILE='AGGFIELD.SYS'
/BREAK VILL HH F4
/C_AVESIZ=MEAN(F2).
```

La meilleure nouvelle est que cette phrase ferait le travail complet quelque soit le nombre de champs qu'il y a. Pour être complètement juste, on doit dire que cette phrase ne serait pas équivalente aux 26 phrases demandées pour l'autre méthode. Pour avoir une structure de fichier comparable pour plus

d'analyse exigerait un total de 5 commandes : deux PROCESS IF, deux AGGREGATE, et une JOIN MATCH. Néanmoins, les économies d'efforts sont substantielles, et, probablement plus important, les risques d'erreurs sont réduits.

Un autre avantage de la division en fichiers séparés est que peu d'espace de disque sera nécessaire parce qu'il y a peu de perte. Souvenez-vous que la structure du fichier unique exigeait 55 variables pour les données de champs semés, variables pour lesquelles de l'espace doit être alloué quelque soit le nombre de champs possédés par le ménage. Le fichier séparé de champs semés a seulement 14 variables, et seulement autant de cas comme demandés pour chaque ménage. À moins que la plupart des ménages aient le nombre maximum de champs, le fichier séparé demandera moins d'espace.

### **Variables clés**

Avant de parler de l'organisation de différents niveaux de données, nous avons besoin d'introduire le concept de variables clés. Ceci est un concept de base de gestion de données qui se pose communément quand on a à faire à des programmes de base de données, mais c'est applicable aussi à l'analyse d'enquête.

Le mot clé est utilisé ici non pas dans le sens d'important ou de fondamental (même si les variables clés sont importantes), mais en analogie avec une serrure ou la clé pour déchiffrer un code secret. Cette analogie est utilisée parce que les variables clés sont un instrument important qui vous aide à extraire les informations dont vous avez besoin à partir de vos fichiers de données.

Notre première définition de variables est assez générale pour s'appliquer aussi bien aux variables clés, mais il y a beaucoup à dire au sujet de cette classe spéciale de variables. Dans un sens, les variables clés agissent plus comme des identifiables que comme des attributs. Utilisons un exemple pour illustrer.

Supposons que nous avons conduit une enquête auprès des membres du département d'agro-économie à MSU. Pour une telle enquête, un cas serait évidemment une personne. Maintenant, supposons que j'ai rempli le questionnaire et répondu à quelques questions comme suit :

:

No. de l'employé:	793278634	Nom:	Chris Wolf
Nombre d'années employé:	16	Classification d'emploi:	7

Nombre d'années employé et Classification d'emploi sont de bon exemples de type normal de variable que nous avons discuté dans la section sur les concepts de fichier de données. Ils sont des attributs qui me décrivent, mais ils ne m'identifient pas comme distinct de quelqu'un d'autre dans l'enquête. Il est très possible, même probable que d'autres personnes se trouveront dans l'enquête et qui aient été employées par MSU pour 16 ans aussi bien que d'autres dans la classification d'emploi 7.

La première variable, No. de l'employé, est plutôt différente. Il n'y a aucun autre employé à MSU qui

a le même numéro que moi de sorte que ce numéro particulier est unique à moi parmi toutes les personnes qui pourraient être incluses dans l'enquête. Cette unicité est une caractéristique très importante d'une variable, laquelle rend une variable très utile à nous. Cette unicité permet à la variable d'être une variable clé.

Remarquer aussi une autre caractéristique intéressante de cette variable, contrairement à la durée de mon emploi et la classification de l'emploi qui révèlent à l'observateur quelque chose à mon sujet, mon numéro d'employé ne me décrit pas réellement d'une manière utile. Vous trouverez souvent que les variables clés ne sont pas descriptives de la même manière que sont les autres variables, même si ceci n'est pas une condition préalable pour une variable clé.

En ce qui concerne la variable restante Nom, qu'en est-il ? Il peut être tentant de l'appeler aussi bien variable clé puisque le nom d'une personne est presque unique. Il y a deux problèmes avec ceci : Premièrement, un nom n'est pas unique; il est possible, même si cela n'est probable, que deux employés aient le même nom. Dans le but d'assigner des variables clés, l'unicité doit être garantie pendant l'élaboration de l'enquête. Si la possibilité existe que la valeur d'une variable sera donnée à deux cas non associés, cette variable ne doit pas être utilisée comme variable clé; Deuxièmement, vous ne devez jamais utiliser une variable alphabétique tel qu'un nom comme une variable clé dans un fichier de donnée d'une enquête. En effet, la plupart des logiciels statistiques ne gèrent pas bien des variables qui ont des valeurs alphabétiques, et vous devez les éviter même pour les variables qui ne sont pas des variables clés.

Ainsi, la condition pour une variable clé est très simple : elle doit prendre des valeurs uniques pour chaque cas dans le fichier de données. Dans l'enquête hypothétique de l'employé décrite ci-dessus, il existait une variable naturelle, pré-existante qui était évidemment candidate pour être une variable clé, mais dans la plupart des enquêtes, ce n'est pas le cas. Vous aurez à choisir une variable clé pour convenir à vos propres données. Ceci est fait en développant un schéma de codage qui assigne des nombres d'identification à chacune des entités dans votre enquête.

Un exemple typique de ceci serait un questionnaire de ménage où vous allez enquêter 45 ménages. La variable clé la plus simple à utiliser ici serait une dont les valeurs vont de 1 à 45, peut-être appelée HH. Avec une variable clé de ce type, il est bon d'assigner les valeurs aux cas (observations) d'une façon arbitraire. Vous pouvez les assigner dans l'ordre par lequel ils ont été choisis pour l'enquête, ou par l'ordre par lequel ils ont été interviewés, ou toute autre méthode qui est essentiellement au hasard.

Il y a des circonstances où il est avantageux d'utiliser plus d'une variable clé. Supposez que les 45 ménages dans l'exemple précédent, étaient localisés dans trois différents villages, avec 15 ménages par village, et que l'enquête allait s'intéresser entre autre aux différences entre villages. Dans ce cas, il serait important d'inclure une variable (VILL) identifiant les villages. Ainsi, la procédure normale serait de numéroter les villages de 1 à 3, et de numéroter HH (ménages) de 1 à 15 dans chaque village. Puis, VILL et HH ensemble deviendront les variables clés pour le fichier de données du



ménage, parce qu'aucune de ces variables par elle seule ne pourra identifier un cas unique. Dans des enquêtes de n'importe quelle complexité, les clés multiples sont plus courantes que les clés uniques.

## **NIVEAUX DE DONNÉES**

Avec l'introduction des variables clés déjà faites, nous pouvons maintenant retourner au sujet de l'organisation des données. L'une des caractéristiques les plus significatives d'un fichier de données d'une enquête sont les unités d'observation ou niveaux. Toute enquête légèrement complexe comportera plusieurs différents niveaux de données, et ainsi demandera plusieurs différents fichiers, avec différentes structures de variables/cas, pour gérer toutes les données. Il est très commun pour les chercheurs de manquer de reconnaître ceci, comme dans l'exemple du début, et d'essayer de traiter toutes les données comme si elles étaient au même niveau.

Ainsi, la question est : comment pouvez-vous reconnaître les différentes unités d'observation dans une enquête en vue d'organiser proprement les données? Tout au long de cette discussion, les termes unité d'observation et niveau seront utilisés sans qu'ils soient définis précisément, mais leur sens devrait devenir clair avec les progrès dans la discussion.

### **Exemple de données à multiples niveaux**

Chaque enquête aura une unité d'observation de base qui est d'importance centrale pour le projet et sur laquelle les autres niveaux de données sont basés. Pour les enquêtes menées par les agro-économistes, l'une des unités d'observation les plus communes est un ménage de ferme, ainsi nous allons l'utiliser dans le reste de la discussion comme base de notre unité d'observation.

Si le ménage est notre niveau de base de données, il est évident que nous serions en train de collecter une certaine quantité d'information au niveau du ménage. Ceci pourrait inclure des variables tels que les revenus, la religion, la durée de résidence dans la localité actuelle, les dettes dues, et d'autres.

Etant donné qu'une enquête est centrée sur le ménage, quels seraient les autres niveaux de données? Pour commencer, nous voudrions connaître probablement quelque chose au sujet des différents membres du ménage. Ceci pourrait inclure beaucoup de différentes caractéristiques tels que âge, sexe, relation avec le chef de ménage, années d'éducation, et plus. Cette information comprendrait les données du niveau de membre du ménage, et comme les autres niveaux que nous allons voir, elles doivent être gardées dans un fichier séparé de celui des données du niveau de ménage.

Un autre niveau de données dans notre enquête hypothétique serait le niveau des cultures du ménage. Pour chaque culture que le ménage produit, nous voudrions savoir le nombre d'ares plantées, la quantité d'engrais utilisée, la production, les ventes, la consommation, etc.

Et finalement, pour ajouter un niveau plus complexe de données, supposez qu'il y avait une composante de multiples visites à l'enquête où nous collectons des informations sur le travail chaque

semaine, montrant combien d'heures chaque membre du ménage a travaillé sur chaque culture pendant la semaine. Quel serait le niveau d'observation pour ces données? Vous pouvez les résumer à partir de la précédente phrase, même si ce serait compliqué : le niveau de ménage-membre-culture-semaine.

Dans chacun des trois exemples où nous avons parlé de niveau de données, ce que nous avons fait (même si nous ne l'avions pas exprimé cette façon) était de décider de ce qui fait que chaque cas était unique parmi les autres cas de l'enquête. En regardant vos propres données, vous trouverez que ceci était la plus facile et la plus sûre méthode pour identifier les différents niveaux impliqués. Regardons comment ça marche pour les trois exemples ci-dessus.

Dans le fichier membre de ménage, nous savons que chaque membre de chaque ménage ne sera pas représenté par plus d'un cas. Si un membre était inclus deux fois, nous saurions que nous avons fait une erreur. Cette unicité nous dit que l'unité d'observation pour ces données est le membre de ménage.

D'une façon similaire, dans le cas des cultures du ménage, nous pouvons voir que le critère d'unicité est une combinaison du ménage et de la culture. Chaque ménage peut planter plusieurs cultures, aboutissant à une ou plusieurs observations dans notre fichier de données pour chaque ménage. Ceci nous dit que les données ne sont pas au niveau du ménage. Cependant, s'il y avait plus d'un cas pour la même culture pour un ménage particulier, ce serait une erreur. Par conséquent, les données dans ce fichier sont au niveau ménage-culture.

Dès que l'organisation des données devient plus complexe, il est facile de négliger un facteur qui contribue à l'unicité des cas. Le fichier sur le travail décrit ci-haut, offre un exemple de ceci. Les deux premiers facteurs, le ménage en question et la personne qui a effectué le travail, sont presque évidents. Mais, naturellement, une personne qui a travaillé sur plus d'une culture sera représentée par plus d'un cas, ainsi la culture est une variable qui identifie.

Ce serait facile de s'arrêter à ce point pour décrire le fichier, mais ceci ignorerait le dernier facteur d'unicité. Puisque l'enquête a été menée durant une certaine période, il peut avoir plusieurs cas pour le même ménage, la même personne, et la même culture, mais représentant différentes semaines. Ainsi, parce que ces 4 morceaux particuliers d'information sont nécessaires pour identifier un cas unique, nous voyons que l'unité d'observation pour ces données est le ménage-membre-culture-semaine.

Ce dernier exemple montre que des données de multiples visites ne sont pas traitées différemment des données de visite unique d'une façon substantielle. Avec n'importe quel type, l'approche de base à l'organisation de données est la même, en commençant avec l'identification des facteurs qui distinguent les cas des uns des autres. Une enquête de multiples visites ajoute simplement un facteur de plus qui compte pour l'aspect de temps ou de date pour les données.

## **Variables clés et niveaux de données**

Vous aurez dû remarquer à présent le lien entre variables clés et les unités d'observation : le fait qu'ils participent à l'unicité de cas dans un fichier de données. Si vous pouvez correctement déterminer les variables clés dans un fichier de données, vous connaîtrez les unités d'observation ou niveaux du fichier. Le contraire est naturellement aussi bien vrai; déterminer les unités d'observation vous indiquera ce qui est nécessaire pour déterminer les variables clés.

Vous allez trouver une autre connexion entre les unités d'observation et les variables clés quand vous utilisez SPSS pour analyser les données d'enquête. Deux commandes que vous allez utiliser d'une façon répétitive quand vous travaillez avec des données de différents niveaux sont AGGREGATE (agrégation de données) et JOIN MATCH soit MERGE (fusion de fichiers de données ou de variables). C'est impossible d'utiliser proprement ces commandes sans connaître quelles sont vos variables clés. JOIN MATCH exige de vous l'usage de la sous-commande BY pour spécifier les variables clés à partir des fichiers que vous êtes en train de joindre. AGGREGATE exige de vous l'usage de la sous-commande BREAK pour spécifier les critères d'agrégation, que vous allez découvrir, sont un sous-groupe des variables clés dans un fichier particulier.

## **La hiérarchie des niveaux**

L'usage du mot niveau dans la description de l'organisation de données implique qu'il ya une sorte de classification ou de hiérarchie impliquée. Ceci peut être mis en diagramme d'une façon qui pourrait aider à comprendre les relations impliquées. Les niveaux de données sont traditionnellement pensés comme arrangés du plus agrégé au sommet vers le moins agrégé au bas. Ainsi, dans une enquête typique d'un seul village, nous pouvions avoir une hiérarchie comme ceci (vous allez reconnaître certains niveaux à partir de l'exemple donné auparavant, mais nous avons ajouté quelques uns) :

Niveau supérieur	Ménage		
Niveau moyen	Moisson dans le ménage	Membres du ménage	Équipement du ménage
Niveau inférieur	Moisson des parcelles	Transactions du moisson dans le ménage	

Dans un sens, les éléments qui sont groupés ensemble dans le rang du milieu ont très peu en commun. Un fichier représentant le niveau ménage-membre-culture ne peut pas être combiné ou utilisé ensemble avec un fichier au niveau ménage-membre actuellement. Pourquoi, est-il approprié de penser de ces deux types de fichiers comme étant dans le même rang dans la hiérarchie? Qu'est-ce qu'ils ont en commun?

Ici, encore une fois, nous voyons la connexion entre les variables clés et les niveaux de données. Les différents niveaux qui sont rangés ensemble dans le diagramme ont le même nombre de variables clés, et ainsi peuvent être imaginés comme ayant un niveau similaire d'agrégation. Nous aurions pu libellé le diagramme peut-être plus proprement comme :

Une clé	Ménage		
Deux clés	Moisson dans le ménage	Membres du ménage	Équipement du ménage
Trois clés	Moisson des parcelles	Transaction de la moisson du ménage	

Ainsi, il est important de se rappeler que la hiérarchie représentée ici est strictement une manière convenable de montrer les relations entre les fichiers de données. Les fichiers qui sont montrés dans le même rang sur le tableau n'ont pas d'unités d'observation identiques. Ils sont cependant similaires dans leur niveau, en ce sens qu'ils ont le même nombre de variables clés.

**Terminologie des niveaux**

Vous trouverez souvent que les gens qui ont travaillé avec des données d'enquête pendant un certain temps, ont développé une terminologie abrégée pour se référer aux niveaux de données. Dans un sens technique, cette terminologie est incorrecte, mais puisqu'elle est si commune, vous devrez être familiers avec elle. Puis, à la fin, si vous commencez à l'utiliser vous même, vous serez au courant de ce que vous faites.

Ces abréviations viennent usuellement en existence avec les enquêtes qui ont un niveau primaire de données, disons le niveau du ménage, qui a plusieurs autres niveaux de données sous lui. Il est presque commun par exemple, d'avoir un groupe de données contenant des informations sur chaque culture plantée par chaque ménage, qui serait correctement décrit comme étant au niveau ménage-culture. Parce qu'il devient fatigant d'utiliser la phrase niveau de ménage-culture de plus en plus, la plupart des gens commencent bientôt à se référer à celle-ci comme le fichier de niveau culture.

Comme vous pouvez le voir, ceci est techniquement incorrecte, parce que si le fichier était au niveau culture, il n'aurait pas un seul cas pour chaque culture distincte. Si par exemple, 15 cultures principales étaient plantées dans cette région, le fichier n'aurait pas plus de 15 cas. Puisque le fichier a actuellement des données sur chaque culture plantée par chaque ménage, il pourrait avoir des centaines ou des milliers de cas.

Il est presque possible d'avoir des fichiers de données à partir de la même enquête qui représentent à la fois le niveau culture et le niveau ménage-culture, ce qui rend spécialement important l'usage de la terminologie correcte pour éviter la confusion. La commande AGGREGATE (agréger) de SPSS peut être utilisée pour prendre des données du fichier de niveau ménage-culture et créer un nouveau fichier de niveau culture en additionnant ou autrement en combinant les cas pour tous les ménages pour chaque culture. Ce nouveau fichier aura un cas par culture, et sera ainsi le niveau culture. Vous pouvez voir que ce fichier est radicalement différent de celui auquel vous vous êtes référés plutôt (incorrectement), comme le fichier de niveau culture.

Ainsi, la terminologie simple est bonne aussi longtemps que vous savez ce que vous êtes en train de décrire. Cependant, quand il y a une chance pour une ambiguïté, vous devez faire attention pour utiliser les termes correctes.

## Vérifier l'organisation de vos données

Une fois que vous avez modelé une organisation pour votre fichier de données de façon tentative, il y a des examens que vous pouvez effectuer pour voir si ce modèle d'organisation est correcte.

### 1. Chaque variable particulière doit apparaître une seule fois dans un fichier donné.

Le premier échantillon de l'enquête utilisé ci-dessus avec les questions sur les **champs plantés** (Fields-planted), montre cette erreur potentielle. Comme vous souvenez, le modèle suivant a été montré comme la mauvaise manière d'organiser les données:

		Code du village.....		Code du ménage.....			
1. Champ cultivé en 1987							
1987							
Champ #	Distance Champ/Maison	Grandeur	Partie cultivée	Parcelle?	Méthode de culture	Comment l'obtenir	Durée Possession
1	H1	H2	H3	H4	H4	H10	H11
2	H12	H13	H14	H15	H16	H21	H22
3	H23	H24	H25	H26	H27	H32	H32
4	H34	H35	H36	H37	H38	H43	H44
5	H45	H46	H47	H48	H49	H54	H55
2. Source primaire de revenue.....							H56
.							
.							
13. Depuis quand vous vivez ici?.....							H67

Dans l'organisation de ce fichier, il y a plusieurs instances de variables de même type. Par exemple, les variables H1, H12, H23, H34, et H45, contiennent toutes les données sur **la distance à partir de la maison**. Ces données peuvent être décrites respectivement comme **Distance du Champ 1 de la maison** (Distance of FIELD 1 from House), **Distance du champ 2 de la maison** (Distance of Field 2 from House) et ainsi de suite. Ce genre de structure de variable est un signe évident que quelque chose ne va pas avec l'organisation des données.

La solution, naturellement, est d'enlever les variables offensantes du fichier et de les mettre dans un autre fichier comme des variables singulières, tout comme nous l'avons fait dans la version correcte de ce premier exemple représenté à nouveau ci-bas :

		Code de la ville.....		Code du ménage.....			
1. Champs cultivés en 1987							
1987							
#	Distance entre champ/maison	Dimension	Partie cultivée	Parcelle?	Méthode plantation	Comment Obtenir	Durée Possession
CHAMP	F1	F2	F3	F4	F5	F10	F11
1	----	----	----	----	----	----	----
2	----	----	----	----	----	----	----
3	----	----	----	----	----	----	----
4	----	----	----	----	----	----	----
5	----	----	----	----	----	----	----

Ce nouveau fichier sera à un niveau plus bas de l'analyse par rapport au fichier original. Le fichier original restera au même niveau comme avant, mais il contiendra seulement les variables relatives aux ménages.

Remarquez que trois choses apparaissent quand vous corrigez ce genre d'erreur : le nouveau fichier au plus bas niveau,

- a) aura peu de variables,
- b) aura plus de cas, et
- c) exigera une variable clé additionnelle.

Le dernier point vaut la peine d'être plus élaboré parce qu'il n'avait pas été discuté quand nous avons originalement présenté ce premier exemple. Remarquez que le groupe de variables définies pour la première organisation (incorrecte) de données, n'inclut pas le nombre de champ comme variable. La colonne du nombre de champ (field) est incluse dans la forme parce qu'elle aide à rendre le tableau plus clair, mais les valeurs n'ont pas besoin d'être saisies dans le fichier de données. Le fichier n'a pas réellement besoin d'une variable relative au nombre de champ parce que les variables elles-mêmes contiennent les informations au sujet du nombre de champ, avec H1 à H11 se référant au champ (field) 1, H12 à H22 pour le champ 2, et ainsi de suite. Il y a seulement un cas par ménage, de sorte que le village et l'identité (ID) du ménage servent comme les variables clés, sans aucun besoin de la variable FIELD (champ).

Cependant, quand nous allons à l'organisation préférée du plus bas niveau, chaque champ devient un cas séparé, de sorte que le fichier qui en découle a des cas multiples par ménage. Ceci nécessite une autre variable clé pour distinguer les cas, ainsi nous devons ajouter la variable FIELD. Quand nous disons que chaque type particulier de variable doit apparaître une seule fois, l'usage du mot type est objectivement vague, laissant beaucoup d'options pour votre détermination. La similarité de type n'est pas basée seulement sur le mot utilisé dans la question, mais sur l'intention qui a motivé la question et la manière dont les données devraient être utilisées.

Par exemple, une enquête qui n'a pas inclus des informations détaillées sur les membres de ménage, pourrait toujours demander l'âge du chef de famille. Il pourrait avoir aussi une question demandant l'âge de la maison elle-même. Dans ce cas, ces informations seraient des variables appropriées pour le fichier de niveau ménage. La présence du mot âge dans les deux questions n'est pas par lui-même assez pour clarifier ces deux variables comme de même type.

## **2. Toutes les variables dans un fichier doivent dépendre du groupe complet de variables clés de ce fichier.**

Nous avons déjà discuté deux façons possibles d'organiser nos données de l'enquête échantillon, mais il existe au moins une autre manière qui pourrait apparaître à d'autres personnes. Elle est malheureusement fautive également mais pour une raison différente que celle que nous avons discutée jusqu'ici.

Beaucoup de personnes reconnaissent l'importance de différents niveaux dans les données, mais ne sont pas encore confortables avec l'idée de diviser leur données entre plusieurs différents fichiers. Elles essaient souvent à tous les prix de mettre toutes leurs données dans un nombre limité de fichiers autant que possible. Cette approche pourrait avoir une organisation de fichier qui ressemble à ceci:

```

Code de la ville..... VILL
Code du ménage..... HH

1. Champs cultivés en 1987
   Distance      1987
   entre        Partie
#  champ/maison  cultivée
CHAMP  F1      F2      F3      F4      F5      ...      F10      F11
1  _____  _____  _____  _____  _____  ...  _____  _____
2  _____  _____  _____  _____  _____  ...  _____  _____
3  _____  _____  _____  _____  _____  ...  _____  _____
4  _____  _____  _____  _____  _____  ...  _____  _____
5  _____  _____  _____  _____  _____  ...  _____  _____

2. Source primaire de revenu..... F12
   :
   :
13. Depuis quand vous vivez ici?..... F23

```

Ceci est une sorte d'hybride de l'organisation correcte et incorrecte présentée ci-dessus. Cette organisation peut être regardée fondamentalement comme la même que celle du fichier de champs plantés ci-dessus, avec un cas par champ, mais avec l'addition des variables de niveau ménage (questions 2 à 13).

Comme il peut avoir plusieurs champs (cas ou observations) par ménage, une décision doit être faite concernant le cas qui doit contenir les informations relatives au ménage. Ces informations peuvent être stockées dans chaque champ (cas) qui appartient à un ménage particulier mais ceci exigerait que les données soient saisies plusieurs fois. L'autre choix est de stocker les données comme partie seulement d'un des cas pour chaque ménage, avec le premier champs comme candidat évident. Ce manque de place claire et évidente pour inclure les données de ménage est une indication que quelque chose ne va pas.

Les variables clés dans ce fichier sont VILL, HH, et FIELD, parce que les trois sont nécessaires pour identifier un cas unique. Ceci crée une situation qui n'est pas usuelle, parce que les variables de niveau ménage, F12 à F23, n'ont aucune relations avec la variable FIELD. Par exemple, la question **Pendant combien de temps avez vous vécu ici** n'est pas un attribut d'un champ particulier d'une façon claire. Même si nous pouvons trouver cela convenable de les assigner au cas contenant le champ numéro 1, il n'y a pas de relation réelle entre ces données et ce numéro particulier de champ.

Ceci montre que cette organisation de fichier ne passe pas le second test qui dit que toutes les variables dans un fichier doivent se rapporter au groupe complet de variables clés. La solution au problème est de mettre les variables offensantes dans un fichier à part, et la structure de fichier qui en découle sera identique à celle que nous avons présenté auparavant.

En corrigeant ce genre de violation, l'un des deux nouveaux fichiers qui doit être créé sera à un niveau qui est au-dessus de celui du fichier original, et montrera deux changements principaux. Le fichier au niveau plus élevé aura :

- a) très peu de cas, et
- b) très peu de variables clés

Contrairement à notre correction du premier exemple de type mauvais, il n'y a aucun changement dans le nombre de variables. Ceci parce que la nature de l'erreur est différente.

### **Définir la signification d'un cas**

Si vous regardez à ces deux tests de plus près, vous pouvez voir qu'ils essaient de faire la même chose : de renforcer une définition claire et consistante d'un cas dans chaque fichier de données.

Dans l'exemple qui n'a pas passé le test ci-dessus, un cas représentait un ménage. Un examen rapproché révèle que chaque cas pour un ménage donné pouvait contenir des données pour des champs multiples cultivés. Ceci peut être regardé comme permettant à un cas de contenir de multiples sous-cas qui sont inconsistants avec la définition du cas lui-même. Ceci est, comme nous l'avons vu, une situation indésirable. Dans l'exemple qui n'a pas passé le second test ci-dessus, un cas avait été désigné pour représenter un champ. Ici, nous avons vu que chaque cas pouvait contenir des données qui ne s'appliquaient pas à ce champ particulier, mais plutôt au ménage comme un tout. Ceci est aussi inconsistant avec la définition d'un cas, et par conséquent indésirable.

Comprendre mieux les variables clés et les niveaux de données vous permettra d'avoir une idée plus claire de la signification d'un cas que vous n'aviez auparavant. Quand l'organisation d'un fichier est correcte, vous devez être capable de reconnaître clairement et sans ambiguïté ce qu'est un cas ou observation dans ce fichier. Quand l'organisation d'un fichier est incorrecte, vous devez être capable d'identifier le problème, et comment placer une partie du fichier dans un autre fichier pour corriger ce problème. Votre habileté de faire ceci s'améliorera avec l'expérience mais vous avez dès maintenant les principes de base nécessaires comme point de départ.