QÜESTIIÓ, vol. 23, 2, p. 321-342, 1999

QUALITÉ DE L'INFORMATION DANS LES ENQUÊTES

L. LEBART
Ecole Nationale Supérieure des Télécommunications*

Cet article tente de montrer les contributions des analyses statistiques multidimensionnelles et des analyses textuelles à l'amélioration de la qualité des résultats d'une enquête. L'idée de base est la suivante: chaque phase de l'enquête et de sa réalisation sur le terrain peut donner à la mesure ou au relevé de certaines variables de contrôles. Le fichier des données individuelles se voit alors doublé d'un fichier de variables de contrôles, pouvant être numériques (chronometrage de l'entrevue) nominales (circonstances de l'entrevue, appréciations) ou textuelles (commentaires libres de l'enquêteur et de l'enquêté). Les méthodes précitées permettent alors de visualiser et de confronter ces deux fichiers, et donc de juger la cohérence globale de l'information, en positionnant les variables de contrôles parmi les variables de l'enquête.

Quality of data in sample surveys

Mots clés: Qualité de l'information. Enquêtes socio-économiques. Analyse des données. Analyses statistiques textuelles.

AMS Classification (MSC 2000): 62H25, 62H30, 62D05, 62P25

* Ludovic Lebart. Centre National de la Recherche Scientifique. Ecole Nationale Supérieure des Télécommunications. 46 Rue Barrault. 75013 Paris.

– Reçu en décembre de 1998.
– Accepté en mai de 1999.

1. INTRODUCTION

Pour les statisticiens, le mot enquête désigne le plus souvent une enquête par sondage (enquêtes démographiques, socio-économiques, socio-politiques, épidémiologiques, enquêtes audimétriques ou de marketing). L'enquête est un recueil d'une certaine ampleur, destiné à mesurer et/ou explorer un phénomène. Il est considéré aujourd'hui comme indispensable d'étudier la cohérence et la validité de l'information produite par cet instrument d'observation complexe et délicat à mettre en oeuvre. Cette étude s'appuie notamment sur les techniques actuelle de traitement des données d'enquêtes. Ces techniques ont été profondément modifiées par l'analyse des données qui intervient, dans une phase préliminaire, pour apprécier la qualité de l'information, synthétiser cette information, et orienter la suite des traitements.

La démarche *Data Mining* reprend cette approche globale des grands fichiers, dans le contexte de la diffusion et de la banalisation de la puissance de calcul (cf. Hand, 1998). Après Fayyad *et al* (1996) on peut définir le Data Mining comme la recherche de *patterns* (de traits structuraux) dans de vastes ensembles de données, ces patterns devant être «valides, nouveaux, potentiellement utiles, et, si possible, compréhensibles ou explicables». Les fichiers peuvent être très grands (des millions d'enregistrements), non structurés et non représentatifs. L'objectif ultime est alors d'extraire des données des informations nouvelles de la façon la plus automatique possible.

L'analyse des données textuelles permet d'étendre ce programme aux informations non numériques (réponses libres, textes). Sous le nom de *Text Mining* elle permet de traiter dans la même optique que le Data Mining des corpus de lettres de réclamations, de questions ouvertes dans les enquêtes de satisfaction ou de marketing, des documents Web et internet.

2. LES CONTRÔLES DE BASE DE L'ENQUÊTE

Chaque phase de la réalisation de l'enquête donne lieu au relevé de *variables de contrôles*, qui vont participer activement au traitement de l'information et à l'évaluation de la qualité de l'information.

2.1. La conception du questionnaire

Le questionnaire est indissociable du contenu de l'enquête et des disciplines concernées. Il serait vain de chercher des règles communes à des questionnaires concernant une enquête nutritionnelle dans un pays en développement, une enquête de satisfaction vis-à-vis de produits financiers, une enquête d'audience de Télévision. Au plus peut-on faire

des recommandations générales concernant: la clarté d'expression et de présentation, la lisibilité, la facilité de manipulation et d'encodage. Pour les enquêtes d'opinions, une série de travaux a porté sur les influences des libellés de questions, de l'ordre des questions, de la nature des questions (fermées ou ouvertes) de la longueur des questionnaires sur les résultats (cf. Schuman et Presser, 1981).

Les premiers contrôles font l'objet d'une étude qualitative sur l'accueil et la compréhension du questionnnaire par les personnes interrogées. D'autre contrôles portent sur la formation des enquêteurs.

Les variables de contrôles attachées à chaque entrevue peuvent concerner l'ordre des questions (cas de permutations aléatoires de titres dans les enquêtes d'audience), les questions ouvertes pourquoi, à la suite des questions fermées dont l'interpretation est susceptible d'être variable (cf. section 5).

2.2. Les modes de questionnement

Actuellement, il existe cinq modes principaux de «passation» d'un questionnaire ou de relevé d'informations de base:

- Le mode externe ou observateur.
- Le mode face à face (direct, ou systèmes dits «CAPI» −computer assisted personal interview−).
- Le mode téléphonique (systèmes dits «CATI» −computer assisted telephone interview−).
- Le mode postal (auto-administré).
- Le mode télématique (panel audimétriques, par exemple).

Les modes informatisés (CATI et CAPI) permettent de relever automatiquement un grand nombre de variables de contrôles, comme, par exemple, le chronométrage détaillé de l'entrevue, la durée totale de l'entrevue, l'heure de l'entrevue).

2.3. Le plan de sondage et son exécution

Cette phase donne lieu à des contrôles par inspecteurs, par contre-visites ou contre-appels. Puis, s'il y a lieu, par une analyse de la distribution des poids de redressement.

2.4. Terrain et recueil

La phase 2.4 donne lieu à des contrôles par inspecteurs, par contre-visites ou contre-appels, et par des questions de contrôles portant sur les caractéristiques et appréciations des enquêteurs (niveau de coopération, atmosphère de l'entrevue). Ces questions peuvent comporter les caractéristiques de l'enquêteur (pour tester le niveau éventuel d'*interaction sociale*), le lieu de l'interview, la présence d'autres personnes. Des questions ouvertes sur l'appréciation de l'interveiw peuvent être posées à l'enquêté et à l'enquêteur.

3. LA VISUALISATION DE DONNÉES D'ENQUÊTES

Il est toujours possible de calculer des distances entre les lignes (individus) et entre les colonnes (variables) d'un tableau rectangulaire de valeurs numériques, mais il n'est pas possible de visualiser ces distances de façon immédiate: il est nécessaire de procéder à des transformations et des approximations pour en obtenir une ou plusieurs représentations planes.

3.1. Méthodes factorielles

C'est une des tâches dévolues à l'analyse factorielle au sens large d'opérer une réduction de certaines représentations «multidimensionnelles». On recherche donc des sous-espaces de faibles dimensions (une, deux ou trois par exemple) qui ajustent au mieux le nuage de points-individus et celui des points-variables, de façon à ce que les proximités mesurées dans ces sous-espaces reflètent autant que possible les proximités réelles. On obtient ainsi un espace de représentation, l'espace factoriel.

Mais la géométrie des nuages de points et les calculs de proximités ou de distances qui en découlent diffèrent selon la nature des lignes et des colonnes du tableau analysé.

Les colonnes peuvent être des variables continues ou des variables nominales ou des catégories dans le cas des tables de contingences. Les lignes peuvent être des individus ou des catégories.

La nature des informations, leur codage, les spécificités du domaine d'application vont introduire des variantes au sein des méthodes factorielles.

On rappelle brièvement ici trois techniques fondamentales:

- *L'analyse en composantes principales (ACP)* (Hotelling, 1933) s'applique aux tableaux de type «variables-individus», dont les colonnes sont des variables à valeurs

numériques continues (exemples: enquêtes de consommation, enquêtes de budget-temps) et dont les lignes sont des individus, des observations, des objets, etc. Les proximités entre variables s'interprètent en termes de corrélation; les proximités entre individus s'interprètent en termes de similitudes des valeurs observées. Elle peut donner lieu à de nombreuses variantes en s'appliquant par exemple à un tableau de rangs (diagonalisation de la matrice de corrélation des rangs de Spearman), ou encore après l'élimination de l'effet de certaines variables (analyses locales ou partielles).

- *L'analyse des correspondances (AC)* (Benzécri, 1973) s'applique aux tableaux de contingences (croisement de deux variables nominales). L'analyse fournit des représentations des associations entre lignes et colonnes de ces tableaux, fondées sur une distance entre profils (qui sont des vecteurs de fréquences conditionnelles) désignée sous le nom de distance du *Chi-deux*.
- *L'analyse des correspondances multiples (ACM)* est une extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques. Son champ d'application est considérable. Elle est particulièrement adaptée à la description de grands tableaux de variables nominales dont les fichiers d'enquêtes socio-économiques ou médicales constituent des exemples privilégiés. Les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers); les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions (cf. Lebart, 1975; Lebart *et al.*, 1995; Saporta, 1990).

3.2. Méthodes de classification

Il existe plusieurs familles d'algorithmes de classification: les algorithmes conduisant directement à des *partitions* comme les méthodes d'agrégation autour de centres mobiles; les *algorithmes ascendants* (ou encore agglomératifs) qui procèdent à la construction des classes par agglomérations successives des objets deux à deux, et qui fournissent une hiérarchie de partitions des objets. On se limitera ici à ces deux techniques de classification:

- Les groupements peuvent se faire par recherche directe d'une partition, en affectant les éléments à des centres provisoires de classes, puis en recentrant ces classes, et en affectant de façon itérative ces éléments. Il s'agit des techniques *d'agrégation autour de centres mobiles*, apparentées à la méthode des «nuées dynamiques», ou méthode «k-means», qui sont particulièrement intéressantes dans le cas des grands tableaux (Ball et Hall, 1967; Diday, 1971).
- Les groupements peuvent se faire par agglomération progressive des éléments deux à deux. C'est le cas de la classification ascendante hiérarchique qui peut fonctionner avec plusieurs critères d'agrégation. La technique d'agrégation «selon la variance»

ou de Ward est intéressante par la compatibilité de ses résultats avec certaines analyses factorielles.

Il est aussi possible d'envisager une stratégie de classification basée sur un *algorithme mixte*, particulièrement adapté au partitionnement d'ensembles de données comprenant des milliers d'individus à classer. Un des avantages des méthodes de classification est de donner lieu à des éléments (les classes) souvent plus faciles à décrire automatiquement que les axes factoriels.

Enfin, la pratique montre que l'utilisateur a intérêt à utiliser de façon conjointe les méthodes factorielles et les méthodes de classification (cf. section 4.3 ci-dessous).

4. LE TRAITEMENT GLOBAL DES DONNÉES D'ENQUÊTES

Rappelons la démarche du statisticien lors du dépouillement traditionnel d'une enquête sur ordinateur avec les outils logiciels disponibles (cf. Grangé et Lebart, 1993). Ce dépouillement met en ouvre des techniques simples, éprouvées, faciles à interpréter: les tris, les tableaux croisés, c'est-à-dire des calculs de pourcentages d'individus pour chaque modalité d'une variable nominale (avec ou sans filtre préalable) et des calculs de moyennes de variables numériques ou quantitatives (qui peuvent être ventilées selon les catégories d'une ou de plusieurs variables nominales).

Des méthodes statistiques plus élaborées viennent parfois compléter ces premiers résultats: régressions, analyses de la variance ou de la covariance, modèles log-linéaires.

Les techniques d'analyse des données (analyses descriptives multidimensionnelles) présentées en section 2 modifient profondément les premières phases du traitement des données d'enquête. Elles vont en fait bouleverser l'enchaînement des tâches, et définir une méthodologie nouvelle.

4.1. Les étapes du traitement

Dans le cadre de cette méthodologie, les étapes du traitement des données d'enquêtes sont, brièvement, les suivantes:

1) Descriptions élémentaires (tri-à-plat, histogrammes, calculs de statistiques élémentaires, moyennes, écarts-types, valeurs extrêmes, quantiles). Retour éventuel aux données de base pour une nouvelle saisie partielle ou pour des corrections.
2) *Épreuves de cohérence globale*; Épreuves d'hypothèses larges (par hypothèses larges, on entend: hypothèses générales permises par les nouveaux outils de description). Structuration des données, typologies, sélection de tableaux croisés.

3) Épreuves d'hypothèses classiques (tests statistiques usuels, régression, discrimination, analyses de la variance, modèles log-linéaires...).

4) Conclusions: Critique de l'information de base: lacunes dans le choix des variables, déséquilibre de l'échantillon ou du champ d'observation, biais ou erreurs. Choix de modèles, énoncés des résultats, rejets d'hypothèses, suggestions de nouvelles hypothèses.

La phase 2 est encore souvent absente des logiciels classiques. Lors de cette phase, la cohérence globale du recueil de données peut en effet être éprouvée de façon systématique, des panoramas globaux peuvent être dressés, permettant de critiquer l'information, mais aussi d'orienter la suite des traitements, de choisir les tableaux croisés les plus pertinents. Les typologies (classification des individus en prenant en compte simultanément plusieurs réponses ou plusieurs caractéristiques de base), les outils de visualisation (plans factoriels) fournissent de nouveaux matériaux d'analyse.

Ces opérations, intervenant au début de la chaîne de traitement, permettent de piloter la suite du dépouillement de l'enquête. Le choix des modèles n'est plus fait de façon aveugle en fonction des hypothèses de base: ces hypothèses pourront souvent être critiquées, d'autres hypothèses pourront être suggérées.

Notons que les règles d'interprétation des représentations obtenues à l'issue des techniques de réduction présentées en section 1 n'ont pas la simplicité de celles de la statistique descriptive élémentaire. Une formation et une expérience pratique s'avéreront nécessaires.

4.2. Le modèle de base: éléments actifs et illustratifs (ou supplémentaires)

Il est très intéressant de positionner dans les sous-espaces de représentation des lignes ou des colonnes supplémentaires du tableau de données (Cazes, 1981). On peut ainsi illustrer les plans factoriels par des informations n'ayant pas participé à la construction de ces plans, ce qui va avoir des conséquences importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelés éléments actifs ou variables actives: ils doivent former un ensemble homogène pour que les distances entre individus ou observations s'interprètent facilement. Ils sont en général relatifs à un même thème de l'enquête. Les éléments illustratifs peuvent être très hétérogènes.

Cette dichotomie entre *variables actives* et *variables illustratives* est du même ordre que la distinction que l'on établit entre variables exogènes (explicatives) et endogènes (à expliquer) dans les modèles de régression multiple. D'un point de vue géométrique,

les deux situations sont d'ailleurs très similaires. Les variables exogènes engendrent un sous-espace sur lequel seront projetées les variables endogènes. Les variables actives engendrent aussi un sous-espace, que l'on va réduire pour le visualiser, et c'est sur cet espace réduit que l'on projette les variables illustratives.

4.3. Complémentarité de la classification

Dans le cas du traitement statistique des fichiers d'enquêtes en vraie grandeur, la démarche précédente fondée sur des représentations graphiques a deux graves inconvénients:

1) Les visualisations sont limitées à deux ou en général à très peu de dimensions, alors que le nombre d'axes significatifs peut être plus élevé.
2) Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles. Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats.

L'utilisation conjointe de la classification automatique et des analyses factorielles permet de remédier à ces lacunes. Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes. Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à 30 dimensions. Autrement dit, l'opération de regroupement va présenter un double intérêt: allégement des sorties graphiques d'une part, prise en compte de la dimension réelle du nuage de points d'autre part.

Une fois les individus regroupés en classes, il est facile d'obtenir une description automatique de ces classes: on peut en effet, pour les variables numériques comme pour les variables nominales, calculer des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales; on peut également convertir ces statistiques en valeurs-test et opérer un tri sur ces *valeurs-test*. On obtient finalement, pour chaque classe, les modalités et les variables les plus caractéristiques.

4.4. Sélection raisonnée des tableaux croisés et noyaux factuels

Prenons l'exemple d'une enquête nationale représentative. Étant donnée la structure de la population, les caractéristiques de base (sexe, niveau de vie, statut matrimonial, niveau d'instruction, profession...) ne sont pas indépendantes. Il est utile de décrire le réseau d'interrelations entre toutes ces caractéristiques de base, puis de positionner les autres thèmes de l'enquête en tant qu'éléments illustratifs.

Les caractéristiques des personnes qui répondent sont alors visibles immédiatement dans un cadre qui tient compte des interrelations existant entre ces caractéristiques. Les consultations classiques (sans visualisation factorielle préalable) de tableaux croisés sont en effet redondantes lorsque les caractéristiques qui servent à établir ces tableaux sont liées entre elles.

Le système de projection de variables supplémentaires permet donc d'économiser du temps et d'éviter des erreurs d'interprétation. Chaque variable illustrative fournit une information qui ne pourrait être acquise que par la lecture de nombreux tableaux croisés.

Les noyaux factuels

On désigne par noyaux factuels des groupes d'individus les plus homogènes possibles vis-à-vis de leurs caractéristiques de base.

On aimerait en effet croiser des caractéristiques telles que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout à fait comparables entre eux du point de vue de leur situation objective (réaliser, dans la mesure du possible, le *toutes choses égales par ailleurs*). Mais de tels croisements conduisent vite à des milliers de modalités, dont on ne sait que faire lorsqu'on étudie un échantillon lui-même de l'ordre de quelques milliers d'individus. De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre ces caractéristiques: certaines sont évidentes (il n'y a pas de «moins de 30 ans» retraités), d'autres ont un caractère plus statistique (il y a plus de femmes dans la catégorie «plus de 60 ans»).

Une classification des individus décrits par la batterie active des caractéristiques de base va permettre de regrouper les individus ayant, dans l'échantillon, le maximum de caractéristiques en commun. En pratique, elle fournira des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2 000 individus.

Le tableau croisant une des variables nominales de l'enquête avec la partition en noyaux factuels résume pratiquement tous les tableaux obtenus en croisant cette même variable avec chacune des caractéristiques de base. De plus, certaines interactions indécelables à partir de ces tableaux binaires peuvent être détectées.

4.5. Itération de traitements. Articulation description-inférence

La plupart des techniques évoquées plus haut, dans le cadre d'une première approche, peuvent être mises en oeuvre directement à partir de logiciels standards. Mais l'exigence de l'utilisateur croît avec la connaissance progressive qu'il acquiert de son sujet. Il lui faut croiser des variables de base, regrouper des modalités d'autres

variables, diviser en classes certaines variables continues... en somme préparer les données en vue d'analyses plus fines.

Les *opérations de recodage* font partie d'un processus itératif qui converge vers une connaissance et une assimilation optimale de l'information de base.

4.6. Estimations de données manquantes

La panoplie du statisticien contient des modèles, permettant, à partir de variables quelconques, de prévoir une variable numérique (régression, analyse de la variance et de la covariance), une variable nominale (analyse discriminante, régression logistique: cf.: Bardos, 1989; Celeux et Nakache, 1994; Hand, 1997), d'étudier les associations dans les tables de contingence (modèles d'association, modèles log-linéaires).

La régression et l'analyse discriminante par arbre (Breiman *et al.*, 1984), qui améliore les méthodes classiques de segmentation utilisées en marketing.

Enfin les réseaux de neurones (Hérault et Jutten, 1994; Thiria *et al.*, 1997) constituent des modèles souples et non-linéaires qui généralisent la plupart des méthodes précitées. Leur fonctionnement en tant que *boite noire*, la difficulté d'interprétation des paramètres, les problèmes de convergence numérique font que ces méthodes ne se substituent pas totalement aux méthodes statistiques plus classiques.

Une des difficultés majeures de l' articulation description-modèles tient au fait qu'on ne peut de façon valide tester sur des données un modèle statistique découvert sur ces mêmes données (Cox, 1977). Il va de soi que le traitement des données d'enquêtes n'est pas le seul domaine où ces problèmes se rencontrent. Des techniques du type «échantillon test» ou «validation croisée» pourront aider à contourner ces obstacles (cf. McLachlan, 1992).

Tous ces modèles permettent dans de nombreux cas d'estimer des données manquantes. Les méthodes de fusions de fichiers (cf. par exemple Aluja *et al.* 1997), permettent également de procéder à des imputations de blocs de variables.

5. VISUALISATION DE DONNÉES TEXTUELLES

Les analyses statistiques de textes peuvent intervenir à deux niveaux dans un contexte industriel et commercial: au niveau du traitement des lettres de réclamations, des cahiers de doléances ou de suggestion, au niveau de questions ouvertes dans des enquêtes postales ou téléphoniques. Les questions ouvertes les plus simples et les plus

fréquentes sont d'une part la question «pourquoi» posée après une question fermée, et d'autre part les questions du type «autre, préciser», comme item de réponse complémentaire à une question fermée. Le traitement proposé va produire de façon automatique des *mots caractéristiques* et des *réponses caractéristiques* pour diverses catégories de répondants.

5.1. Questions ouvertes dans les enquêtes

Il peut donc être intéressant, dans un certain nombre de situations d'enquête, de laisser ouvertes des questions, dont les réponses se présenteront sous forme de textes de longueurs variables. Les outils de calcul et les méthodes statistiques descriptives multidimensionnelles apportent une aide au traitement de ce type d'information, évidemment complexe. Plus généralement, le *Text Mining* désigne l'analyse exploratoire de très grands recueils de textes. Le cas des questions ouvertes est un cas favorables de text mining, puisque les textes ont une homogénéité exceptionnelle (réponses à une même question) et sont accompagnés d'informations complémentaires très riches (questions fermées).

Bien que les réponses libres et les réponses aux questions fermées fournissent des informations de natures différentes, les premières sont plus économiques que les secondes en temps d'interview et génèrent moins de fatigue. Une simple question ouverte (par exemple: «Avez-vous des réclamations à formuler concernant ce produit?») peut remplacer de très longues listes d'items. Notons que les questions ouvertes sont considérées comme peu adaptées aux problèmes de mémorisation de comportement. «Quels sont les noms des magazines que vous avez lus la semaine dernière?». Pour ces questions qui font l'objet d'enquêtes périodiques, il a été prouvé maintes fois que les questions fermées donnent des taux d'oubli plus faibles (Belson et Duncan, 1962).

5.2. Les traitements statistiques de textes

Il existe deux grandes séries d'applications des analyses statistiques de textes, selon que l'on s'intéresse à la forme ou au contenu:

- Les applications à des textes littéraires (attributions d'auteurs, datation, par exemple) qui cherchent à saisir des caractéristiques de forme et de style à partir des distributions statistiques de vocabulaire, d'indices ou de ratios, ou encore à partir de corpus partiels de mots-outil (articles, conjonction, etc.). (cf. par exemple Holmes, 1985, pour une revue de ces travaux).
- D'autre part les applications réalisées en recherche documentaire (Salton, 1988), en codification automatique, dans le traitement des réponses à des questions ouvertes, qui s'intéressent principalement au contenu, au sens des textes.

Cependant, lors du traitement statistique de réponses à des questions ouvertes, ou lors des analyses d'entretiens, le socio-linguiste peut être aussi intéressé par la forme, par les connotations véhiculées par exemple par certains synonymes, certaines tournures (cf. par exemple: Achard, 1993). Les méthodes d'analyses de réponses libres dans les enquêtes relèvent de cette seconde famille d'application (Lebart et al, 1999).

5.3. Les unités statistiques découpées dans les textes

Les formes graphiques

L'unité statistique de base est la forme graphique, suite de caractères non-délimiteurs (en général des lettres) entourée par des caractères délimiteurs (blanc, points, virgules...). Un même mot pourra souvent donner lieu à plusieurs formes graphiques, selon son cas ou son genre dans le texte. Une même forme graphique peut renvoyer à plusieurs mots (en français, avions renvoie à un nom, mais aussi au verbe avoir). Cela n'est pas toujours un inconvénient grave, car les formes graphiques ne seront pas traitées isolément. Les traitements statistiques concerneront en effet les profils de fréquences de formes graphiques, c'est-à-dire les vecteurs dont les composantes sont les fréquences de chacune des formes utilisées par un individu ou un groupe d'individus. Ces profils contiennent une information extrêmement riche. Plus précisément, les techniques mettront en évidence les différences entre profils de formes graphiques.

«**Mots-outil**», parties du discours

Des progrès importants ont été réalisés dans le domaine de l'analyse syntaxique automatisée des textes, comme en témoigne, par exemple l'amélioration constante des correcteurs orthographiques. Des analyseurs syntaxiques permettent de calculer la proportion de noms, de verbes, d'adjectifs, etc.

Notons que si l'isolement de mots-outil (encore appellés mots vides ou mots grammaticaux) demande une désambiguïsation du texte —cas de la forme pas en français, par exemple— il existe aussi des locutions contenant des mots pleins qui sont des substituts de mots-outil (*de façon que, en même temps que, sans oublier*, etc.).

Les unités lemmatisées

Un autre type de traitement préliminaire du texte consiste à procéder à une lemmatisation. Cette opération, difficile à réaliser de façon entièrement automatique, consiste à remplacer les formes par l'entrée du dictionnaire correspondant (infinitif pour les verbes, masculin singulier pour les adjectifs, formes non élidées à la place des formes

élidées, etc.). Elle est parfois complétée par la suppression de certains mots-outils (articles, conjonctions, etc., cf. par exemple Reinert, 1986). En documentation automatique, cela permet de travailler avec un nombre restreint de mots-clé dont les occurrences sont fréquentes. Une lemmatisation complète demande une analyse morpho-syntaxique approfondie, et ne peut être entièrement automatique (cf. Charniak, 1993). En traitement de questions ouvertes, cette opération n'est pas toujours souhaitable a priori car elle détruit certaines locutions. En revanche, elle peut intervenir comme complément, car elle fournit un point de vue différent de celui fourni par une analyse entièrement automatique sur les formes graphiques du texte. Dans le cas d'entretiens non directifs peu nombreux, la lemmatisation permet de travailler avec des seuils de fréquences de mots plus élevés que ceux nécessités par l'analyse des formes graphiques.

5.4. Les analyses statistiques; les trois outils de base

Une numérisation préliminaire (qui est aussi une compression) consiste à affecter à chaque nouvelle forme graphique un numéro d'ordre qui sera associé à toutes les occurrences de cette même forme. Ces numéros seront stockés dans un dictionnaire de formes, ou vocabulaire, propre à chaque exploitation. Les trois outils de base sont l'analyse des correspondances des tableaux lexicaux, les sélections de formes caractéristiques, les sélection de réponses modales.

a) *Analyse des correspondances des tableaux lexicaux*

Les analyses des correspondances (cf. section 3) peuvent décrire les tables de contingence croisant les réponses et les formes graphiques, ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories.

Avec ce type de représentation, la présence de mots-outils est parfaitement justifiée: si ces mots caractérisent électivement certaines catégories, ils se positionnent dans leur voisinage, et peuvent être intéressants à interpréter; si au contraire leur répartition est aléatoire, ils s'abîmeront dans la partie centrale du graphique, sans en encombrer la lecture.

b) *Formes ou segments caractéristiques (ou spécificités)*

Il est tentant de compléter les représentations spatiales fournies par l'analyse des correspondances par quelques paramètres d'inspiration plus probabiliste: les spécificités ou formes caractéristiques. Ce seront les formes «anormalement» fréquentes dans les réponses d'un groupe d'individus (cf. Lafon, 1980). Un test simple fondé sur la loi hypergéométrique permet de sélectionner les mots dont la fréquence

dans un groupe est notablement supérieure (ou inférieure pour les mots anti-caractéristiques) à la fréquence moyenne dans le corpus.

c) *Les sélections des réponses modales*

Pour un groupe d'individus donné, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents-type, selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux la classe. On peut, pour chaque regroupement, calculer la distances du profil lexical d'un individu au profil lexical moyen du groupement. On peut ensuite classer les distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. On obtient ainsi une sorte de résumé des réponses de chaque regroupement, formé de réponses originales.

5.5. Stratégie de traitement

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et de signification que ceux des réponses isolées. Ce regroupement a priori peut être réalisé à partir des variables disponibles, retenues en fonction de certaines hypothèses. Mais ceci suppose une bonne connaissance préalable du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites exploratoires.

Regroupement par noyaux factuels

La technique dite des «noyaux factuels» déjà évoquée en section 4 va permettre de donner des éléments de réponse à ce problème. Étant donnée une liste de descripteurs ou de variables caractérisant les individus, le problème est de regrouper les individus en groupes les plus homogènes possibles vis-à-vis de ces caractéristiques, sans en privilégier certaines a priori. La partition obtenue est une sorte de «partition moyenne» qui résume les principales combinaisons de situations observables dans l'échantillon, et qui permet donc de procéder à des regroupements de réponses textuelles les moins arbitraires possibles.

Analyses directes sans regroupement

Une telle analyse produit une typologie des réponses, en général assez grossière, et produit de façon duale une typologie de mots ou de formes graphiques.

Il est donc possible d'illustrer ces typologies par les caractéristiques des individus interrogés qui auront le statut de variables supplémentaires ou illustratives. Ce traitement

direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé.

5.6. Conclusions sur les analyses de textes

Il s'agit avant tout d'une confrontation de questions ouvertes et de questions fermées. L'analyse est essentiellement différentielle, comparative, et en cela se distingue de l'analyse de contenu classique. Elle ne vise en effet qu'à décrire les contrastes entre plusieurs textes, ces textes étant les réponses originales, ou des regroupements de réponses réalisés à partir des questions fermées de l'enquête.

Pour une question ouverte et pour une partition de la population, on obtient donc, de façon intégralement automatisable:

- Une visualisation des proximités entre formes et catégories, par analyse des correspondances du tableau lexical agrégé, éventuellement complétée par une visualisation similaire des proximités entre segments et catégories.
- Les formes (et/ou segments) caractéristiques de chaque catégorie.
- Les réponses modales de chaque catégorie.

Ces résultats, obtenus sans codification ni intervention manuelle, fournissent des compléments et donnent des éléments critiques nouveaux pour juger à la fois la cohérence et la pertinence du questionnement, la compréhension des réponses, ainsi que le niveau d'implication ou de participation des répondants. Ils participent donc à l'amélioration de la qualité de l'information, et fournissent des éléments originaux au dossier des analyses de satisfaction.

6. PROBLÈMES DE QUALITÉ D'INFORMATION

Les visualisations de variables numériques ou textuelles qui viennent d'être évoquées permettent de prendre en compte certaines déficiences de l'information de base (non-réponses par exemple), ainsi que des variables de contrôles liées à la qualité du recueil de l'information de base (cf. ASU, 1992).

Conjectures sur les non-réponses

Les non-réponses sont des modalités comme les autres, qui peuvent être positionnées dans les espaces factoriels des thèmes, comme dans les espaces de la structure de base. Le traitement des non-réponses qui se prête mal aux tests statistiques usuels reçoit ici

une importante contribution, dans la mesure où l'on peut étudier le contexte de ces refus ou lacunes, soit en termes de caractéristiques des répondants, soit à partir des réponses effectives à d'autres thèmes.

Le positionnement des variables techniques

Dans l'espace des caractéristiques de base ou dans celui des principaux thèmes, que ceux-ci soient résumés par un plan factoriel ou par une partition, il est possible de placer les modalités de variables nominales dites «techniques» telles que: numéro ou nom de l'enquêteur, caractéristiques diverses de l'enquêteur, heure de l'interview, lieu et durée de l'interview, appréciation de l'enquêteur ou de l'enquêté sur l'interview, etc.

On obtient ainsi un panorama de la fabrication de l'information, permettant de rapprocher globalement les circonstances des interviews et les caractéristiques des personnes interrogées. Cette confrontation permet souvent d'apprécier la validité des données de base et de nuancer l'interprétation des résultats.

Les questions ouvertes

Alors que la question ouverte *pourquoi?* après une question fermée permet de vérifier la compréhension de la question, des questions de commentaires libres à l'issue de l'interview (questions qui peuvent être posés à l'enquêté, mais aussi à l'enquêteur) permettent de critiquer aussi bien le questionnaire que les conditions de sa passation.

Finalement, dans le traitement des données d'enquêtes l'approche globale et exploratoire est un élément important d'une *démarche qualité*. En effet, détecter des patterns, c'est évidemment dans une première phase détecter des anomalies, des incohérences, des points aberrants (*outliers*), des hétérogénéités inattendues. Dans le cas des données d'enquêtes, particulièrement en présence de questions ouvertes, cela peut amener non seulement une critique de la réalisation pratique de l'enquête et du recueil de l'information sur le terrain, mais cela peut dans certains cas aller jusqu'à une remise en cause de la conception de l'enquête et de son questionnaire.

RÉFÉRENCES

Achard, P. (1993). *La sociologie du langage*. PUF. Paris.

Aluja Banet, T., Rius, R., Nonell, R. and Martínez-Abarca, M.J. (1997). «Data Fusion and File Grafting», *Proc of the IV ème Congrès International d'Analyse Multidimensionnelle des Données. NGUS-97,* Bilbao, 22-28.

ASU, (Lebart, L., ed.) (1992). *La qualité de l'information dans les enquêtes*. Dunod, Paris.

Ball, G.H. and Hall, D.J. (1967). «A clustering technique for summarizing multivariate data», *Behavioral Sciences*, 12, 153-155.

Bardos, M. (1989). «Trois méthodes d'analyse discriminante», *Cahiers Economiques et Monétaires*, 33, 151-190.

Belson, W.A. and Duncan, J.A. (1962). «A Comparison of the check-list and the open response questioning system», *Applied Statistics*, 2, 120-132.

Benzécri, J.P. (1973). *L'Analyse des Données. Tome 1: La Taxinomie. Tome 2: L'Analyse des Correspondances*, (2de. éd. 1976). Dunod, Paris.

Breiman, L., Friedman, J.H., Ohlsen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Cazes, P. (1981). «- Note sur les éléments supplémentaires en analyse des correspondances», *Les Cahiers de l'Analyse des Données*, 1, 9-23; 2, 133-154.

Celeux, G. and Nakache, J.P. (eds). (1994). *Analyse discriminante sur variables qualitatives*, Polytechnica, Paris.

Charniak, E. (1993). *Statistical Language Learning*, The MIT Press, Cambridge.

Cox, D.R. (1993). «The role of significance tests», *Scandinavian Journal of Statistics*, 4, 49-70.

Diday, E. (1971). «La méthode des nuées dynamiques», *Revue Statist. Appl.*, 19, 2, 19-34.

Fayyad, U., Piateski-Schapiro, G., Smyth, P. and Uthurasamy, R. (Eds) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press/The MIT Press.

Grangé, D. et Lebart, L. (1993). *Traitement statistique des enquêtes*. Dunod, Paris.

Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. J. Wiley, Chichester.

Hand, D.J. (1998). «Data mining: Statistics and more?», *The American Statistician* (May issue).

Hérault, J. et Jutten, C. (1994). *Réseaux neuronaux et traitement du signal*. Hermès. Paris.

Holmes, D.I. (1985). «The analysis of literary style - A Review», *J.R. Statist. Soc.*, 148, 4, 328-341.

Hotelling, H. (1933). «Analysis of a complex of statistical variables into principal components», *J. Educ. Psy.*, 24, 417-441, 498-520.

Lafon, P. (1980). «Sur la variabilité de la fréquence des formes dans un corpus», *Mots*, 1, 127-65.

Lebart, L. (1975). «L'orientation de dépouillement de certaines enquêtes par l'analyse des correspondances multiples», *Consommation*, 2, 73-96.

Lebart, L., Morineau, A. y Fenelon, J.P. (1985). *Tratamiento Estadístico de Datos*. Marcombo, Barcelona.

Lebart, L., Salem, A. y Becue, M. (1999). *Análisis Estadístico de Textos*. Ediciones Milenio, Lleida.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. J. Wiley, New York.

Reinert, M. (1986). «Un Logiciel d'analyse lexicale», *Les cahiers de l'analyse des données*, 4, Dunod, 471-484.

Salton, G. (1988). *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, New York.

Saporta, G. (1990). *Probabilités, analyse des données et statistiques*. Technip, Paris.

Schuman, H., Presser (1981). *Questions and Answers in Attitude Surveys*. Academic Press, New York.

Thiria, S., Gascuel, O., Lechevallier, Y. et Canu, S. (1997). *Statistique et méthodes neuronales*. Dunod, Paris.

ENGLISH SUMMARY

QUALITY OF DATA IN SAMPLE SURVEYS

L. LEBART
Ecole Nationale Supérieure des Télécommunications*

This paper aims at highlighting the contribution of both exploratory multivariate methods and textual analysis to the assessment of data quality in surveys.
The basic idea is that each step in carrying out a survey, from the conception of the questionnaire to the very end of the interview, can provide us with a series of control variables. These variables could be numerical (e.g. duration of each interview), nominal (e.g. circumstances, location of the interview) or textual (e.g. open ended questions about the content of the survey).
The methods mentionned above allow for confrontations of initial data files and the control file, leading to an assessment of the obtained information.

Keywords: Data quality, Socioeconomic Surveys, Exploratory Multivariate Data Analysis, Textual Data Analysis.

AMS Classification (MSC 2000): 62H25, 62H30, 62D05, 62P25

* Ludovic Lebart. Centre National de la Recherche Scientifique. Ecole Nationale Supérieure des Télécommunications. 46 Rue Barrault. 75013 Paris.

– Received December 1998.
– Accepted May 1999.

1. INTRODUCTION

Survey data processing will involve exploratory multivariate data analysis and textual data analysis, that are closely related to Data Mining and Text Mining.

2. BASIC CONTROLS OF A SURVEY

When carrying out a sample surveys, each step leads to specific controls. Besides, it provides the researcher with a set of control variables that can be added to the variables relating to the content of the survey.

3. VISUALIZATION OF SURVEY DATA

The techniques of visualization dealt with in this section are those adapted to the processing of large data tables. Principal axes methods (Principal component analysis, simple and multiple correspondence analysis) are the most common methods.

Among clustering techniques, the k-means method is certainly the most largely used, thanks to its ability to tackle huge data sets.

4. THE GLOBAL PROCESSING OF SURVEY DATA

The most common phase of the processing requires using the demographic variables to design the frequency distributions and the cross-tabulations. These same demographic variables can also be used to build a typology of the respondents. In fact, several typologies of the respondents can be derived from several sets of the so-called active variables (homogeneous set of variables relating to a particular theme of the questionnaire).. Each of these step defines a specific point of view. The basic procedure consists in positionning supplementary variables in the maps corresponding to each typology.

The complementarity of clustering is stressed, due to the limitation of the visualization to a small number of dimensions. Moreover, these visualizations can include several hundred points, and give rise to crowded or illegible graphs, and to lengthy lists of coordinates.

Thus it is important to make use of the data management and computational capabilities of the computer to complete and clarify the presentation of the results. The combined use of clustering methods and correspondence analysis can fill in the gaps. When there

are too many points on a graph it may be useful to group the data into homogeneous families. The algorithms used for developing these groups work the same way whether the points are located in a two-dimensional or a ten-dimensional space.

In other words, the process has two objectives: to minimize graphical printouts, on the one hand, and to work with the real dimensionality of the configuration of points, on the other hand. Once the individuals have been grouped into clusters, it is straightforward to obtain a description of these clusters: indeed, statistics related to differences between internal values for each cluster and overall values for the sample can be calculated for numerical variables and categorical variables. These statistics can also be converted into *test-values* and sorted on these test-values.

Finally, the most characteristic response categories and variables can be displayed for each cluster.

All the technical variables collected can be projected as illustrative (or: supplementary) variables. This helps us to detect possible interactions between some aspects of the content of the surveys and some apparently lateral features such as the gender of the surveyor, the day or the time of the interview, the presence of other persons during the interview, the duration of the interview, the opinion of the surveyor about the interview, etc.

5. VISUALIZING TEXTUAL DATA

Open ended questions in surveys can contribute to the assessment of data quality. Open-ended questions can be found in the questionnaires of many surveys performed in socioeconomy, epidemiology, advertising, business or political marketing. They become an essential part of these questionnaires when the scope of research goes beyond a simple tally, and when a complex and unknown topic is being explored.

Open ended questions are less costly in terms of interview time, and generate less fatigue and tension. But a fundamental advantage lies in their use *to probe the response to a closed-end question.*: This is the follow up additional question «*Why?*». Explanations concerning a response already given have to be provided in a spontaneous fashion. A battery of items might suggest new ideas that would only mar the authenticity of the explanation given.

The main tools are the *correspondence analysis of lexical tables*, showing on a series of mappings the associations between the words used and the characteristics of the respondents, the *characteristic words*, (list of words or phrases most associated with a specific category of respondent) and the characteristic responses, pinpointing the responses most characteristic of a category.

6. DATA QUALITY ASSESSMENT

The powerful tools briefly described above allow for a new level of assessment of consistency in survey data processing. All sorts of control variables, no-response items, open responses (through characteristics words or responses) can now be positionned at a low cost among the variables relating to the content of the survey.