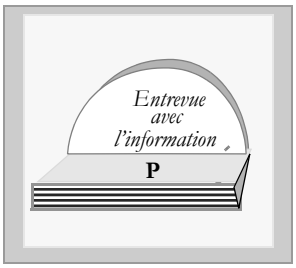


ENTREVUES AVEC L'INFORMATION

Comme il est rassurant de constater ici qu'il se confirme que l'écriture a été inventée par des analphabètes.

Il y a pourtant des antécédents historiques. Ne pouvant lire ni donc comprendre les hiéroglyphes des Égyptiens, sans doute accessibles tout au plus à une classe privilégiée, les Anciens ont commencé à les transcrire pour les exprimer en sons syllabiques, ce dont une des versions est le cunéiforme ("en coins") des Hittites et autres Belges préliminaires.

Le cunéiforme... Il est beau d'écrire ainsi en traces de roseau dans l'argile souple : la poésie est déjà sur le support, et ces petits autels en plaques de terre n'attendent que le sacrifice de la Pensée du Lecteur. Les autres écritures vont suivre, jusqu'à la merveilleuse HTML, la pictographie et l'exotique Java-script.

Serait-il possible que 3000 ans plus tard une histoire tout-à-fait semblable se reproduise et que, incapable de comprendre les problèmes des EAH de son temps, perplexe devant le complexe, écarté des hiéroglyphes des sciences par l'ésotérisme des savants, ce soit un ignare qui ait conçu et mis au point un moyen de produire de la connaissance, dit "système d'investigation"?

C'est possible, mais sur une portée musicale, un ruisseau argenté dont on n'aurait dessiné que les ondes, on peut écrire les sons, mais pas leur contenu.

Quant aux écritures, elles ne donnent que le contenu des messages, mais pas leurs sons, pas leur musique. Or, chez les Roast-beefs, mais avant l'époque des bœufs fous, quelqu'un a entendu dire par Shakespeare : « Méfie-toi de l'homme qui n'a pas la musique en lui ».

Il est curieux, alors, que les mathématiques disent des "certitudes", car cette discipline est l'étude des "vérités hypothétiques" (selon Peirce, 1870), alors que l'information, elle, dit des incertitudes, justement parce qu'elle aussi communique des "vérités hypothétiques"...

ENTREVUES AVEC L'INFORMATION

Sommaire

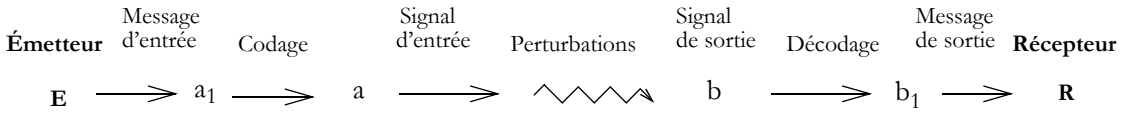
1 Le réseau et ses agents

1.1 Schéma de la communication entre un émetteur et un récepteur

Un émetteur **E** envoie des signaux à un récepteur **R** via la simple ligne de communication de la Figure **1**. Dans le contexte de la gestion, l'*émetteur* sera assimilé à un *investigateur* (un "chercheur-informateur") et le *récepteur* à un *manager*. Celui-ci est dans une situation d'*intention*, et il faudra éclairer ses choix, surtout s'il est Chef. Cette ligne de communication implique la séquence suivante:

- L'émetteur **E** envoie un message (a_1), lequel subit un codage, donnant un signal (a) cheminant sur une ligne de communication;
- Le signal peut y être affecté de perturbations, passer des obstacles ou des filtres;
- Il en résulte le signal de sortie (b) à décoder pour obtenir le *message à la sortie* (b_1) en faveur du récepteur **R**.

Figure 1. Ligne de communication Émetteur-Récepteur



1.2 Les mesures d'information sans téléonomie

1.2.1 Les champs d'incertitude et leurs variations

Un champ d'incertitude est un ensemble de lieux des points dans un espace de probabilité ou de crédibilité. La théorie mathématique de l'information suppose que le montant d'information communiquée par un émetteur à un récepteur est lié à la *variation d'incertitude* produite par des variations de probabilités dans les espaces probabilistes considérés. Plus récemment s'est développée l'approche des *degrés de croyance* dans un espace de *crédibilité*, approche qui n'est pas présentée ici.

Or, dans le schéma décisionnel, il existe 3 espaces probabilistes, associés respectivement:

- Aux *événements* possibles, ou aux contextes aléatoires appelés "états de la nature";
- Aux lignes d'*actions* disponibles (ce qui implique les décisions potentielles);
- Aux *conséquences* résultant de l'interaction des deux premiers.

Ceci implique 3 champs d'incertitude qui ne sont d'ailleurs pas indépendants. Les mesures d'information associées au processus de communication sont alors classées en deux groupes:

- Celles qui sont seulement orientées vers la *connaissance* du système de communication et des champs d'incertitude des événements du preneur de décision; on pourrait qualifier cet aspect de physique plutôt que d'économique;
- Celles qui font intervenir la notion de *valeur*, et par suite soient liées directement à la *situation d'intention*, ou, de façon équivalente, à l'information en relation avec une *téléonomie*. Cet aspect relève plus de l'économie et de la gestion.

Seul le premier champ est concerné par la théorie des communications et de l'information, laquelle développe des mesures construites sur des probabilités et variations de probabilités. Depuis HARTLEY [1928; les références, plus nombreuses, sont cette fois en fin d'exposé], puis surtout SHANNON & WEAVER [1949] ainsi que BRILLOUIN [1956], la plupart des auteurs qui présentent l'information en ces termes mathématiques se concentrent sur le *montant* d'information qui *peut* être communiqué plutôt que sur le montant qui est *réellement* communiqué. Un fabuleux exposé de cet aspect est «L'Information et sa thermodynamique», apothéose de Tome Nord.

Ce montant d'information qui peut être communiqué est une contribution quantitative à la valeur du réseau d'information indépendamment de la signification des messages transmis par le système, ce qui a même conduit à se demander si le terme "information" était bien adéquat pour ce contexte. Quoiqu'il en soit, ces mesures s'adressent donc bien à l'information sans contenu, c'est-à-dire..."qui ne veut rien dire" – mais on va... l'utiliser!

1.3 Relation manager-investigateur

La version froide du processus d'information est un enchaînement, en séquence et en parallèle, de liens de communications entre émetteurs et récepteurs. La version de systémique en *gestion* apparaît avec des *agents*. Le processus devient alors *interactif*, en ce sens qu'en général le récepteur s'enquiert auprès de son émetteur ("informateur"), en fonction de ses propres intentions de "manager", ce qui provoque un aller-retour de messages.

En gestion, l'informateur lui-même peut avoir aussi ses propres intentions; elles ne sont pas nécessairement identiques à celles du "manager", et il peut consulter lui aussi des sources fournissant différents types d'informations.

Ceci sera porté à la Figure **2**, qualifiée de processus d'information "étendu". Chacune des communications qui y figure peut donner lieu à des biais, des perturbations et des déformations dont on donnera plus loin une liste pour l'exemple. Le résultat global peut être un sérieux manque de correspondance entre les attentes du manager et son interprétation des messages effectivement reçus. En effet la réponse adéquate à une demande d'information suppose pour le moins que le manager connaisse lui-même ses intentions.

Toutefois, les intentions du manager-décideur peuvent être:

- Mal définies, ce qui peut être dû à leur nature complexe, par exemple si elles sont multicritères ou conflictuelles;

- Instables, donc varient sur la période d'information et d'analyse;
- Mal communiquées, pas claires pour l'informateur:
  - Soit parce que le message du récepteur à l'informateur est ambigu;
  - Soit parce que la communication est sciemment incomplète ou déformée.

Dans ces cas, il y a déjà ambiguïté quant à la finalité de l'information et le chercheur-informateur va prendre lui-même l'initiative de l'ensemble d'information qui satisfait sa finalité. Il en résulte que l'information rendue reposera souvent plus sur la conception des choix et de l'analyse du chercheur-informateur que sur celle du manager-récepteur.

L'informateur peut aussi avoir des intentions propres qui l'influencent; parfois on lui impose des objectifs différents de ceux du (ou des) récepteur(s) des messages: ce sera souvent le cas de l'information politique, avec comme cas-limite l'"intoxication". Dans ce dernier cas, cependant, le récepteur n'est pas nécessairement demandeur et n'a pas communiqué au préalable ses souhaits: il subit l'emprise de l'information.

2 Les sources d'information

En économie et gestion, les sources d'information reconnues officiellement (c'est-à-dire, par les politiques, la radio, la télé, les vendeurs de tapis, les statistiques et les sondages, bref, par tout ce qui vous apprend tout sauf la vérité), sont, paraît-il:

- Les sources secondaires;
- Les sources primaires:
  - Les expérimentations naturelles;
  - Les expérimentations contrôlées;
- Les systèmes abstraits;
- Les répondants.

Pour savoir quelles rumeurs ils colportent, autant les interroger.

2.0.1 Les sources secondaires

Une information est dite de source *secondaire* si elle est recueillie par des personnes ou des organisations dans des intentions autres que celles de la résolution du problème concerné.

- Les sources secondaires *internes* se trouvent dans les documents et communications associés à l'activité de l'organisation;
- Les sources *externes* sont les documents accessibles, établis ou délivrés par ailleurs.

Les raisons de consultation de sources secondaires sont les suivantes:

- Il se peut que la collecte d'information primaire ne soit plus nécessaire, si l'information secondaire suffit pour traiter le problème;
- Les coûts de la recherche secondaire sont inférieurs à ceux de la recherche primaire.

La recherche secondaire a d'importantes contributions supplémentaires:

- Mieux définir et comprendre le problème, et formuler des hypothèses;
- Planifier la collecte de données primaires;
- Définir la population et l'échantillon dans une collecte primaire.

2.0.2 Expérimentation "naturelle"

Une expérimentation "naturelle" est telle que l'investigateur n'y intervient que dans la mesure requise pour l'appréciation des résultats. Les approches en sont:

- Les approches *statiques* (dites en anglais "cross-sections"), où les données et observations concernent un même ensemble échantillonné à une période spécifique;
- Les *chroniques* et les *panels*, lesquels observent un même échantillon mais à des repères temporels successifs;
- L'approche des *tendances* qui implique l'obtention de données issues d'échantillons différents de la même population, prélevés à des repères temporels successifs.

2.0.3 L'expérimentation contrôlée

Dans le cas de l'expérimentation contrôlée, l'investigateur peut couvrir toute la procédure: organiser, structurer, mesurer, interpréter, conclure. Un exemple classique en est l'étude des "plans expérimentaux".

2.0.4 Le système abstrait

La source d'information qualifiée ici de "système abstrait" a pour output les résultats d'analyses conduites sur des modèles symboliques et formalisés – par exemple un modèle économétrique, ou un simulateur. Il est permis, en forçant quelque peu son destin, de considérer un modèle ou un système abstrait comme un "répondant". On pourrait donc "interroger" un système abstrait, en lui injectant un stimulus et puis en recueillant sa réponse. La manipulation de modèles économétriques, la simulation, ainsi que certains modèles d'aide à la décision sont des moyens qui peuvent fournir de telles contributions.

2.1 Le processus de communication étendu

Sur la Figure **2**, la ligne de communication de la Figure **1** devient une *chaîne d'information*. L'*ordre* d'une chaîne est le nombre de stations de la séquence initiée par le demandeur jusqu'à l'interprétation du message final par le récepteur. Cette notion de chaîne permet d'étudier les erreurs et les pertes d'information sous forme modulaire.

Une telle approche, initiée en théorie des communications, a été généralisée vers les "chaînes" d'information, puis la "traçabilité" de produits. Le "parcours du patient" dans un hôpital, dans l'exposé sur les «Modèles de processus», en est une forme récente.

La chaîne implique les entités et opérations suivantes.

a Les agents et leurs intentions

- Le *récepteur* **R**, client de l'information, qui peut être un groupe;
- L'*émetteur* **E** (chercheur-informateur), qui peut aussi être multiple;
- Les *sources* d'information, citées ci-dessus, dont les répondants (E').

b Les liens entre les ensembles et les signaux associés

1) Le *modèle* mental de la situation d'intention, l'*explicitation* de l'objet de l'information demandée (sur quoi?), et l'objectif poursuivi (pourquoi?). Il en résulte le message d'entrée;

2) Le *codage*: formulation de l'objet et de l'objectif de l'information. Il en résulte le signal "a", expression formelle du message;

3) La *communication* du signal d'entrée à l'émetteur, lequel recevra un signal de sortie identique dans la mesure où des obstacles et perturbations de nature physique ou humaine n'auront pas altéré le signal;

4) Le *décodage*: l'interprétation du signal b tel que celui-ci est perçu par l'émetteur E;

5) Le *modèle* mental des objectifs de l'émetteur et la c*onfrontation* du message avec son expression;

6) La *sélection* des sources et *appel* aux sources impersonnelles;

7) Le *retour* des observations;

8) L'*appel* aux répondants et départ d'un sous-cycle d'information;

9) Le *retour* des réponses issues de ce sous-cycle;

12) L'*élaboration* du message a_1'. L'émetteur fait la synthèse, l'agrégation ou le filtrage;

13) Le *codage*: transformation du message en signal, en fonction de la nature de la ligne de communication et de la réceptivité souhaitée de la part de R;

14) Le *transfert* de a' et b' compte tenu des obstacles et perturbations;

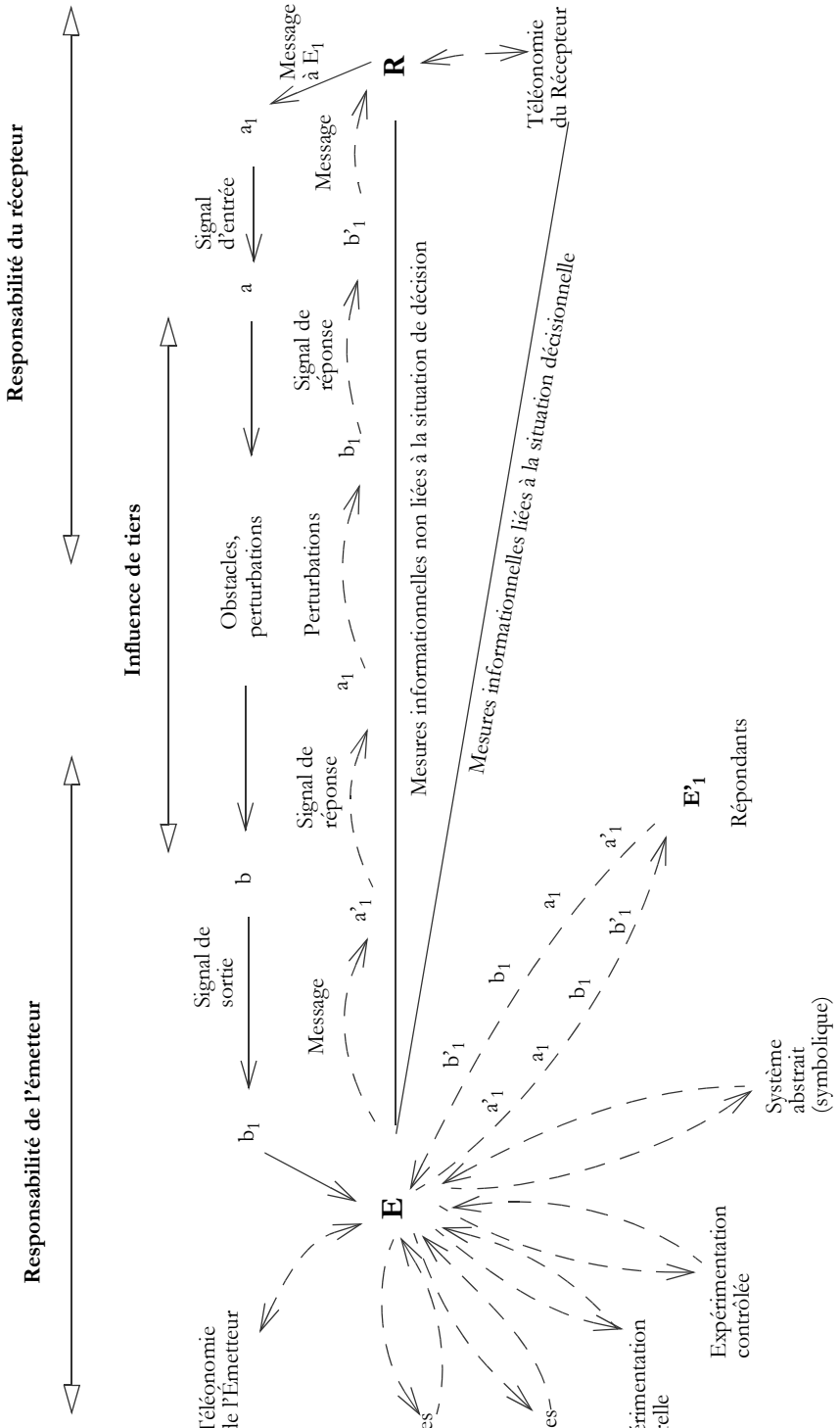
15) Le *décodage* du signal;

16) La *confrontation* par le récepteur du message de sortie b'_1 avec les objectifs d'information d'abord, avec la situation d'intention ensuite.

Le schéma proposé à la Figure **2** est le schéma étendu de cette communication entre un émetteur ("chercheur-investigateur") et un récepteur ("manager"). Il contribue au repérage et à la distinction des mesures d'information ainsi qu'à celui des erreurs qui perturbent et gênent le processus d'information. Il offre aussi l'avantage de lier le processus aux intentions du récepteur et aux intentions "perturbatrices" de l'émetteur et des tiers, et son étude doit permettre au récepteur d'améliorer la fiabilité de son processus d'information.

Certaines mesures d'information, comme il a été montré dans l'exposé «L'Information et sa thermodynamique», aident d'ailleurs à en évaluer la fiabilité, mais seulement du point de vue des correspondances entre un émetteur et un récepteur.

Figure 2. Processus de communication étendu



2.2 Les sources d'erreurs et... les erreurs dans les sources

La Figure **2** situe des "dysfonctions", erreurs et perturbations potentielles. Celles-ci vont être précisées en commençant par une liste de biais et erreurs à considérer dans les sources de données statistiques, selon une énumération initialement due à O. MORGENSTERN [1972]. Elles concernent aussi bien les données que les observations et informations, bien que ces trois notions soient différentes.

Dans la même veine de la non-fiabilité, qui gangrène les "systèmes d'information" sans épargner les plus chers, on citera et développera ensuite des problèmes associés à l'information issue de "répondants". En particulier, elle donne (avec COCHRAN [1963] et DE BRUYN et EK [1974]) une estimation statistique de la portée d'erreurs de non-réponse et une procédure qui tient compte des "non-réponses" et des "sans opinion" appliquée à l'appréciation des patients dans le domaine hospitalier.

2.2.1 Sources d'erreurs des statistiques économiques et de gestion

- Absence d'expériences organisées;
- Informations soustraites, mensonges;
- Manque de formation des observateurs;
- Erreurs dues aux questionnaires;
- Définitions ou classifications insuffisantes;
- Erreurs dues aux instruments;
- Le facteur temps: repérage inadéquat et retards de saisie;
- Observation de phénomènes uniques;
- Illusion de la précision;
- Inapplicabilité des concepts utilisés;
- Statistiques fonctionnellement fausses et statistiques dénuées de signification, notamment en comptabilité d'entreprise et comptabilité nationale.

3 L'information issue des répondants

3.1 Objet de l'information

L'*objet* de l'information issue de répondants est le domaine du témoignage et de l'opinion, et donc relève de l'appréciation. On peut y distinguer:

- L'information par *communication* (donc interactive) avec les *répondants*. Ce sont des agents supposés capables de fournir des réponses à une enquête informative. Remarquons que cette voie, par ailleurs efficace, présente des obstacles et biais dont il sera fait mention dans un bref instant;
- L'information par *observation* des comportements présents ou passés des répondants et de leurs résultats.

3.2 Erreurs et biais des données, observations et opinions

Aux informations issues de répondants correspondent les erreurs suivantes:

- Erreurs d'échantillonnage;
- Biais de réponse (Inexactitude, ambiguïté);
- Biais de "non-réponse".

3.2.1 Erreurs d'échantillonnage

De telles erreurs se produisent parce que la population concernée n'est pas représentée exhaustivement par l'échantillonnage. Il en résulte que l'échantillon choisi n'est pas tout à fait représentatif, eu égard aux caractéristiques de la population dont il est extrait.

3.2.2 Biais de réponse en recherche d'opinion

Les biais de réponse en recherche d'opinion, prélevés sur des épaves de publications de marketing, peuvent être résumés selon la liste suivante:

- Le biais d'ambiguïté;
- Le biais de compétence;
- Le biais de téléonomie;
- Le biais de caractère;
- Le biais dit "de métonymie";
- Le biais dit de dissonance cognitive;
- L'erreur de réponse et de... non réponse.

a Le biais d'ambiguïté

La multiplicité des points de vue des individus rend leur opinion contingente au contexte ou à la situation présente. Par exemple, un jugement issu du public sur une implantation hospitalière belge est "on a un bel hôpital" et, par le même contribuable, "un luxe et un gaspillage scandaleux".

b Le biais de compétence

Le biais de compétence touche le répondant à la fois juge et partie relativement à la question posée. Ainsi, en milieu hospitalier, il va de soi que c'est au corps médical qu'on doit demander d'évaluer l'efficacité thérapeutique. Mais pour conforter son opinion, le corps médical soucieux de qualité fait appel à la *revue par les pairs*, l'*audit médical*, le *technology assessment*, c'est-à-dire les approches d'étude *collégiales* de la qualité des soins médicaux, appliquant, cela va de soi, le dicton de Monseigneur BERKELEY (Évêque, vers 1750) disant que «L'observateur gagne son objectivité par le fait d'être observé».

Il recherche de cette façon une validation, légitime et recommandée, par *caution d'experts*, car le seul danger éventuel est que ce gain en compétence collégiale puisse devenir un biais corporatiste.

Le biais de non-compétence intervient, lui, presque par définition, quand on interroge des individus sur ce qui n'est pas leur domaine. Si ce sont des patients que l'on interroge sur la qualité des soins, ils présentent un biais de non-compétence en répondant plus souvent à la question "a-t-on été *aux petits soins* pour vous"?

c Le biais de téléonomie

Les gens répondent selon la mesure dans laquelle leurs propres attentes sont satisfaites. Le biais de téléonomie est donc celui par lequel on apprend non pas "ce qui est bien" (ou pas bien) mais ce qui *plaît* ou non à tel répondant selon ses besoins et ses attentes.

d Le biais de caractère

Les mêmes personnes sont contentes ou mécontentes de tout.

e Le biais dit "de métonymie"

La *métonymie* a été en systémographie spécifiée comme un type de *métaphore* par laquelle on dit une partie pour évoquer un tout. Appliquée ici, elle veut dire qu'une seule cause, ou des causes partielles, de mécontentement peut conduire à une attitude de rejet sur tous les autres points.

f Le biais dit "de dissonance cognitive"

Un répondant atténue sa critique négative quand l'objet considéré a été choisi par lui-même, et il n'est pas enclin à dénigrer sa propre décision.

g Le biais de rumeur

L'opinion peut être emportée par celle que répand la rumeur, ou encore s'opposer à celle-ci par une teigneuse vanité d'individualisme; dans ces deux, cas l'opinion est biaisée par défaut d'indépendance.

Ces différents biais seront repris dans l'exposé sur «L'Évaluation», précisément dans le point de vue *appréciatif*, et seront illustrés dans l'exposé sur «L'Évaluation hospitalière». Ce qui est irritant, et une lâcheté du chercheur, est de voir dans un rapport: «Le lecteur est prié de considérer ces résultats avec prudence». Organiser un recueil d'opinion pour éviter de tels biais, ou alléger leurs effets, est affaire de spécialiste. Si on ne peut se faire trop d'illusions sur la pureté des jugements, il est toujours bon d'au moins connaître les pièges, y compris la bêtise. Il est vrai qu'il faut un peu de tolérance dans ce domaine, mais si un résultat est invalide il faut le jeter.

3.3 La formulations des questions et le traitement des réponses

Les questionnaires soumis aux répondants peuvent présenter différentes formulations, dont des exemples figurent ci-après dans le contexte de l'appréciation hospitalière. Ce qui intéresse le systémicien, c'est la restriction de degrés de liberté de la réponse, permettant de construire des indicateurs et homogénéiser le traitement quel que soit l'objet de l'appréciation.

Des formulations de questions posées ici sont les suivantes:

- Commentaire libre;
- Question ouverte;
- Question fermée;
- Non-orientée;
- Orientée;
- Orientée et codée.

Une question est *ouverte* si l'expression de la réponse est libre: le répondant ne "coche" pas une proposition déjà rédigée; sinon, elle est qualifiée de *fermée*. On voit que les formulations de la liste ci-dessus sont rangées, dans le sens du plus libre au plus contraint. On s'attend alors à ce que la méthode de traitement puisse être d'autant plus automatisée que la contrainte est serrée; effectivement la méthode formelle donnant *des scores* décrite à la section **3** ne s'appliquera qu'au cas de question orientée et codée. Les exemples suivants, de l'appréciation hospitalière, illustrent les différentes expressions proposées.

a Commentaire libre

Dans la section "Que pensez-vous de la propreté générale de l'hôpital?", la question

Avez-vous des suggestions à faire? appelle un commentaire libre, par exemple:

On lave l'évier avec le gant de toilette du malade!

b Question ouverte

Un exemple de question du type *ouverte* est:

À votre arrivée, comment jugez-vous la signalisation pour vous guider?

c **Question fermée, non-orientée** (cochez le "O")

Que pensez-vous de l'allure générale de l'hôpital, de sa conception?

Vieillot	Moderne	Rébarbatif	Accueillant	Sécurisant
O	O	O	O	O

Pour ces catégories (libre, ouverte et fermée par une liste) l'interprétation se fait par une approche non formalisée, mais les suivantes se prêtent à l'élaboration d'un "score".

d Question fermée, orientée

Au service des urgences: *Comment sont les relations avec les ambulanciers?*

Agréables	Correctes	Sympa	Froides	Désagréables
O	O	O	O	O

L'"orientation" est de gauche à droite; la question est *ordonnée* par le fait que l'appréciation est de moins en moins gratifiante.

e Question fermée, codée

a. Question codée à deux niveaux, trois modalités; par exemple:

Saviez-vous avant cette admission que ce service existait à l'hôpital?

Oui	Non	Sans opinion (S.O)
O	O	O

ou: *êtes-vous favorable à la consultation le soir?*

Pour	S.O	Contre
O	O	O

b. Question codée à trois niveaux (quatre modalités)

Très content	Content	Mécontent	Sans Opinion
O	O	O	O

Cette formulation présente un biais de *déplacement* (vers le "content"). D'autre part les modalités de réponses y sont exprimées en terme de *satisfaction* (content, mécontent) et non en terme d'*appréciation* (bon, agréable, facile). Dès lors les réponses y dépendent plus de l'état mental du juge que des qualités de l'objet jugé; ainsi un patient peut apprécier positivement la qualité des soins prodigués en radiologie mais être mécontent d'avoir attendu deux heures dans le couloir.

c. Question orientée codée à quatre niveaux (5 modalités)

Mécontent	Pas bon	Satisfaisant	Très bien	Sans opinion
O	O	O	O	O

Ici la *satisfaction* et le *jugement* sont "combinés" en alternant les réponses relatives au *sujet* (content...) et à l'*objet* (bien...). Cette façon de faire est recommandée malgré ces deux composantes car elle permet une grande simplification par une présentation homogène de la formulation pour presque toutes les questions. Le traitement et la construction d'un indicateur y sera plus rapide et surtout moins biaisée par la formulation même de la réponse imposée par l'enquêteur.

Par contre, une formulation à 5 niveaux, du type:

Très mauvais	Mauvais	Neutre	Bon	Très bon

n'est pas conseillée, car le "neutre" est une perte d'information et est certainement récepteur des "sans opinion" que la statistique ne lui donne pas le droit de remplacer.

3.3.1 Les options d'exploitation des réponses

L'exploitation de résultats d'enquête peut être faite selon cinq approches, qui peuvent être complémentaires.

1 Disposer d'une personne qui les lit, les classe, découvre et communique les points importants et récurrents qu'il est utile de mettre en évidence.

   Ceci est édifiant et irremplaçable par une procédure automatique, mais demande du temps (même aidé par une base de données bien rédigée) et de l'objectivité. De plus, il n'est guère facile de diagnostiquer l'*évolution* de certains aspects. Aussi, dans le monde plus fermé du *comptage* de réponses, les approches suivantes peuvent être utilisées;

2 Éditer les tableaux d'effectifs et pourcentages de réponses à chaque question dans l'ordre d'interrogation et aider le lecteur à trouver les points remarquables;

3 Traiter les tableaux de fréquences variables-réponse par analyse statistique multivariée et en dégager les groupements et les facteurs principaux;

4 Faire appel aux *modèles linéaires généralisés* pour estimer la signification statistique des paramètres d'influence sur les variables;

5 Construire une méthode et une présentation spécifiques pour tenter de répondre aux questions essentielles posées par le "client du problème" en conservant l'information de base et en construisant des indicateurs aidant à une lecture rapide et pertinente selon le thème d'investigation choisi.

L'approche (2) est une phase obligée, mais, s'arrêtant là, elle aurait la lâcheté de confier au lecteur le soin de l'interprétation de la masse des résultats. Les approches (3) et (4) sont élégantes et scientifiques, mais, dans le cas présent, perdent les réponses libres et s'accommodent mal de la diversité des expressions utilisées pour aider les répondants à exprimer leur opinion.

C'est l'approche (5) qui sera élue ici. Il en résultera que l'information contenue se trouve directement lisible sur les tableaux de synthèse, spécialement reconstruits à cet effet. Ces tableaux ne concernent que les questions "fermées" (donc à choix multiples). Les réponses ouvertes, les commentaires et ce qu'a appris l'enquêteur sont à fournir dans une section séparée du rapport et dans ses conclusions.

3.4 Le problème posé par la non-réponse

3.4.1 L'erreur de non-réponse

Une erreur de "non-réponse" se présente lorsqu'un individu est inclus dans un échantillon mais n'a pu être atteint par l'enquête. L'inconvénient d'une telle situation est accru par le fait que la direction de l'erreur est inconnue et que son amplitude est malaisée à estimer. Il est fréquent aussi que des répondants ne s'expriment pas sur certains thèmes, ce qui rend le message "sans opinion". Trois interprétations peuvent être avancées à ce sujet:

- Soit le répondant a été témoin de l'événement qu'on lui demande de juger mais, faute d'opinion, ne peut pas s'exprimer: il est indécis; ceci est donc une *modalité*, à considérer au même titre que les autres niveaux de réponse;
- Soit le répondant n'a pas été témoin de l'événement et, faute d'information, ne peut s'exprimer;
- Soit le répondant ne désire pas s'exprimer.

Un même vide peut dès lors trouver son origine dans deux faits différents: l'attitude et la non-information. Il convient donc de les distinguer dans la méthode d'interprétation: la première concerne la *variance* de la réponse (incertitude plus ou moins marquée) alors que la seconde concerne le *montant* d'information, ce qui se traduit par la *puissance d'opinion*. Celle-ci affecte d'une part la *crédibilité* du résultat de l'analyse et d'autre part la *notoriété* relative au thème de la question citée.

3.4.2 Les deux attitudes statistiques

- La première attitude est de ne faire état que des jugements de ceux qui ont répondu (ce qui va de soi) mais précisément le fait de répondre ou non peut être associé à une obédience ou à une orientation d'opinion, ce qui contredit évidemment la neutralité de l'échantillon;
- La deuxième version est statistique: le comptage de la répartition des réponses (par exemple dans un questionnaire à choix multiples) mentionne les "sans-réponse" sans en tirer parti. Ceci conduit à un problème d'inférence statistique, en particulier pour la définition des intervalles de confiance.

3.4.3 Le danger d'interprétation et de lecture hâtive

Les enquêtes d'appréciation par des patients (dans «L'Évaluation hospitalière») montrent un grand nombre de "sans réponse", ou "sans opinion". Ainsi, peu de gens savent si l'endoscopie (la quoi?) est meilleure au CHX, ou si on fume dans la salle d'attente du dentiste pendant les accouchements à l'étage du-dessus, ou encore si la scintigraphie (hein??) est douloureuse. De plus, la non-réponse ou le "je ne sais pas" est aussi une information, et il faut trouver le moyen d'en tenir compte.

Les interprétations d'enquêtes sont donc biaisées par ce problème. Il implique aussi que la lecture directe des pourcentages de telle ou telle réponse est très dangereuse lorsqu'il s'agit de nombre de réponses et de notoriété. Ainsi, soit le cas d'un échantillon de 50 personnes dont à la question ouverte sur le choix de l'hôpital 5 répondent qu'elles sont venues "pour la bonne réputation du médecin", les 45 autres ne citant pas cet argument. Une caricature d'interprétation aberrante serait que 90% des gens pensent que le médecin n'a pas bonne réputation! Dans ce cas, on a par contre un beau score positif (l'argument favorable a été cité cinq fois) mais la *puissance* d'opinion est faible, car elle n'est affectée que d'un paramètre de notoriété de 5/45.

4 Une méthodologie d'analyse tenant compte des "sans réponse"

Le manque de contributions théoriques au problème des sans-réponse dans les enquêtes, pourtant courant, donne un stress grelottant au statisticien de garde. On repère que la validité de la majorité des votes en fonction du nombre d'abstentions au Conseil Athénien préoccupait déjà ARISTOTE; on le retrouve un peu plus tard dans COCHRAN (1963), puis dans DE BRUYN et EK (1974). Enfin on a pu puiser ici dans H.S. KONIJN (1973), malgré ses notations inextricables, une inspiration pour un indicateur à trois modalités. Ces références sont précisées en fin de cet exposé.

L'explication des indicateurs construits ici sera brève mais illustrée sur tableur, montrant les résultats sur des cas typiques simplifiés et utilisant le nombre de réponses (de l'échantillon) situées dans chaque modalité de la liste présentée au répondant.

4.1 Méthode pour le type non-orienté, à deux niveaux

4.1.1 Un indicateur de dominance

Les notations utilisées sont les suivantes:

- N : l'effectif de l'échantillon, soit N = A+O+B+S
- A : le nombre de réponses "A";
- O : le nombre de Sans Opinion;
- B : le nombre de réponses "B";
- S : le nombre de "Sans Réponse" (= N - R);
- R = A+O+B : le nombre de réponses;
- E = (A + B) : le nombre d'opinions exprimées soit pour "A" soit pour "B";
- C = A / (A+B): le "rapport d'acceptation".

Soit à présent une question à deux niveaux, disons A et B, et une modalité "O" pour "Sans Opinion", et la possibilité de Sans Réponse, "S"; pour chaque modalité on dispose du nombre de réponse. Les données pour une telle question se présentent, pour un échantillon de taille N, selon la ligne (1) suivante:

Modalités	A	O	B	S
(1) Effectifs	40	25	10	20

Le *rapport d'acceptation* est construit par C= A / (A+B). La réponse "A" est dominante dans la mesure où C est supérieur à 1/2, c'est-à-dire si la distribution est plus en faveur de "A" que la référence "neutre" {0,5; 0; 0,5}. C'est le cas ci-après pour (3), où C(3)=0,80 alors que la référence (2) donne 0,5:

Modalités:	A	O	B
(2)	50	0	50

par exemple:	A	O	B
(3)	40	25	10

L'expression de la dominance par C = A / (A+B) est toujours préférable à A/B parce que le résultat est plus clair (compris entre 0 et 1), que la variance est plus faible et surtout parce qu'on affaiblit le biais appelé "l'effet coussin", qui est issu du rapport d'erreurs de mesure situées à la fois au numérateur et au dénominateur.

Un premier indicateur peut être construit par C(3)-C(2), soit ici 0,80 - 0,50 = 0,30.Toutefois ce *même* résultat pourrait être obtenu par la répartition (4) ci-dessous, car C(4) = 56 / (56+14) = 0,80 également:

	A	O	B
(4)	56	5	14

Cependant, à la vue d'un lecteur, le cas (4) paraît plus "significatif", en faveur de A. Effectivement, ce qui le différencie de (3) c'est la *variance*, laquelle, pour une telle distribution, augmente en fonction de la proportion de Sans Opinion.

4.1.2 La variance de l'indicateur de dominance

Soit v la variance de l'espérance mathématique de "A" (par exemple "A" serait "oui", et "B" serait "non") pour cette distribution statistique {A;O;B,} de la famille multinomiale, et soit w l'écart-type. Elle est calculée comme suit:

$$v = [(O/R) / E] *C* (1-C)$$

4.1.3 La valeur d'acceptation

L'indicateur construit ici, appelé "valeur d'acceptation de A", est calculé comme suit:

> Le nombre de fois l'écart-type de l'espérance mathématique du rapport d'acceptation A/(A+B) que celui-ci s'écarte de sa médiane, laquelle est le "neutre" 0,5. Celui-ci correspond à {50%, 50%}.

Disposant de cela, la lecture est: l'opinion est en faveur de A si l'indicateur est positif (il est négatif en faveur de B), et ce d'autant plus que la valeur de l'indicateur est élevée. Comme l'écart-type est plus grand quand il y a plus de "Sans Opinion" (ce qui varie à chaque question, d'où ce sérieux problème théorique), l'indicateur est donc effectivement plus faible quand la participation à l'opinion est plus faible.

En effet, soit à comparer à présent les cas (3) et (4), c'est-à-dire:

	A:	O:	B:
(3)	40	25	10
(4)	56	5	14

Les valeurs obtenues sont respectivement:

(0,80 - 0,50) / 0,0326 = 9,20 pour (3), et

(0,80 - 0,50) / 0,0130 = 23,1 pour (4).

L'indicateur est ici près de 2,5 fois supérieur pour (4), lequel a d'ailleurs cinq fois moins de sans opinion. Il serait négatif si le rapport en faveur de "A" était inférieur à 0,50.

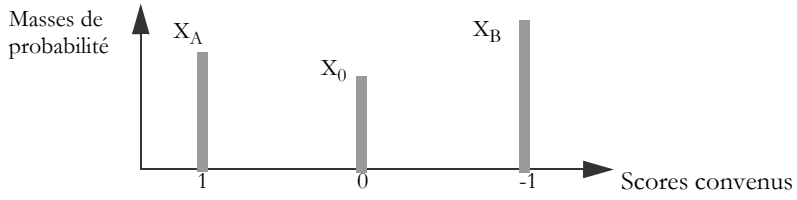
Les valeurs de l'indicateur pour ce type de question peuvent alors être rangées de la plus grande à la plus faible *par thème* sur le tableau de synthèse des résultats, et le lecteur n'a qu'à prendre ce tableau en mains pour voir immédiatement les dominantes de réponses.

Un cas particulier se présente quand toutes les réponses sont exclusivement "A" ou "B", donc il n'y a pas de Sans Opinion. Dans ce cas la variance est nulle, mais il suffit alors d'afficher le pourcentage d'opinions dépassant 50% en faveur de A ou B, et d'attirer l'attention du Lecteur sur cette valeur remarquable.

4.1.4 Score de discordance

Soit que A, O, B soient exprimés en pourcentage du nombre de réponses, respectivement X_A, X_O, X_B. Attribuons aux réponses de type A, O et B respectivement les valeurs indicatrices 1, 0, -1. Dans ce cas nous associons X_A, X_O, X_B à des masses de probabilité sur ces valeurs, conformément à la Figure **3**.

Figure 3. Masses de probabilité sur les indicateurs



L'espérance de cette distribution est $(X_A - X_B)$, et le second moment est $(X_A + X_B)$. Dès lors, le second moment centré est:

$$v = (X_A + X_B) - (X_A - X_B)^2.$$

Ce v est la *variance* d'une distribution de signes, appelée aussi indice de *disconformité*. Elle est minimale lorsque tous les participants donnent la même réponse, et maximale lorsque la répartition est [1/2; 1/2] entre les A et les B (ou les "oui" et les "non").

Un score de *dominance* facile et intéressant peut alors être construit par le rapport d, qui est la dominance absolue révisée par la disconformité:

$$d = \frac{X_A - X_B}{X_A + X_B - (X_A - X_B)^2}$$

- Lorsqu'il est plus élevé cela signifie une dominance à la fois plus marquée et moins discordante;
- Il est nul lorsque la répartition est égale;
- Il est négatif lorsque les "B" (ou les "non") dépassent les A (ou les "oui").

La confrontation de (3) et (4) ci-dessus utilisant "d" donne respectivement les valeurs de 0,789 et 0,904 en faveur de A, ce qui est un résultat très clair.

Une répartition de {99%, 1%, 0%} donnerait la valeur extrême de 1,00.

4.2 Méthode de traitement d'une réponse codée à plusieurs niveaux

Soit une question orientée codée à plusieurs niveaux, dont différents exemples de libellés et de pourcentages de réponses sur chaque modalité figurent ci-dessous.

Les modalités répondent aux remarques sur les biais formulées à la section **3**:

	Mécontent	Pas bon	Satisfaisant	Très bien	Sans.Op	
ou:	Désagréables	Froides	Sympa	Cordiales	Sans.Op.	
ou:	**"--"**	**"-"**	**"+"**	**"++"**	**"S"**	
(5) :	16	4	18	42	20	(R=0,8)
(6) :	2	10	40	8	40	(R=0,6)

Le problème posé n'est pas élémentaire:

- Déterminer si "(5) est meilleur que (6)" pour toute série de telles formulations;
- Construire un indicateur de jugement qui ait une valeur de *référence*:
  - permettant de *comparer* les questions et aussi les objets concernés,
  - dont l'évolution est lisible dans le temps,
  - si les enquêtes sont répétitives,
  - pour des échantillons de taille différente,
  - et pour des nombres "s" de "sans réponse" différents.

On désire au fond savoir si l'item concerné par cette question est jugé comme "bon" ou s'il est moins bien ou meilleur que "bon", et dans quelle mesure. À cette fin est construit un *vecteur flou* dont les éléments sont les valeurs de la fonction d'appartenance au sous-ensemble "bon"; celui-ci servira de référentiel et c'est la *distance à ce vecteur* qui en sera l'indicateur.

- Il est logique de proposer que ce vecteur "bon" ait des scores de valeur nulle pour les modalités "Très Mauvais" (ou "--") et "Mauvais", (ou "-").
- Les modalités favorables, telles "Bon" (ou "+"), "Très Bon" (ou "++") doivent avoir une valeur de référence positive.

Le vecteur retenu comme satisfaisant ces conditions est le suivant:

	"--"	"-"	"+"	"++"	S.O.
réf. "r":	0	0	0,66	0,33	s

Pour cinq niveaux, dont un neutre ("="), on propose la référence "bon" suivante en %:

"--"	"-"	"="	"+"	"++"S.O.
0	0	0,33	0,33	0,33s

Les deux cas de "référence" présentent une dissymétrie, qui vise à corriger un biais des répondants par lequel ceux-ci seraient enclins à considérer le "bon", que l'on espère atteint par tout service, comme le neutre (ou la médiane) de la répartition.

La procédure créée ici est de calculer la *distance de Hamming* du vecteur de réponses à la frontière du sous-ensemble flou de référence, que l'on peut imaginer visuellement comme le nuage "bon". Cette distance, qui exprime les écarts entre les éléments de deux vecteurs, est un être mathématique correct, ayant les propriétés d'être positive, réflexive, symétrique et satisfaisant l'inégalité de transitivité, dite inégalité "triangulaire".

Cependant, il est souhaité de plus:

- D'obtenir une valeur négative lorsque la répartition est moins bien que la référence "bon";
- De donner une pondération supérieure aux opinions plus marquées, comme "Très Mauvais" et Très Bon".

En conséquence, chacun des écarts est pondéré selon les poids indiqués ci-dessous:

	Exécrables,	Shi_Han_Te,	Froides,	Sympa,	Très sympa
Poids:	-2	-1	(0)	+1	+2

Faisant cela, nous perdons la qualification de *distance* car certaines des propriétés mathématiques nécessaires en sont perdues, mais nous obtenons un *écart* qui donne une grande pertinence de lecture.

Pour l'exemple, les répartitions (5) sont reprises ci-dessous et amenées à 100%, mais les Sans Opinion seront traités en **4.3** par une démarche complémentaire.

(5) :	0,16	0,04	0,18	0,42	0,20
(5)':	0,20	0,05	0,225	0,525	0,00

Appliqué aux réponses (5)', l'écart donne **- 0,675**.

Aux fins de comparaison, il donne pour les réponses (5) la valeur de **- 0,633**, ce qui n'a effectivement "pas l'air" très différent. On a pourtant 42% de très bon en (5) contre 8% en (6) mais ce qui coûte cher à (5), ce sont ses 16% de "très mauvais", ainsi que le fait que le taux de réponse est plus élevé, soit R=0,8 contre R=0,6. Ceci montre bien la puissance de la méthode.

4.3 La puissance d'opinion et l'indicateur "Score"

Admettre que le taux de réponses réel est un paramètre déterminant dans le calcul de l'indicateur appréciatif de performance revient à refuser l'hypothèse selon laquelle les non répondants auraient répondu de la même manière que ceux qui se sont exprimés. Il serait illogique d'assimiler les non-réponses à des répondants effectifs; par contre, elles peuvent être considérées comme une perte d'information par rapport à la situation certaine où tous les répondants se seraient exprimés. Dès lors, le problème est double:

- D'une part, évaluer la perte d'information due aux non-réponses par rapport à la situation certaine (taux de réponses réel de 100%);
- D'autre part, exprimer cette perte d'information en un modérateur capable d'affaiblir l'écart et de la sorte obtenir l'indicateur "Score".

Ce modérateur est utilisé non pas pour diminuer la qualité de "meilleur" ou "moins bien" que "bon", mais pour tenir compte du fait que tous les répondants *ne se sont pas exprimés* sur certains thèmes d'appréciation. Comme le problème est celui de la notoriété, on acceptera le rôle de ce paramètre en tant que *puissance d'opinion*.

Deux approches de la *modération* peuvent être considérées: l'affaiblissement "linéaire" et l'affaiblissement "exponentiel".

a L'affaiblissement linéaire

Cette méthode considère simplement la perte d'information comme une fonction linéaire du taux de réponses réel. Un tel modérateur pourrait n'être en fait que le taux de réponses réel "p". Ainsi un écart de 0,521 avec un taux de 75% donnerait au "Score" la valeur de 0,521 $*$ 0,75, soit 0,391.

Cette méthode a cependant le défaut d'accorder le même poids à un répondant effectif qu'à un non répondant, ce qui sera modifié par l'affaiblissement "exponentiel".

b L'affaiblissement exponentiel

Cette méthode se base sur la théorie de l'information de SHANNON et WEAVER. Celle-ci a montré à la section **3** que le contenu en information d'un message annonçant un événement certain (de probabilité 1) par rapport à la probabilité d'apparition a priori "p" octroyée par un individu à cet événement peut être mesuré par:

$$-p\,\mathbf{l}_{g2}(p)$$

Ceci donne le gain en information dû au fait de savoir que l'événement se produira de manière certaine alors qu'on ne lui prédisait qu'une probabilité d'apparition de p.

Le cas de non-réponse cependant demande l'inverse: on a affaire à un gain en information négatif, c'est-à-dire à une perte d'information, depuis la situation "certaine" (donc un taux de réponses "fictif" de 100%, assimilable à une probabilité a priori 1) vers le taux de réponses réel "p" (assimilé à la probabilité à posteriori). Dès lors, la relation devient trivialement:

$$\mathbf{l}_{g2}(p/1) = \mathbf{l}_{g2}(p) - \mathbf{l}_{g2}(1) = \mathbf{l}_{g2}(p)$$

Exprimée en bits (Binary digits), cette relation exprime la *perte d'information* du taux de réponses réel "p" par rapport à la situation où tous les répondants se seraient exprimés. Le premier problème est donc résolu. Cependant, cette grandeur est d'une autre nature que l'indicateur "Écart" qui le précède, et par conséquent ne peut être un paramètre additif ou multiplicatif de cet indicateur.

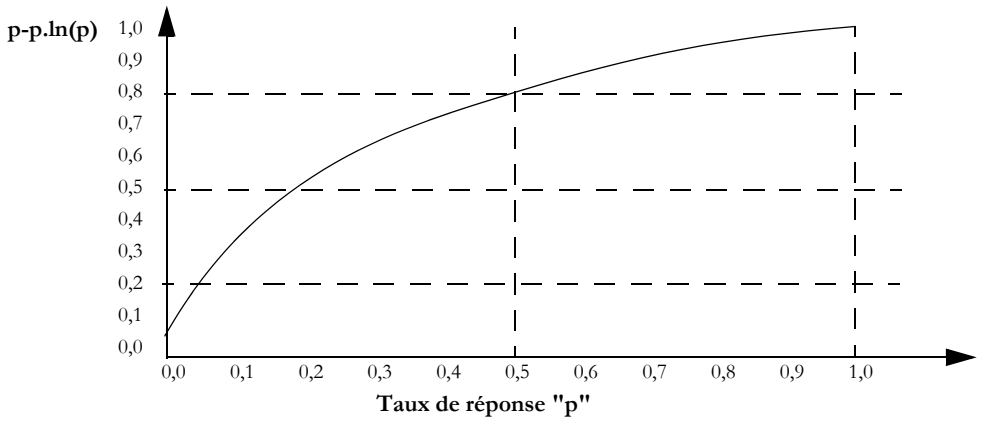
Dès lors, pour exprimer cette mesure de la perte d'information en un modérateur de l'écart, on calcule le rapport de la perte cumulée d'information jusqu'au taux "p" à la mesure de la perte cumulée maximale, c'est-à-dire pour le taux "1".

La fonction de modération proposée ici est le rapport de la valeur de l'intégrale de la fonction jusqu'à p (le taux de réponses réel), à la valeur de l'intégrale jusqu'à 1:

$$\text{Infor} = \frac{\int_0^p \mathbf{l}g_2(r)dr}{\int_0^1 \mathbf{l}g_2(p)dp}$$

Ce rapport vaut $p - p.\mathbf{l}_{g2}(p)$ et est présenté à la Figure **4**.

Figure 4. Fonction de modération selon le taux de réponse



Cette fonction de modération possède des propriétés intéressantes:

- Elle accorde plus d'importance aux répondants effectifs qu'aux non-réponses. En effet, pour un taux de réponses réel disons de 50%, la modération est de 0,85, donc diminue de 15% la notoriété de l'écart. Plus formellement, sa dérivée vaut - $\mathbf{l}_g(p)$;
- Elle conserve le neutre 0;
- L'interprétation n'est pas altérée par cette fonction: une valeur positive est toujours meilleure que "bon" tandis qu'une valeur négative est moins bonne, et d'autant moins bonne qu'elle est plus basse.

En terme marginal, la perte d'un répondant dans le cas de taux de réponses réel faible est plus fortement pénalisée par la fonction de modération que dans le cas d'un taux de réponses élevé, ce qui se voit sur la Figure **4**. Donc, plus le taux de réponses est faible (plus les répondants se raréfient), plus la perte d'un de ceux-ci est considérée comme une diminution de la puissance d'opinion.

Cette méthode d'élaboration des indicateurs "Écart" et "Score" ne relève pas de la statistique paramétrique, et dès lors, ne permet pas une inférence fondée sur ses propriétés.

5 Les mesures d'information liées à la situation téléonomique

5.1 Rôle de l'information dans une situation téléonomique

Selon ACKOFF et EMERY [1972], on peut dire d'un individu (I) qu'il est dans une situation de décision dans un environnement (**Ω**) si un ensemble de conditions sont satisfaites. Celles-ci définissent formellement une problématique de choix à options discrètes multiples, dans une formulation devenue depuis lors le "classique" de la théorie des décisions.

Ces conditions sont:

- Il y a dans **Ω** au moins deux lignes d'action exclusives disponibles; donc m > 2;
- Pour deux au moins des lignes d'action disponibles dans N, la probabilité de choix de l'individu est supérieure à zéro; elles sont alors appelées lignes d'*action potentielles*;
- Il y a dans l'ensemble des résultats (défini comme exclusif et exhaustif) un résultat (R_a), pour lequel deux des lignes d'action potentielles ont quelque efficience.

  Donc: $E_{1a} > 0$ et $E_{2a} > 0$; de plus: $E_{1a} \neq E_{2a}$;

- Le résultat pour lequel la condition précédente est satisfaite a une valeur pour I;

  Donc: $\boldsymbol{V}(a) > 0$.

Cette définition peut être résumée moins techniquement comme suit: on peut dire d'un individu qu'il est dans une situation d'intention et de décision s'il désire quelque chose et qu'il dispose de plusieurs voies d'efficience inégale pour s'efforcer de l'obtenir.

Dans ce contexte, on peut situer la proposition D'ACKOFF selon laquelle l'information est le montant de choix potentiels de lignes d'action que possède un individu. Selon la même source, l'action d'informer consiste alors à convertir des choix disponibles en des choix "potentiels". Ceci est compatible alors avec la proposition de W. CHURCHMAN [1961]: l'"Information" se réfère à l'*expérience enregistrée qui est utile à la prise de décision*.

Par contre l'acception courante de "l'expérience enregistrée qui réduit le *degré d'incertitude* dans la prise de décision" n'est pas nécessairement correcte en *théorie* de l'information: les mesures d'information de la section **1** ont montré en effet que l'information apportée par un message peut accroître l'incertitude du fait que la dispersion des probabilités a priori peut augmenter de par la réception de ce message.

Dans le cas d'une situation téléonomique, caractérisée par le fait qu'il y a un ou plusieurs agents munis d'une *intention*, les trois champs d'incertitude cités à la section **2** sont donc effectivement considérés, avant de prendre en compte les coûts de l'information et de la communication:

- L'espace probabiliste (ou de crédibilité);
- Les conséquences potentielles;
- L'évaluation des conséquences (la "valeur", au sens d'ACKOFF).

5.2 Les regrets et les risques

Par le calcul des espérances de valeur dans l'espace de probabilité des événements (ou "états de la nature"), on peut évaluer les risques et regrets associés à chacune des lignes d'actions. Comme l'état informatif du preneur de décision varie selon les messages reçus, on peut calculer le risque avant le message ("a priori") et les risques après le message ("a posteriori"). C'est la variation du risque qui mesure l'espérance de valeur du message [COLSON & DE BRUYN, 1973]. On obtient de la sorte les concepts classiques suivants:

- EVIE ("Espérance de Valeur de l'Information Expérimentale"), donc l'espérance de gain en valeur brute;
- ENVIE ("Espérance Nette de Valeur de l'Information Expérimentale") en valeur nette. Celle-ci suppose que l'on a pu faire le calcul du coût (ou de l'espérance de coût) de l'ensemble du système de communication;
- Dans le cas de message direct, on obtient l'EVIP ("Espérance de Valeur de l'Information Parfaite") si le mécanisme générateur des données est déterministe. Cette démarche est l'*Analyse Pré-postérieure*, dont une version formelle est située dans l'exposé «Vers le décideur artificiel».

5.3 La probabilité d'inversion de la décision

Le calcul de la probabilité d'inversion de la décision présuppose qu'un choix a été fait parmi les critères de valeurs, ceux-ci déterminant la fonction de décision, donc les conditions d'optimalité d'une décision. On peut montrer (COLSON & De BRUYN [1973]) que la probabilité d'inversion est la somme (ou l'intégrale, en continu) des probabilités des stimuli d'inversion. Un stimulus d'inversion est défini comme toute issue d'un message qui oblige à changer de décision sous la condition du critère de choix adopté et sans compter le coût d'inversion. Une application opérationnelle de cette approche et de ces mesures peut être lue dans COLSON et DE BRUYN [1970, 1973, 1989] et concerne la gestion d'un portefeuille de titres boursiers.

En conclusion, on est encore loin ici d'un modèle de gestion du processus d'information qui comporterait ses propres mesures et ses propres critères de performance. Cependant, un parcours utile des réseaux d'information, dans une optique très générale, figure dans WEISMAN [1972] ainsi que dans MUCHIELLI [1973].

6 L'information dans un processus bayesien de gestion

6.1 Liaison information-décision

La littérature concernant la gestion livre couramment des propositions de schémas de liaison de l'information à la décision qui se ramènent au type suivant:

Informations → modèle → réflexion → critères → décision → action → résultat → contrôle

Une autre présentation est celle de S. EILON (1969) qui suggère:

> Information → analyse → mesures de performance → modèle → stratégies → prévisions des résultats → critères de choix → résolution

Chacun peut d'ailleurs proposer sa propre séquence; toutefois ces séquences ne sont pas des processus, tels que définis dans l'exposé sur les «Modèles de processus».

En effet:

- Les composantes sont hétérogènes et ne sont pas des unités de comportement: "décision" "action" et "mesures de" ne se prêtent pas à l'intégration, étant de natures différentes, et les liens indiqués ne s'interprètent donc pas ;
- Le mot "information" est confus dans cette présentation. Il semble recouvrir à la fois "préhension des événements", "données", "observations", "connaissance des faits", "outputs d'un Système d'Information de Gestion";
- L'information est présentée comme premier élément de la séquence, et serait donc initiatrice de cette séquence ainsi que stimulus du "processus", ce qui est loin d'être le cas général. Ainsi par exemple, la téléonomie (les projets, les objectifs) est absente ici, alors que c'est elle qui peut être initiatrice d'un appel à l'information.

En réponse à ces remarques critiques, cette deuxième partie de l'exposé vise à améliorer et préciser de telles "séquences" en s'aidant de la présentation d'un processus bayesien de gestion explicitant l'apprentissage par l'expérience, et qui sera repris à d'autres fins dans l'exposé sur «Le Domaine de la gestion». Ce sera la Figure **5**, modifiée d'après MORRIS (1968), qui forme un *processus de gestion.*

Ce processus est tel que l'expérience est enrichie par la confrontation des résultats de décisions aux conséquences espérées, et, de la sorte, augmente l'efficience ou l'adéquation de la conception des choix lors du traitement d'un nouveau "stimulus". Cette efficience serait logiquement accrue par le fait que plus rapidement la conception des choix est claire, moins il faudra appeler les processus figurant dans la partie gauche de la figure, à savoir les processus qui sont requis pour améliorer cette conception.

Selon une telle configuration, on pourrait considérer que l'"information" au sens large comprendrait toutes les acquisitions de connaissances qui enrichissent l'expérience et peuvent contribuer à la conception des choix, et dès lors les arcs décrivent l'ensemble des flux gérés par un processus d'information de gestion. En outre, cela comprendrait l'ensemble des supports, des sources, destinations et mémoires utilisés pour engendrer et gérer ces flux. Les entités transférées par ces flux sont notamment des *données* et des *observations*:

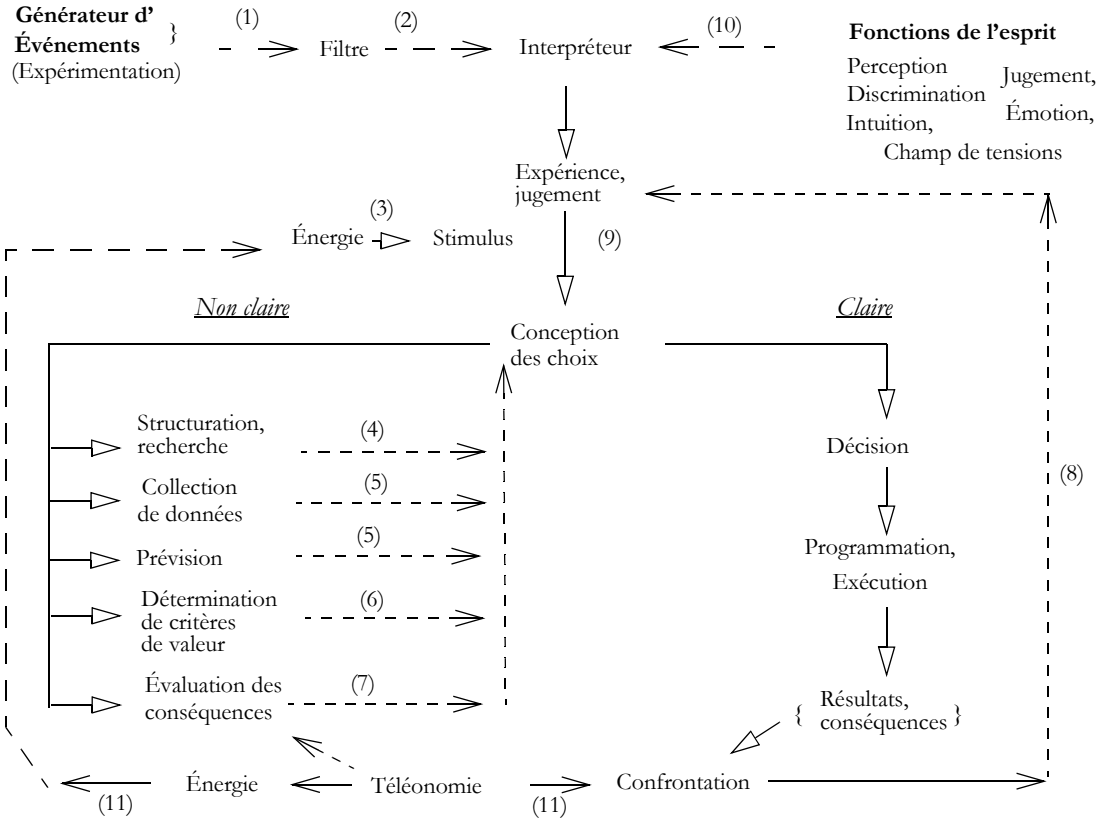
- Par *donnée*, on peut entendre la description totale ou partielle d'un fait, événement, phénomène ou objet, ou de caractéristiques de ceux-ci, qu'elle soit qualitative ou quantitative et concerne ou non le problème considéré;
- Par *observation*, on peut entendre une donnée ou un ensemble de données recueillies volontairement pour améliorer la connaissance du fait, événement, phénomène ou objet concerné;
- Le support de transmission des données et observations est le *message* (au sens propre; au *sens figuré* le "message" est le contenu significatif que l'on a voulu transmettre).

6.2 Distinction des flux d'information dans un processus d'apprentissage

La figure initiale (de MORRIS) a dans cet exposé été garnie et enrichie de différents types de contributions de l'information. En effet, le contexte de la gestion étant toujours celui de situations d'intentions (la gestion est dans un champ de téléonomie), le processus d'apprentissage représenté à la Figure **5** implique des distinctions entre les arcs orientés:

- Les arcs en trait plein orientent le *temps*, donnant le *sens* de lecture de la configuration;
- Les arcs en pointillé sont des flux d'information;
- Les flux (1) et (2) portent des données (et observations et constatations);
- Les flux (4), (5), (6), (7) et (8) portent des informations de la nature de résultats d'une phase d'analyse, qui, dans les cas (6), (7) et (8), seraient des "outputs d'un modèle";
- Entre l'interpréteur et la conception des choix (flux (9) et (10)), il y a un flux de signaux qui "actionne" un comportement: c'est de l'information au sens de *fonction d'énergie*;
- Le rôle du "filtre" est de faire passer les données qui peuvent créer, par innovation, un stimulus engendrant une situation d'intention.

Figure 5. L'information dans un processus "bayesien" de gestion



La conception des choix s'exerce par la fonction de penser (penser, c'est simuler?). Elle fait d'abord appel aux fonctions de l'esprit (selon JUNG et SINGER), à savoir l'*intuition*, la *perception-discrimination*, le *jugement*, l'*émotion*. En gestion, ces fonctions seraient (mais JUNG n'y est pour rien) exercées dans un *champ de tension*. De plus, la transmission des outputs de ces fonctions de l'esprit à la fonction de penser est la *rationalisation*.

On peut se demander alors quel est le type d'"information" qui est transmise entre les fonctions de l'esprit et la fonction de penser. L'acception de J.A. BECKETT [1971, traduite ici] est:

> «Ensemble ou série d'ensembles de patterns de relations impliquant énergie et matière; cela peut impliquer aussi certains types de mémoire».

Si ces relations ont à présent un *pattern*, elles ne sont plus "n'importe quoi"; elles auraient un "schème". Ce seraient donc des "schèmes" (mise persistante de l'idée harmonieuse dans l'espace mental, a dit la Systémographie) qui cheminent sur le flux (10). Or, si une entité dans l'espace a un pattern, elle a plus de forme que le désordre qui lui a donné naissance; il a donc fallu de l'énergie pour apporter et conserver cet ordre, sinon cet ordre se disloquerait naturellement par application du principe d'entropie maximale de la thermodynamique.

Le flux (10) transporte donc aussi de l'énergie.

D'autre part considérons le flux (11): il transfère de l'énergie issue de la *téléonomie*, c'est-à-dire d'attracteurs et répulseurs ("on voudrait", "on ne veut pas"). Les flux (10) et (11) montrent donc aussi une acception de l'information en tant que fonction d'énergie.

On s'interroge enfin sur le type d'information transférée entre le générateur d'événements et les *fonctions de l'esprit*, dont la perception-discrimination. Le *sujet* (ici le concepteur des choix) a déjà de l'information a priori: soit par exemple sous forme de "réseaux de faits" (selon la proposition de LEIBNIZ), soit de "connaissances à priori" (selon KANT), soit sous forme d'"ensembles de possibles" (selon CARNEADES, 2000 ans avant KANT et, plus récemment, vers 1970, selon SCHACKLE).

Lorsqu'on retient la dernière version, selon laquelle les messages concernant les événements modifient dans le chef du récepteur ses degrés de croyance quant à ces événements, on adopte une version de l'information qui est un *modificateur de probabilité* subjective, ou de degré de crédibilité. C'est celle qui donne la version quantifiée figurant dans l'exposé «L'Information et sa thermodynamique».

La gestion présentée sous forme de processus d'apprentissage a donc servi de support pour avancer plusieurs acceptions souvent reprises sans raffinement sous le terme générique d'information.

Évidemment, de nombreuses *données* et *observations* figurent implicitement dans le processus et ce, notamment dans les entités "collection des données, structuration et recherche", mais seul a été débattu ici ce qui enrichit la *conception des choix* du décideur, et pas encore sa *décision*. Ce sera cet aspect de *structuration* (en nuages?) et d'*exploitation* qui formera la deuxième partie, très synthétique, de cet exposé.

7 Nuages de données

La complexité des phénomènes et des situations en économie et gestion demande souvent à l'investigateur de décrire ceux-ci par un ensemble d'observations prises simultanément sur plusieurs variables. L'ensemble des méthodes élaborées pour élucider les informations contenues dans de telles collections de données multi-dimensionnelles relève de l'*analyse multivariée.*

7.1 Les objectifs généraux

- Mesurer les phénomènes;
- Analyser:
  - les collections de données secondaires;
  - les résultats d'expériences;
  - les données de sondages;
- Pour concevoir et élucider:
  - la globalité des relations impliquées par les données;
  - les spécificités des relations présentes dans l'ensemble.

L'analyste a pour tâche de *parler* de cet ensemble; il est confié au gestionnaire d'*agir* sur cet ensemble. Ainsi, en gestion commerciale, on distinguera une phase d'*analyse* de marché d'une phase d'*action* de marketing.

7.2 Les sources

- Les collections de données sont dites *secondaires* lorsqu'elles n'ont pas été recueillies par l'investigateur précisément dans le propos de sa présente étude; elles pré-existaient, et il les utilise à ses propres fins;
- Les collections de données sont dites *primaires* lorsqu'elles ont été recueillies par l'investigateur spécifiquement pour son étude: ainsi en est-il généralement des observations directes, des résultats d'expériences et des données de sondages.

7.3 Les structures de données

Les données multi-dimensionnelles se présentent normalement sous forme de tableaux à deux entrées, et, dans certains cas, à trois entrées. Les deux entrées sont spécifiées par convenance visuelle comme les lignes et les colonnes. Ainsi, deux des exploitations de tableaux de données rectangulaires en marketing sont de situer un produit sur le marché:

- Aider à définir un produit nouveau (trouver un "créneau"). Le *créneau* concerne un ensemble d'attributs qui sont évalués par les clients (potentiels), lesquels ont alors le statut de "juges". La configuration des données serait alors celle du Tableau **1.A**.

- Aider à définir une politique de marque (un "label"). Le *label* concerne un ensemble d'attributs qui sont évalués par rapports à la concurrence; la configuration des données serait celle du Tableau **1.B**.

Tableau 1. Image de marque d'un produit sur le marché

1.A Vu par la clientèle: Créneau						**1.B Relatif aux concurrents: Label**				
	Attribut 1	...	Attribut k	...	Attribut p		Attribut 1	...	Attribut k...	Attribut p
Juge 1						Sujet 1				
...						...				
Juge i						Sujet i				
...						...				
Juge n						Sujet n				

Un élément x_{ik} du tableau de gauche désignerait par exemple l'intensité avec laquelle un juge i (client, donneur d'opinion) associe l'attribut k à l'objet concerné, ce qui peut former des *scores*.

Dans le cas de plusieurs juges, exprimant une opinion quantifiée sur plusieurs (mêmes) critères et plusieurs (mêmes) sujets, le tableau serait tri-dimensionnel et il faudrait donc trois indices. Ce serait en fait un ensemble de tableaux combinant A et B.

Le Document P.2 montre un tel tableau de type "Label". Il présente deux entrées: "Institutions financières - Attributs" vues par un échantillon de répondants à un sondage d'opinion. Un élément x_{ik} d'un tel tableau désignerait le nombre de fois que l'attribut k est associé au sujet i. Il s'agit alors d'un tableau de *fréquences*, ou d'effectifs, et son élucidation par le "Multidimensional Scaling" y est présentée.

7.4 Les échelles de variables

Lorsque les attributs et propriétés sont portés sur une échelle («de mesure»), ils peuvent acquérir le statut de *variable*. Il est conventionnel – et mathématiquement pertinent – de distinguer quatre échelles de variables, à savoir: *nominale*, *ordinale*, *intervalle*, *ratio*.

7.4.1 L'échelle nominale

L'attribut est spécifié par le nom d'une catégorie, et l'objet par le fait d'appartenir ou non à cette catégorie: par exemple "acheteur", "en bois", "de revenu élevé".

En sciences humaines, ces catégories sont en général des *modalités* d'une propriété, d'une caractéristique, ou d'un attribut. Ainsi, l'attribut "type d'habitat" aurait trois modalités – disons "citadin, extra-urbain, rural"; des modalités de "catégorie socio-professionnelle" peuvent être "employé, ouvrier, inactif-étudiant...". Ces exemples relèvent du type nominal "pur", en ce sens qu'il n'y a pas de raison analytique d'établir la liste dans tel ordre plutôt que tel autre.

L'échelle nominale peut être qualifiée d'*ordonnée* lorsque des permutations des modalités sont significatives. Ainsi, des "classes de revenu", ou "tailles du ménage" sont aussi catégorielles, mais ont des modalités *ordonnées*, à savoir rangées dans un suite logique de dominance ("élevé, moyen, faible", ...; "0, 1, 2, ... enfants").

Les données élémentaires (c'est-à-dire non-combinées) figurant dans des tableaux sous l'échelle nominale sont, pour chaque objet et modalité, de type *dichotomique*: "appartient" ou "n'appartient pas" à cette modalité; "être" ou "ne pas être" de cette catégorie, "en avoir ou pas"... De telles informations peuvent être, et sont souvent, codées en *binaire*, à savoir "1" ou "0".

Le remplissage exhaustif de chaque modalité de chaque attribut en code binaire forme un tableau *disjonctif complet*, typiquement le Tableau **2**.

Tableau 2. Un tableau disjonctif complet

Objets	**Attribut A**					**Attribut B ..**		
	Modalité A_1	...	Modalité A_k	...	Modalité A_p	Modalité B_1	...	Modalité B_m
Objet 1	0	1	0	0	0	0	0	1
...	0	0	0	1		0		
Objet i			x_{ik}= 1 ou 0					
...								
Objet n								

Comme cette présentation est lourde; elle est ramenée en général au "nombre de fois que", c'est-à-dire à un tableau de *fréquences*.

Soit à présent que les deux entrées du tableau (les lignes et colonnes) soient des attributs, ou *catégories*. Dans ce cas, les catégories sont *croisées*, et les cellules du tableau sont occupées par des fréquences *jointes*, c'est-à-dire le nombre d'objets relevant simultanément de telle catégorie en ligne et de telle catégorie en colonne.

En divisant ces fréquences par le nombre total d'observations, on obtient un *Tableau de contingence*, ce qui est rendu au Tableau **3**.

Tableau 3. Tableau de contingence

	Attribut A et ses modalités			**Attribut B et ses modalités**			**Attr. C ...**
Modalités	A_1	... A_k ...	A_p	B_1	... B_j ...	B_m	
A_1							
... Ai ...		f_{ik}			f_{ij}		
A_p							
B_1...							
... Bi ...		f_{ik}					
B_m							

Les opérations arithmétiques sur ces tableaux sont toutes celles que l'on peut faire à partir des fréquences, donc des "comptages". La structure de groupe mathématique de cette échelle est celle des *permutations*, à savoir:

$y = \beta(x)$, où $\beta(x)$ indique une correspondance bi-univoque.

Des "statistiques" permises y sont le mode, les associations de contingence, et les mesures issues de la théorie de l'information.

7.4.2 L'échelle ordinale

L'échelle ordinale est plus riche en information que la nominale, en ce sens qu'elle exprime un *rangement* des données selon une relation de *dominance*. Ceci signifie que les unités d'observation, les objets, sont présentés dans un *ordre* selon les attributs concernés, par exemple du plus petit au plus grand, du meilleur au plus mauvais, du "préféré" à l'"exécré", etc. Les expressions numériques figurant dans les cellules d'un tel tableau sont des *rangs*, tels qu'on les lit au Tableau **4**.

Il va de soi que les rangs des objets ne sont pas les mêmes sur tous les attributs (souvent appelés *critères* dans une telle situation). Un des problèmes qui se présentent est alors de rechercher quel est le rangement des objets sur tous les critères simultanément, soit un problème d'*ordre médian*, typique de l'analyse des *préférences*.

L'échelle ordinale ne permet pas d'opérations arithmétiques sur les valeurs, c'est-à-dire les rangs qu'elle exprime; on ne divise pas, ne multiplie pas, des rangs. Si l'on soustrait le rang i (par exemple 10), du rang j (par exemple 15), cela exprime seulement le nombre de rangs à franchir pour passer de i à j, soit le nombre d'objets intermédiaires (ici 4) plus 1.

Tableau 4. Rangements multicritères d'objets

Objets	Critère A	...	Critère k	...	Critère p
Objet 1	12	3	9	4	21
...	3	5	11	5	4
Objet i	2	7	r_{ik}= rang de i sur critère k	8	6
...	1	8	10	9	5
Objet n	9	4	7	11	13

Les écarts de rangs n'expriment *pas les distances* séparant les objets: ainsi le troisième coureur cycliste peut arriver très loin du second, mais en avoir une dizaine juste dans son sillage de feu...

La structure de groupe mathématique de l'échelle ordinale est celle du groupe *isotonique*:

$y = \mu(x)$, où $\mu(x)$ signifie toute fonction monotone croissante.

Des statistiques permises y sont la médiane, les percentiles, les corrélations de rang, les tests de signe et de séquences.

7.4.3 L'échelle d'intervalle

Les variables portées sur l'échelle d'intervalle sont des nombres rationnels, disons tels qu'on en voit d'habitude, mais qui n'ont pas de neutre et ne sont pas définies par rapport aux opérations de multiplication et de division.

Ainsi en est-il par exemple des cotes d'étudiants sur différents cours; les quotients de deux cotes n'ont pas de signification, et d'ailleurs les moyennes arithmétiques des cotes ne s'interprètent pas en principe, le *neutre* n'étant pas défini et les référentiels de niveau n'étant pas nécessairement identiques (c'est ce qu'on qualifie de cotes "subjectives". D'autres cas sont les mesures de température (centigrade et Fahrenheit, dont on n'interprète pas le rapport), des mesures d'énergie, des dates calendaires.

La structure de groupe mathématique de l'échelle d'intervalle est celle du groupe linéaire:

$y = \alpha + \beta x$, où β est positif.

Des statistiques permises y sont – en plus des autres citées – la moyenne, l'écart moyen, l'écart-type, les corrélations moment-produit, les tests t et F de FISHER.

7.4.4 L'échelle de ratio

Les valeurs portées sur une échelle de ratio correspondent aux variables dites habituellement "quantitatives", où toutes les opérations arithmétiques sont permises et ont une signification. Ainsi en est-il d'évidence des variables courantes en gestion, telles les prix, revenus, quantités consommées etc. Elle est caractérisée par la définition d'un zéro significatif, c'est-à-dire *interprétable*. La structure de groupe mathématique de l'échelle de ratio est celle du groupe de similarité:

$y = \gamma x$, où γ est positif.

De nouvelles statistiques permises y sont la moyenne géométrique, harmonique, et les coefficients de variation. Ainsi, si l'on voulait établir un régime de cotes d'étudiants telles que l'on puisse de façon légitime en calculer les moyennes et les corriger pour les niveaux subjectifs différents selon les profs, il suffirait de diviser chaque cote par la moyenne géométrique des trois meilleures, et calibrer le résultat sur 20 par un facteur de dimension.

On entend parfois l'expression de variables "qualitatives" pour celles portées sur les échelles nominale et ordinale, et "quantitatives" pour les autres. En américain, certains utilisent "discrete data" pour les nominales (et leurs comptages).

Les échelles citées sont dans un ordre d'information croissant; cela implique que l'expression des données dans l'échelle "suivante" peut être transformée en les échelles "précédentes" parmi celles qui ont été citées.

7.5 Les tableaux de données chronologiques

La version la plus courante de tableaux de données en économie et gestion est celle de la séquence de valeurs d'un objet observées à des repères temporels discrets en général séparés par des intervalles constants. Il s'agit alors de *séries chronologiques*, où les principaux problèmes posés sont l'analyse de comportement, les associations, les prévisions.

8 Contributions de l'analyse de données multivariée

8.1 Identifier, classer et grouper

- *Identifier* signifie allouer chaque objet, en fonction de ses attributs ou propriétés, de façon adéquate dans une *classification prédéfinie*. Ainsi en est-il de reconnaître un minerai, d'insérer un être vivant dans une typologie, de placer une entreprise dans les "débiteurs douteux".
- *Classer* signifie à la fois créer des classes et insérer les objets dans les classes créées. Il y a donc une typologie à élaborer, ce qui peut se faire en "découvrant" les objets de par leurs attributs. Ainsi pourrait-on classer des ratios de gestion, ou des produits. Cependant, lorsque le domaine est connu, les classes sont souvent pré-existantes; par exemple les classes de ratios de *solvabilité*, de *liquidité* sont familières aux analystes financiers.

- *Grouper* des objets se fait en pratique de façon itérative: les n objets sont considérés au départ comme des groupes d'isolés, et les objets sont progressivement réunis en fonction de leur association, de leur similarité, les groupes formés cessant de s'enrichir lorsqu'ils doivent absorber des objets trop dissemblables. Ceci est typique de l'*analyse des groupements*, ou "Cluster Analysis", formant soit une hiérarchie ascendante, soit des groupes séparés appelés "clusters".

On peut aussi procéder par *décomposition* d'une globalité d'objets: ils forment un seul groupe initial, puis ce groupe est progressivement décomposé en fonction des attributs différenciant les objets. Ceci revient à former une *hiérarchie descendante*, dont une démarche typique est celle de la *segmentation*, qui a beaucoup de succès en marketing dans une optique de structuration de la clientèle.

Les structures de données conduisant à ces démarches sont des données de *similarité*, ou de *proximité*; il s'agit donc de comparaisons par paires. Le tableau à utiliser est dès lors un croisement des objets, formant une matrice carrée où les cellules sont occupées par un indice d'association, ou de similarité, tel qu'en montre le Tableau **5**.

Tableau 5. Indices de similarités entre des objets

	Objet 1	**...**	**Objet k**	**...**	**Objet n**
Objet 1	s_{11}				
...					
Objet i			s_{ik}= similarité de i et k		
...					
Objet n					s_{nn}

Le tableau de similarité peut être une donnée initiale, ou peut être construit à partir des données initiales objets-attributs, telles que présentées au Tableau **1.B**. Dans ce cas, chaque objet est décrit par un *vecteur* de p valeurs sur les p variables, vecteur qui représente en termes familiers son *profil*.

La comparaison deux à deux des objets se fait par celle de leurs profils, et il existe une panoplie d'indices de similarité, ou de proximité, indiquant dans quelle mesure ces profils sont semblables.

Comme ces objets peuvent être considérés comme des points dans l'espace (si les variables définissent un espace métrique), on pourra exprimer ces similarités en termes de *distances*, qui montreront que les objets les plus *similaires* seront les plus *proches* dans l'espace p-dimensionnel des variables.

8.2 Étude des dépendances entre les variables

L'intérêt se porte à présent sur les relations entre les variables, et notamment les dépendances de certaines variables vis-à-vis d'autres. Une vue globale des associations peut être obtenue en exprimant la matrice carrée des p*p associations, par exemple des corrélations de rang ou de moment-produit. Ces indices sont obtenus sur base de l'échantillon formé par les *n* observations sur les *p* variables ce qui forme un tableau initial comme celui de type **1.B**; les associations se présentent alors selon le Tableau **6**.

Ces relations peuvent s'exprimer sur des partitionnements: soit deux groupes de variables, disons les x_j, et les y_k; la question posée à l'analyse de corrélation canonique, par exemple, est celle de la corrélation entre ces groupes.

Tableau 6. Indices d'associations entre variables

	Variable 1	**...**	**Variable k**	**...**	**Variable p**
Variable 1			r_{1k}		
...			...		
Variable j			r_{jk} = associations, corrélations...		
...			...		
Variable p			r_{pk}		

8.3 La prédiction

8.3.1 Inspiration: les modes et les sources

Au-delà du sérieux de la systémique, il y a bien des mots de la langue française qui concernent des assertions relatives à l'avenir. En un coup de rateau, on peut ramasser par exemple:

- Projection;
- Prophétie;
- Prédiction;
- Pronostic;
- Perspective;
- Prospective;
- Prévision;

- Anticipation;
- Conjecture;
- Divination;
- Extrapolation;
- Futuribles...

- Une première façon de les classer est de former deux groupes, dont l'un comprend ceux qui commencent par "P", ou même "Pr"; leur effectif est surprenant, mais le propos s'arrête là;
- Une deuxième façon est de repérer les approches qui pourraient recevoir l'agréation de "scientifiques"; cette qualité les poserait parmi les candidats à figurer dans des "systèmes" de prévision.

À cette fin dernière, cinq conditions au moins doivent être remplies :

- Il faut que l'objet de la prévision soit clairement identifié. Ainsi, "le marché", "la conjoncture", "l'immobilier" sont des ensembles composites aux comportements variés. Et d'ailleurs,

  «*Là où il y a diversité, il ne peut y avoir de vérité*»
  Saint Jérôme

- L'horizon concerné doit être précisé. Il est pourtant chanté que:

  «*Le poète a toujours raison*
  Qui voit plus loin que l'horizon»

  Mais ce que la chanson (de Jean FERRAT) ne dit pas, c'est où se trouve l'horizon en question, surtout quand la fenêtre donne sur le mur d'un HLM de 23 étages;

- Il doit être possible de confronter les prévisions aux réalisations. Par exemple "le harcèlement sexuel des professeurs va se tasser", se prête mal à une telle confrontation, les statistiques à ce sujet étant peu fiables;
- L'*expression* de la prévision ne peut être conditionnelle ou contenir sa propre contradiction. Ainsi, le filosofe qui a dit: le «XXIe siècle sera mystique ou ne sera pas» mérite tout au plus d'être vu à la télé;
- Le *mode de génération* des prévisions doit être explicite; les boules de pétanque (en cristal?), les foies d'oies du Capitole, les excréments de crapaudes sont des bases malsaines, tandis que les illuminations manquent souvent de... clarté. Pour qu'une prévision puisse entrer en systémique, il faudrait à la limite que deux prévisionnistes exploitant les mêmes données fournissent les mêmes assertions sur le futur.

Pour l'illustrer, et étendre les sources d'information, voici d'autres exemples (où l'on s'aide de A. PAUL: *Et l'homme créa la Bible*, BAYARD, 2000, p.234).

La *divination* est un processus par lequel des assertions non vérifiées à ce jour sont issues de *prémisses*. Celles-ci sont plus précisément appelées *protase*, et la conclusion qui en est issue est l'*apodose* (qui vient du grec – ancien! – "apodosis", c'est-à-dire la restitution, l'explication, ou la solution d'un problème). Un exemple (donné par A. Paul p.234) est édifiant :

> «Si la vésicule biliaire du mouton sacrifié est démunie du canal cholédoque, l'armée du roi, au cours d'une expédition, souffrira de la soif».

Comme mode de génération d'assertions sur le futur, on peut aussi essayer la *prophétie*, qui serait assez prometteuse selon PHILON D'ALEXANDRIE (au 1er siècle):

> «Un prophète n'a rien qu'il ignore, puisqu'il recèle à l'intérieur de lui-même un soleil spirituel [noèton hélion] et des rayons dépourvus d'ombre, afin d'appréhender avec une netteté parfaite les *réalités* qui échappent au regard sensible, mais qui sont appréhendées par la pensée ["dianoïa"].»

Ce que ce bon Monsieur PHILON ne précise pas, c'est ce qu'il entend par les "réalités" (souligné ici) dans son contexte; aussi, mieux vaut s'en référer encore à A. PAUL («La génération des interprètes "prophètes"», *op.cit.* p.209):

> «Cela arrive à la gent prophétique: l'intellect [nous], en nous, est chassé au moment où arrive le souffle divin [théou pneumatos]; lorsque celui-ci repart, le nôtre est réintroduit car il n'est pas permis que le mortel cohabite avec l'immortel [...].»

Le fondement de ce mode de génération d'assertions est dû à PLATON; il élabora le concept d'*inspiration* sous ses deux aspects:

- La *possession* (par un esprit...);
- Le *souffle divin*, c'est-à-dire «un dieu dont la personnalité se substitue à la leur» (selon PAUL, *op.cit.)*. Ce serait le cas typique de Moïse – s'il était repéré, et de quelques autres – qui en ont acquis la réputation. Un peu moins "divin", mais plus chatoyant à l'oreille, est le souffle des Muses, qui inspire bien des choses, surtout Ερατο, qui chantait si bien.

  Mieux: cela explique l'assertion faite, dans les présents exposés et par un certain auteur (qui ne mérite pas d'être cité), selon laquelle celui-ci disposerait d'une "intelligence pneumatique". Cette expression laissée en suspens est facile à comprendre maintenant que la relation entre le "pneumos" grec et le *souffle divin* est clairement établie.

Ces conditions pour obtenir le label "scientifique" peuvent pour l'essentiel être ramenées à la *vérifiabilité*; ceci implique donc qu'il doit être possible de montrer qu'elles sont fausses – ce qui est bien le critère "scientifique" de *falsifiablilité* de Karl POPPER. De plus, la vérifiabilité est évidemment plus nette dans le cas d'expressions prévisionnelles quantitatives... et voilà qui nous fait perdre bien des systémiciens prophètes et prédicateurs.

C'est dommage, et il faudra donc en venir à une autre version de la prédiction; mais qu'on ne s'y trompe pas: dans les sociétés humaines, les EAH, le nombre d'assertions du futur et de leurs exploitations fondées sur le divinatoire, le prophétique et autres extra-lucidités est certes bien plus élevé que celui des "systèmes" de projections auxquels on ne croit guère plus. Ils sont largement remplacés par le plus puissant, le plus olfactif des organes prévisionnels, le P.I.F. (Previsional Information and Forecasting?).

8.3.2 Expiration: l'apodose

En analyse des données, la *prédiction* au sens anglo-saxon est l'*assertion* de *valeurs* de variables présumées *dépendantes*, disons les y_k, *conditionnelle* aux valeurs expérimentales de variables présumées *déterminantes*, disons les x_j. La contribution est alors d'estimer les valeurs des **y** pour des valeurs données des **x**. La classe écrasante de tels modèles est celle des relations *linéaires*, ou qui peuvent être ramenées à une forme linéaire par transformation de variables. L'analyse de *régression* en est la forme la plus courante.

Un modèle peut être développé, et ses paramètres estimés, pour des variables mesurées sur différentes échelles. Des exemples en sont donnés ci-après.

- La variable à prédire peut être nominale, conduisant à des modèles de type *logit* et *probit* ;
- Les variables déterminantes (le vecteur **x**) peuvent être catégorielles (des modalités, des facteurs), tandis que la variable à prédire (y) est mesurée sur une échelle par intervalle ou de ratio, ce qui conduit aux analyses de *variance* et de *covariance*;
- Les deux entrées peuvent aussi être *catégorielles*, et on est alors appelé à exploiter des *modèles linéaires généralisés*;
- L'analyse de *covariance* est un modèle linéaire estimant la réponse d'une variable y à des valeurs de variables déterminantes x_j, où on aurait par exemple x_1 catégorielle et x_2 "quantitative" (voir le Tableau **7**). Ce serait le cas où on viserait à estimer la consommation y d'un produit par des ménages, x_1 étant la catégorie socio-professionnelle et x_2 le revenu.

Récemment, la classe des modèles linéaires généralisés a été étendue à des variables de plusieurs types d'échelles.

Tableau 7. Prédiction de y par p variables prédictives x_i

Valeurs de la dépendante y	**Variables prédictives x_k catégorielles**			**Var. prédictives x_j quantitatives**
Observations	$\mathbf{x_1 \cdots}$	$\mathbf{x_k}$	$\mathbf{\cdots x_p}$	$\mathbf{x_{p+1} \cdots x_j \cdots x_q}$
y_1	x_{11}=c	x_{1k}	x_{1p}	$x_{1,p+1} \cdots x_{1j} \cdots x_{1q}$
12	b	...		... 0,84 ...
y_i	x_{i1}	x_{1k}	x_{ip}	$x_{i,p+1} \cdots x_{ij} \cdots x_{iq}$
7,8	a	...		... -14,08 ...
y_n	x_{n1}=b	x_{nk}	x_{np}	$x_{n,p+1} \cdots x_{nj} \cdots x_{nq}$

Un cas particulier relevant à la fois de la classification, de la dépendance et de la prédiction est celui de la *segmentation*. Il s'agit d'un tableau catégoriel, du type du Tableau **1**, mais modifié selon le Tableau **8**.

La segmentation est un modèle qui a du succès en pratique, car il répond au problème de partitionner au mieux les composantes de populations décrites par des modalités. Sa mise en oeuvre itérative demande toutefois de lourdes explorations combinatoires

Tableau 8. Un tableau disjonctif pour la segmentation

Observations	Dichotomie de Y variable maîtresse		Attribut A					Attribut B ..		
	Y_1	Y_2	Mod. A_1	...	Modalité A_k	...	Mod. A_p	Modalité B_1	...	Mod. B_m
Obs 1			0	1	0	0	0	1	0	0
...	"est Y_1"		1	0	0	0	0	...	...	...
Obs i			0	0	x_{ik}= 1 ou 0	0	1	0	1 ou 0	
...		"est Y_2"	0	0	1	0	0	...	...	...
Obs n			0	1	0	0	0	1	0	0

Une fraction des observations est donc "Y_1", par exemple "acheteur" ou "Y_2", par exemple "non-acheteur". La recherche est celle des modalités qui *conjointement* effectuent au mieux la séparation des observations en celles qui sont Y_1 d'une part et celles qui sont Y_2 d'autre part.

Dans un tel modèle, un partitionnement selon un ensemble de modalités ou d'attributs "prédit" l'appartenance à Y_1 ou à Y_2, mais aussi constitue des sous-groupes de l'échantillon de n observations qui ont conjointement les mêmes modalités, et sont appelés *segments*; des exemples en sont "avoir une profession libérale", "habiter une commune rurale", "avoir plus de deux enfants". Ce résultat est une information typique pour le *marketing différencié*, qui considère la correspondance entre un "segment de marché" et une "cible".

8.4 Élaboration et tests d'hypothèses

La contribution de tests d'hypothèses est de valider des assertions en les soumettant à l'expérience des données, ou de renforcer des convictions a priori.

Le sport le plus pratiqué dans ce domaine est l'*inférence statistique*. Celle-ci fait parler d'une population qui n'est pas exhaustivement observable, sur base d'un échantillon observé qui en est extrait. Il est demandé de le faire de la façon la plus crédible possible.

Que veut dire "crédible"? On ne peut parler que de certains aspects de la population-source, tels qu'ils sont perçus par l'analyse de l'échantillon. Ces aspects sont résumés par des paramètres, ou des confrontations de *paramètres*, qui concernent par exemple la *tendance centrale* (moyenne, médiane), la *dispersion* (écart-type, entropie), la *concentration* (intervalles inter-quartiles), des *associations* (coefficients de contingence).

Ces paramètres peuvent être calculés sur les observations contenues dans l'échantillon.

8.5 La réduction dimensionnelle ou "simplification structurelle"

La réduction dimensionnelle n'est en principe pas explicative, mais est *exploratoire*, en ce sens qu'elle aide l'investigateur à mettre en évidence des relations globales entre des données, et à faire apparaître des *facteurs* sous-jacents (on dit des variables "latentes") qui prennent en charge les comportements communs des variables spécifiques.

Ceci est un apport typique des approches *factorielles*, où nous retrouvons parmi les plus connues l'analyse en composantes principales, l'analyse factorielle "classique" et *oblique*, l'analyse des *correspondances*, et la panoplie de méthodes de *Multidimensional Scaling*. L'aide apportée est de ramener l'espace de projection à deux ou trois dimensions, de sorte que les points concernés peuvent être portés sur une surface – une feuille, un écran – en perdant le moins possible d'information contenue dans les variances et covariances. Ainsi, soit un tableau de données de n objets z_i (lignes) sur p variables v_k, tel que le Tableau **9**.

Tableau 9. Scores de n objets z_i sur p variables v_k

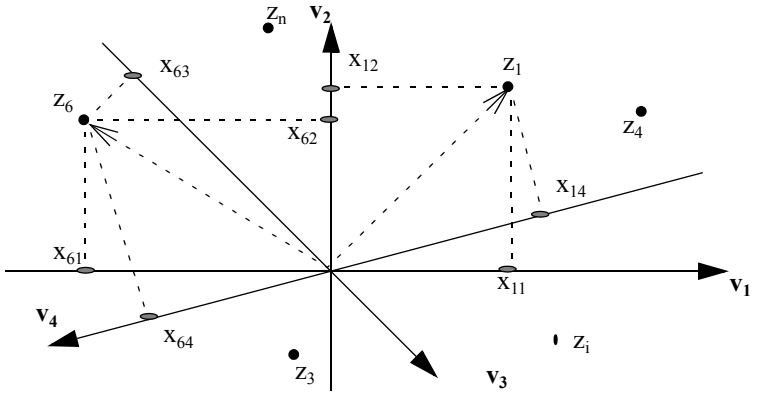
Objets	**Variables prédictives x_j catégorielles**			**Var. prédictives x_j quantitatives**
	x_1 ...	**x_k**	**... x_p**	**x_{p+1}... x_j ...x_q**
z_1	**x_{11} ...**	x_{1k}	x_{1p}	$x_{1,p+1}$... x_{1j} ... x_{1q}
...				
z_i	x_{i1} ...	x_{1k} = valeur de z_i sur v_k	x_{ip}	$x_{i,p+1}$... x_{ij} ... x_{iq}
...		...		...
z_n	x_{n1} ...	x_{nk}	x_{np}	$x_{n,p+1}$... x_{nj} ... x_{nq}

Graphiquement, tout objet z_i peut être décrit par un vecteur de p éléments x_{ik} porté sur un référentiel R^p, le nombre de variables en donnant les dimensions. Il y a donc n points z dans l'espace R^p. La Figure **6** présente une telle configuration, où des vecteurs issus de l'origine, conduisant par exemple à z_1 ou à z_6, sont dessinés

Les projections orthogonales sur les axes v sont les coordonnées des points; nous constatons par exemple que z_1 et z_6 sont opposés sur l'axe v_1, mais associés sur l'axe v_2; z_3 et z_n sont opposés sur v_2, mais associés sur v_1. Les *profils* de ces points sont donc différents. Si les profils étaient semblables, les points seraient groupés dans l'espace, du fait que leurs coordonnées (les *scores* sur les variables) seraient peu différentes. La similarité est donc, du point de vue topologique, associée à la *proximité*, laquelle est indiquée par une *distance*.

Des conditions doivent cependant être satisfaites pour que des distances soient mathématiquement des *mesures*, et elles ne le sont pas nécessairement pour les différentes échelles de variables, notamment les variables nominales et ordinales.

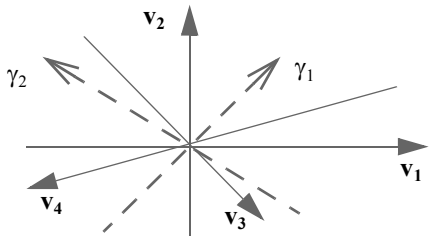
Figure 6. Configuration de points-objets z_i dans un espace R^4 des variables v_k.



La configuration de la Figure **6** n'est qu'une illusion d'optique, en ce sens que les proximités ne peuvent apparaître visuellement dans un espace multi-dimensionnel non représentable évidemment. La réduction structurelle consiste en l'élaboration d'un nouveau référentiel, qui conserve le maximum d'information contenue dans les p variables initiales en le minimum de dimensions.

Ce nouveau référentiel est celui des *facteurs*, ou des *composantes principales*. La propriété essentielle des facteurs est de prendre en charge l'*association* entre les variables v, de sorte que ces facteurs représentent les dimensions pour lesquelles l'information est en quelque sorte originale, spécifique à cette dimension, et non "redondante" au sens que la même information n'est pas commune à plusieurs variables. L'entrée se fait donc par les variables v. Celles-ci sont projetées sur les nouveaux axes, disons les γ_q, ce qui fournit une éventuelle interprétation des associations entre les variables.

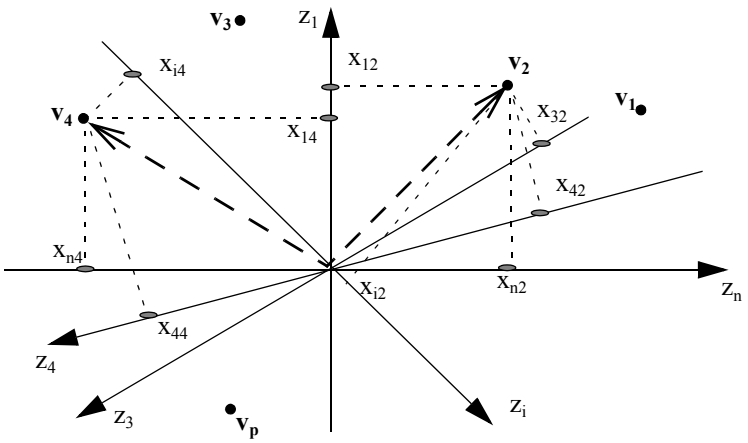
Figure 7. Du référentiel des variables v_k vers l'espace R^m des facteurs γ_q.



La Figure **7** montre cette modification du référentiel. par rapport aux variables. Si l'on entre dans l'analyse par les variables v_k, chacune de celles-ci peut être lue comme un vecteur de n éléments qui figurent en colonnes: les points-variables "v" sont donc situés dans l'espace R^n des observations, les n coordonnées de chacune de ces variables étant les valeurs, ou les scores, des observations sur celles-ci.

Dès lors, les variables seront, pour cet échantillon, d'autant plus associées que les listes de valeurs seront plus *similaires*; les v_k seront alors plus *proches* dans l'espace R^n, tel qu'à la Figure **8**.

Figure 8. Les variables v_k dans l'espace R^n des objets z_i, i=1,... n.



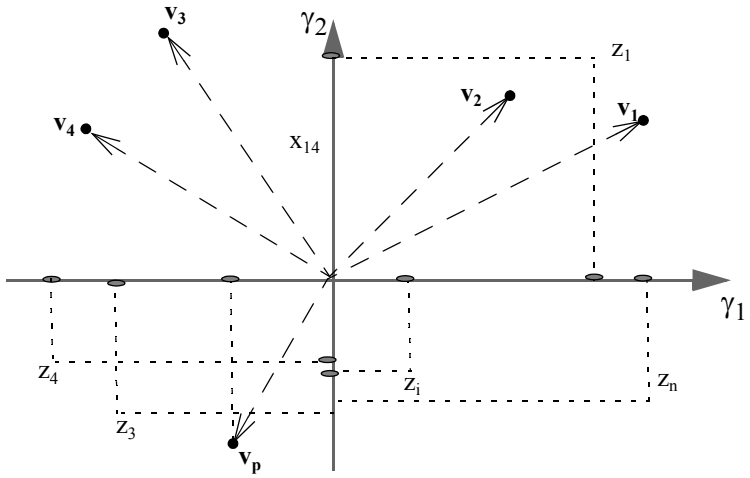
Cette association peut être exprimée par des coefficients adéquats, dont aux premières loges les coefficients de corrélation. L'expression de toutes ces associations entre les variables peut être portée sur une matrice p*p, qui était celle du Tableau **6** ci-avant.

Sur base de ces associations peut se construire (par des techniques de projection) le nouveau référentiel des **γ**. Ensuite, les variables v_k sont projetées sur ce référentiel, comme à la Figure **9**, et leur configuration dans l'espace, par rapport aux axes peut donner lieu à interprétation.

Dans certains modes d'analyse de données, les objets z_i peuvent également être projetés sur ce nouveau référentiel des facteurs **γ**. Comme cette nouvelle configuration est plus parcimonieuse (on ne retiendra que deux ou trois nouveaux axes) que celle des p dimensions de l'espace original, on pourra "voir" les positions des objets par affichage, ce qui justifie l'expression "réduire pour visualiser". Une telle contribution est apportée notamment par l'analyse des *correspondances* (sur base de tableaux de *contingence*), et par le *Multidimensional Scaling*, où les tableaux de données peuvent être sur échelle métrique ou non-métrique.

La Figure **9** montre une projection simultanée des variables v_k et des objets z_i dans l'espace $\boldsymbol{R}^m$ des facteurs γ_q, q=1,... m (ici m=2) et des objets sur les facteurs **γ**.

Figure 9. Projection des variables v_k et des objets z_i dans l'espace R^m des facteurs



9 Exploitations des analyses de données

9.1 Les classes d'analyse

9.1.1 L'exploratoire

Les analyses *exploratoires* tendent à suggérer des hypothèses plutôt que de les tester; elles c servent donc surtout à identifier et comprendre des structures de données complexes. Il n'y a en principe pas de modèle statistique ou économétrique préalablement défini, qui soit candidat à l'acceptation ou au rejet par l'épreuve des données.

Cette classe comprend une variété d'approches, parmi lesquelles l'analyse des *groupements*, les analyses en *composantes principales* – qu'elles concernent des données ordinales ou continues – l'analyse *factorielle*, l'analyse des *correspondances*, le *Multidimensional Scaling*, et certaines formulations particulières comme les modèles *log-linéaires* symétriques. Des visions globales d'associations sont aidées par les tableaux de similarités, les matrices de corrélations.

9.1.2 L'explicatif

Les analyses *explicatives*, en revanche, visent à éclairer l'investigateur sur des processus et structures du monde réel, sur la base d'hypothèses spécifiées a priori, qui sont soumises à l'épreuve des données.

Ce qui est recherché, ce sont des associations, si possibles justifiées par des liens de causalité, entre des variables exogènes et endogènes, ou entre des variables endogènes mutuellement. Plus pratiquement, on y distinguera des variables "dépendantes", ou "à expliquer" (souvent de symbole "**y**"), et des variables déterminantes, parfois dites "explicatives" (souvent de symbole "**x**").

Dans certains contextes, les variables dépendantes sont des réponses, par exemple d'agents économiques et sociaux, tels des consommateurs, des ménages, des opérateurs en bourse, à des stimulus, tels des besoins, des promotions.

Une classe particulière et intéressante sont les analyses explicatives de choix discrets, par lesquels une aide est apportée à l'analyse de comportements d'individus ou de groupes d'individus dans certains contextes de décision, tels ceux du choix d'une résidence, du choix d'un mode de transport ou de tel type d'emploi.

Le Tableau **10** en présente des exemples pour les deux entrées, selon les échelles des variables. Cette partition est cependant incomplète, et n'est pas sans recouvrement.

Tableau 10. Exemples d'analyses "explicatives" selon les échelles de variables

<table>
<tr><th>Variables "dépendantes" y</th><th colspan="4">Variables "prédictives" x</th></tr>
<tr><td></td><td>Binaires</td><td>Discrètes (nominales)</td><td>Métriques</td><td>Mixtes</td></tr>
<tr><td>Binaires</td><td>Tableaux de contingence 2*2, Régression logistique, Modèles log-linéaires</td><td>Régression logistique généralisée, Modèles log-linéaires</td><td>Réponse quantale à des "doses"</td><td></td></tr>
<tr><td rowspan="2">Discrètes nominales</td><td colspan="2">Tableaux de contingence, Modèles log-linéaire</td><td></td><td></td></tr>
<tr><td></td><td colspan="3">Régression linéaire logistique généralisée, "logits", réponse quantale à des "doses", modèles "probit"</td></tr>
<tr><td>Discrètes ordinales</td><td></td><td>"Rapports de chances" ("proportional odds")</td><td></td><td></td></tr>
<tr><td></td><td colspan="4">Modèles linéaires généralisés (plusieurs catégories)</td></tr>
<tr><td rowspan="2">Métriques</td><td>Tests "t"</td><td>Analyse de Variance</td><td>Régression multiple classique</td><td>Analyse de covariance, Régressions mixtes (binaires)</td></tr>
<tr><td></td><td></td><td colspan="2">Modèles économétriques classiques</td></tr>
</table>

9.2 Synthèse des exploitations des analyses de données

Une liste de contributions de l'analyse des données à l'information est résumée dans l'Encart **1**.

Encart 1. Contributions d'analyses des données à l'information

Mesurer	:	Poser des échelles adéquates, uni- et multi-dimensionnelles;
Expliquer	:	Confronter une ou des variables "dépendantes", à des variables déterminantes; construire un modèle, "reproduire" les données par ce modèle;
Ordonner	:	Établir et comparer des rangs, par relations de dominance; sélectionner des objets, individus selon des critères, formant une structure d'ordre.
Décrire	:	Établir des paramètres, des présentations explicites d'aspects pertinents;
Synthétiser	:	Établir des paramètres de synthèse; réduire pour simplifier, pour interpréter;
Préférer	:	Ranger selon la dominance subjective, ou selon les concordances et discordances d'opinions;
Comparer	:	Établir des relations d'associations et d'oppositions, de similarité, de correspondances;
Grouper	:	Former des agglomérats, des sous-ensembles cohérents;
Typer	:	Classifier sans hiérarchie; établir des typologies, des partitions emboîtées;
Identifier	:	Intégrer dans une classification, discriminer selon un critère a priori;
Agréger	:	Former des ensembles additifs, minimisant la perte d'information interne;
Segmenter	:	Partitionner selon une hiérarchie descendante;
Visualiser	:	Projeter sur un référentiel parcimonieux, aider la lecture graphique;
Valoriser	:	Rechercher la valeur de l'information de l'échantillonnage.

10 Les Voix de l'information

10.1 Des rumeurs dans des champs de tension

L'information a de multiples aspects; ceux dont on a parlé en sont quasi des extrêmes.

On l'a tout d'abord abordée comme une *mesure*, adressée au nombre de bits d'information que l'on peut transmettre: c'est l'aspect relevant de la *théorie de l'information,* apanage des "systèmes" techniques-symboliques qui ne font qu'assurer la logistique des processus impliquant des événements et leurs aléas.

Peut-être un apport des mesures citées est-il de nous aider à appréhender le degré d'ordre dans nos idées? Mais attention: des fois, on échange ses idées avec quelqu'un, et on en revient tout bête.

À l'autre extrême, l'information est un *champ de tension* qui affecte les jugements et les comportements des agents.

Dans ce champ se distinguent des voix de l'information qui sont avant tout celles de la *rumeur*. Alors qu'en théorie de l'information ce sont des signaux, froids et digitalisés, donc sans piment et sans saveur, qui cheminent bit à bit sur les canaux de communication qui irriguent ces champs. En pratique de la gestion, ce sont les rumeurs qui répondent le mieux au schéma de base de cette théorie.

Entre le récepteur à l'écoute et les émetteurs diffus et non-repérables dans un champ, la rumeur place son codage, ses filtres, ses interpréteurs, ses biais, et les paradoxes de l'ouï-dire par sous-entendus... mais au fait, la diffusion de la rumeur, sa répartition maximale, n'est-elle pas une forme de maximisation d'entropie? C'est très souvent cette rumeur qui constitue le point d'entrée de l'investigateur, lequel commence à s'éveiller quand il en entend des voix... et l'output de l'investigation peut initier un processus de gestion, exerçant par là une fonction téléonomique.

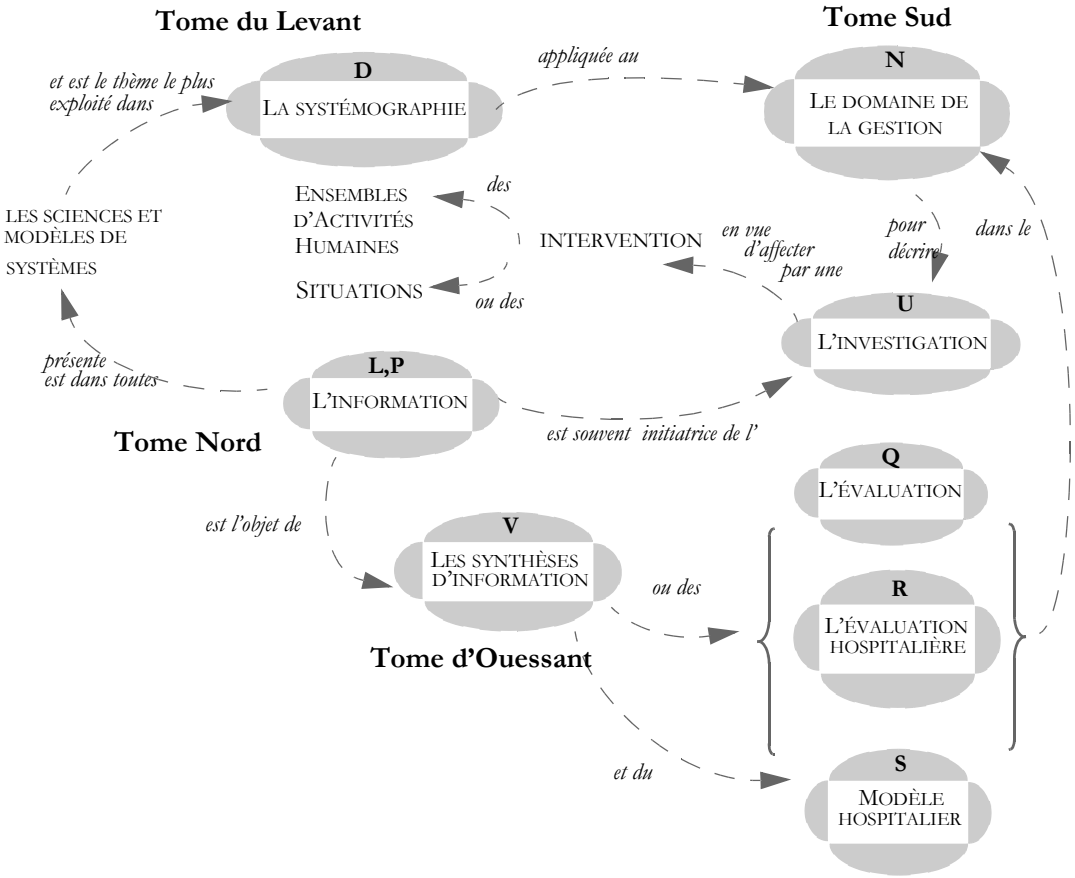
À cette dernière acception de l'information, qui se doit d'être complétée par des observations et des analyses, on attribue même des propriétés de *fonction d'énergie* quand elle stimule ses clients pour qu'ils s'excitent sur quelque chose, par exemple mettent en œuvre un processus de choix suivi d'action, ou parfois fassent cesser un processus oiseux. En ce sens, l'information téléonomique peut aussi "inverser" une décision, c'est-à-dire orienter vers une autre préférable.

Ceci fait apparaître un indicateur de performance d'un "système d'information" de gestion, à savoir *minimiser la probabilité d'inversion de décision*.

Ceci dit, alors qu'un *processus* a été défini comme une interaction continue impliquant un échange de matières et d'énergie, il est étendu ici aux symboles ou ensembles de symboles, c'est-à-dire des phrases, des codes, des nombres, des images. Son exécutif, entité opérationnnelle du processus d'information, en est le *processeur*.

10.2 Les membres du réseau

Figure 10. L'information comme initiatrice et réceptrice



11 Bibliographie

ACKOFF, R.L. [1958]
«Towards a behavioral Theory of Communications»
Management Science, **4**.

BECKETT, J.A. [1971]
Management Dynamics
McGraw Hill, p.91.

CHURCHMAN, W. [1961]
Prediction and Optimal Decision: Philosophical Issues of a Science of Values
Prentice Hall.

COCHRAN, MOSTELLER & TUCKEY [1954]
Statistical Problems of the Kinsey Report
American Statistical Association, Washington.

COCHRAN, W. [1963]
Sampling Techniques, 2e éd.,
Wiley, 356-359.

BIRNBAUM & SIRKEN
«Bias due to non-availability in Sampling Surveys»
Journ. of American Statistical Association, **45**, 98-111.

COLSON, G. & DE BRUYN, CHR. [1989]
«An Integrated Multiobjective Portfolio Management System».
In: *Mathematical and Computer Modelling*, Pergamon Press, **12**, N°10-11.

DE BRUYN, CHR. & COLSON, G.1973]
«Essai de schématisation d'une démarche informationnelle de gestion»
Revue belge des Sciences Commerciales, **5**.

DE BRUYN, CHR. [1970]
«Peut-on "se" connaître vis-à-vis d'un marché spéculatif?»
Revue belge des Sciences Commerciales, **3**.

DE BRUYN, CHR., & EK, C. [1974]
«Analyse statistique et qualitative d'une enquête sur la départementalisation»
Revue des Sciences Économiques, ULg, 4e trim, 1974.

DE BRUYN, CHR. & PIRNAY, F. [1992]
«L'appréciatif en gestion hospitalière»
Méthodes Quantitatives de Gestion, ULg.

DE BRUYN, CHR., & TRAN, Q.D. [1991]
«Monitoring des performances d'une institution de santé»,
in *Bases de données périnatales et assurance de qualité*, A.U.D.I.P.O.G.

GREEN P. & THULL, N. [1971]
Research for Marketing Decisions
Prentice Hall.

KONIJN, H.S. [1973]
Statistical Theory of Sample Survey Design and Analysis
North Holland.

MORGENSTERN, O. [1972]
Précision et incertitude des données économiques
Dunod, Paris (1e et 2e édition en 1950 et 1963).

MORRIS, W.T. [1968]
Management Science, A Bayesian Approach
Prentice Hall Int. Series in Management.

MUCCHIELLI, R. [1973]
Communication et réseaux de communication
2e éd., Entreprises Modernes d'Édition, Paris.

PICHAULT, F. [1991]
Le conflit informatique
Université de Liège.

WEISMAN, H.M. [1972]
Information Systems, Services and Centers
Wiley, New York.

12 Document P.1. La force des poils

a Tableau de données

Soit le Tableau **1** de données présentant les scores d'un test d'aptitude physique d'étudiants ("Test"), les résultats d'épreuves (force, vitesse, endurance) et leur poids et leur taille. Les observations distinguent les cheveux Clairs (C) ou Foncés (F), et c'est ce qui est passionnant. Pour une vision plus directe, les colonnes suivantes sont ces variables centrées (leur moyenne a été soustraite), qui seront seules utilisées. La priorité de ce document est d'illustrer – très succintement – quelques notions figurant dans cet exposé, en donnant un résultat d'analyse descriptive, associative, prédictive et visionnaire.

Tableau 1. Données des tests d'aptitude physique

Observ.	Chev.	Test	Force	Vitesse	Endur.	Poids	Taille
C1	C	3,17	9,08	-4,67	-4,08	15,25	-12,92
C2	C	2,17	4,08	-2,67	-0,08	3,25	-3,92
F1	F	-0,83	0,08	6,33	-5,08	-3,75	-10,92
C3	C	-2,83	-0,92	-2,67	3,92	-2,75	6,08
F2	F	-5,83	-3,92	4,33	0,92	-7,75	-0,92
F3	F	2,17	-1,92	2,33	2,92	-1,75	11,08
F4	F	-2,83	-6,92	4,33	-3,08	-8,75	-8,92
C4	C	5,17	6,08	-4,67	0,92	13,25	5,08
F5	F	6,17	1,08	-0,67	1,92	0,25	-0,92
F6	F	-3,83	-2,92	0,33	-4,08	1,25	13,08
C5	C	-0,83	-0,92	-3,67	3,92	-5,75	1,08
F7	F	-1,83	-2,92	1,33	1,92	-2,75	2,08

b 2. Description

Le Tableau **2** présente les moyennes des scores selon la couleur des cheveux.

Tableau 2. Scores moyens selon la couleur des cheveux

Cheveux	Test	Force	Vitesse	Endur.	Poids	Taille
C	1,36	3,48	-3,66	,91	4,65	-,91
F	-,97	-2,50	2,61	-,65	-3,35	,65

Un test non-paramétrique (il y a peu d'observations) comparant les "variables" Clair et Foncé montre une discordance très significative entre ces moyennes, d'ailleurs très visible. Mais y a-t-il un naïf dans la salle? Quelqu'un qui croirait à une "causalité"?

c Association

Le troisième tableau illustre une des associations que l'on peut calculer. Il s'agit de la corrélation habituelle entre les variables, ici pour les seuls cheveux clairs.

Tableau 3. Matrice des corrélations entre les variables

Chev.	Test	Force	Vitesse	Endur.	Poids	Taille
Test	1,00	0,86	-0,70	-0,68	0,87	-0,35
Force	0,86	1,00	-0,68	-0,94	0,96	-0,68
Vitesse	-0,70	-0,68	1,00	0,52	-0,74	0,32
Endur.	-0,68	-0,94	0,52	1,00	-0,84	0,87
Poids	0,87	0,96	-0,74	-0,84	1,00	-0,49
Taille	-0,35	-0,68	0,32	0,87	-0,49	1,00

d Proximité

L'illustration de la *proximité* ("numérique") est faite au Tableau **4**, présentant la *distance* de degré 1 (somme des valeurs absolues des écarts) entre les variables. Il est rassurant d'y constater l'opposition, annoncée à la section **8** de l'exposé, entre les *associations* (du Tableau **3**) et les *distances* présentées au Tableau **4**. Évidemment, à la moindre distance correspond une plus grande proximité.

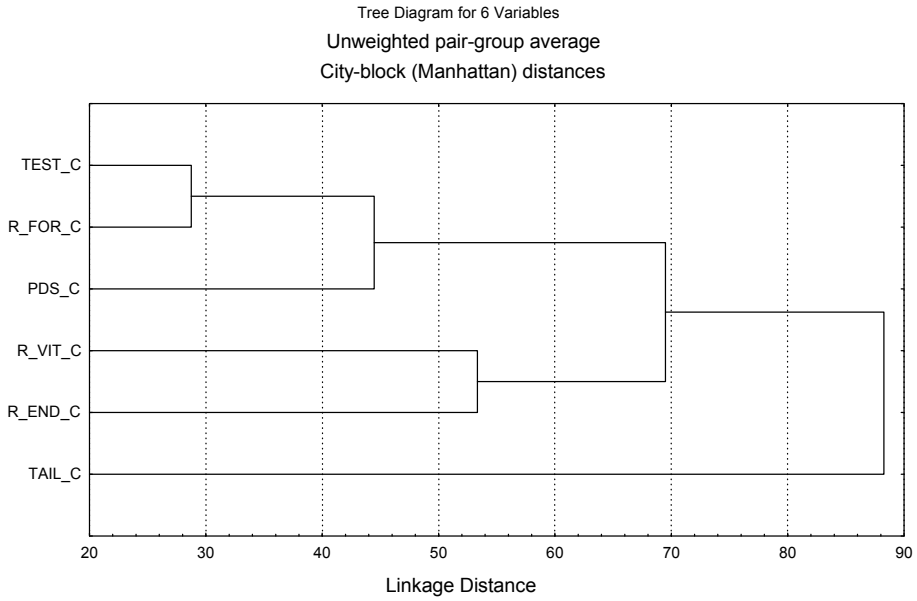
Tableau 4. Matrice des distances (arrondies) entre les variables

Chev.	Test	Force	Vitesse	Endur.	Poids	Taille
Test	0	28	64	45	53	91
Force	28	0	75	57	36	92
Vitesse	64	7	0	53	91	91
Endur	45	57	53	0	84	64
Poids	53	35	90	83	0	104
Taille	91	91	91	63	104	0

e Groupement

La Figure **1** présente une *agglomération* progressive des variables fondée sur les distances calculées selon la métrique de degré 1 – c'est-à-dire celles du Tableau **4**. Il est clair que la *taille* a les valeurs le plus autonomes – donc elle se "raccroche" au plus tard aux autres variables, tandis que les variables de *force* sont les premières associées.

Figure 1. Groupement progressif des variables ("_c" indique "centrées")



La Figure **2** montre l'agglomération des *observations* du Tableau **5**, lequel n'en donne que trois lignes très typiques (par exemple C1 est "très loin"). La distinction selon la couleur des cheveux y est évidente, mais F5 est une observation "déplacée".

Figure 2. Groupement hiérarchique progressif des observations

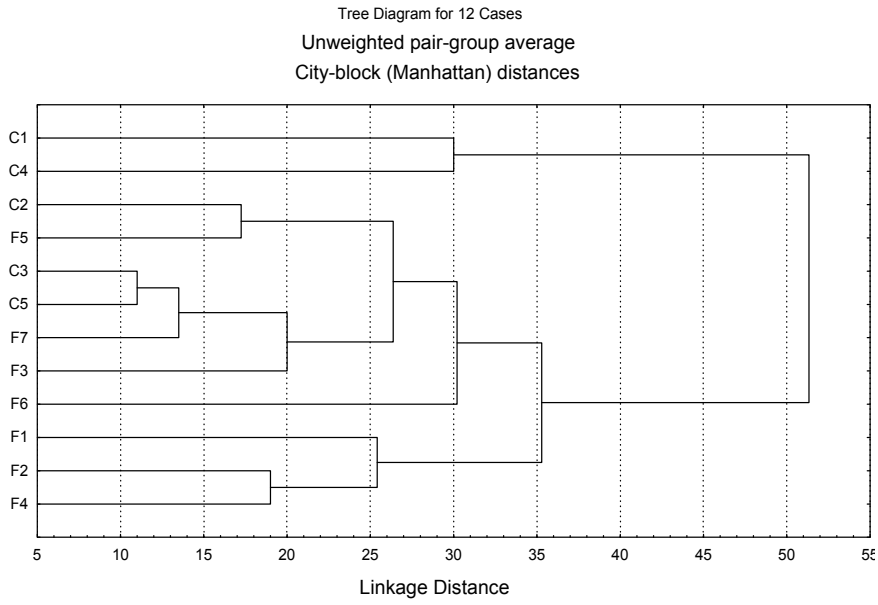


Tableau 4. Matrice des distances (arrondies) entre trois observation

	C1	C2	F1	C3	F2	F3	F4	C4	F5	F6	C5	F7
C1	0	33	46	63	71	67	60	30	48	64	58	62
F2	71	38	30	28	0	32	19	57	30	35	23	17
F7	62	29	29	13	17	17	30	42	19	24	14	0

f Prédiction: analyse discriminante

Dans cette section, la variable dépendante (nominale) est la couleur des cheveux (Clair ou Foncé), ce qui correspond à Y_1 et Y_2 dans l'exposé. On cherche à "prédire" Y_j par les variables déterminantes x_k, ici les scores obtenus aux tests. L'idée est, si on recueille un ensemble de tels scores, de "prédire" leur type de propriétaire. À cette fin, on a appliqué l'analyse discriminante classique à deux groupes, bien que les hypothèses (dont le nombre d'observations) soient flageolantes, mais cela permet de présenter les résultats sur une seule page – le critère absolu d'admissibilité pour tout document adressé à un Chef.

Le Tableau **6** donne les résultats techniques standard de la discrimination paramétrique; on y lira que la discrimination est "significative", en ce sens que la probabilité associée à l'obtention d'une telle valeur de F(soit 8,12) par le hasard des degrés de liberté (ici 4;7) sans effet des facteurs discriminants est inférieure à 0,01. De plus, il est "le plus significatif" de retirer la variable Vitesse, ce qui implique que c'est cette variable qui contribue le plus à la discrimination entre les deux catégories.

Cette assertion repose sur le fait que le paramètre "F-remove" (littéralement "F-retirer") est le plus élevé (4,639), et la probabilité d'atteindre une telle valeur par le hasard des degrés de liberté (4;7) sans effet de ce facteur est inférieure à 0,07 (le "p-level").

À l'opposé, la variable *endurance* a la plus faible contribution.

Tableau 6. Résultats techniques de la discrimination

Wilks' Lambda: 0,177 $F_{4,7} = 8,12$ $p< 0,0091$

	Wilks' Lambda	Partial Lambda	F-remove (1,7)	p-level
Test	,25	,70	2,97	,12
Force	,24	,72	2,61	,15
Vitesse	,29	,60	4,63	,06
Endur.	,19	,92	,55	,48

La distance de Mahalanobis entre les centroïdes des observations "Clair" et "Foncé" est de 19. Ceci implique que leur écart est globalement significatif et que la discrimination peut être efficace. À présent, le modèle va donc être utilisé aux fins de prédiction. Comme annoncé dans l'exposé, il s'agit d'un modèle linéaire, disons en simplifié:

$C \Leftarrow \beta 1$ Test + $\beta 2$ Force + $\beta 3$ Vitesse + $\beta 4$ Endurance + constante

Le modèle est bien sûr le même pour F. Les paramètres β_k sont donnés au Tableau **7** pour les deux variables à prédire.

Tableau 7. Coefficients de discrimination de C, F

	C	**F**
	$q_C = 0{,}416$	$q_F = 0{,}583$
Test	-,65	,47
Force	,73	-,52
Vitesse	-,95	,67
Endur.	,27	-,20
Constante	-3,57	-1,92

À part la question de la standardisation – que nous ne traitons pas ici – l'exploitation des coefficients se fait pas l'expression linéaire annoncée: lorsque les scores de nouvelles observations sont obtenus, l'homme-système prédira la couleur des cheveux sans les voir. Ceci avec plus ou moins de fiabilité, mais que c'est beau! Et malin!

Comme dans cet échantillon il y a 5 "C" et 7 "F", la distribution a priori serait 5/12 contre 7/12, que l'on retrouve au Tableau **7**, sous $q_C = 0{,}416$ et $q_F = 0{,}583$.

La combinaison linéaire (par les β_k) des scores observés est projetée sur un seul axe (parce qu'il n'y a ici que deux groupes à séparer) appelé discriminant. Les coordonnées des observations sur cet axes sont appelées *scores canoniques* (non-standardisés). Ceux-ci sont portés au Tableau **8**.

Tableau 8. Scores canoniques des observations

Observ.	C1	C4	C3	C5	C2	F6	F7	F5	F2	F3	F1	F4
Groupe prédit	*C*	*C*	*C*	*C*	*C*	*F*	*F*	*F*	*F*	*F*	*F*	*F*
Score	-3,39	-2,48	-2,08	-1,92	-1,76	0,46	0,71	0,90	1,24	1,82	2,97	3,53

Par la procédure présente, la valeur "zéro" est le seuil de séparation des deux groupes; aux coordonnées négatives sont associées les prédictions "Clair", et positives pour les Foncés, ce qui donne le "groupe prédit" sur le Tableau **8**. On remarquera que F1 et F4, par exemple, rejoignent au plus tard les groupes où figurent les "C" dans le groupement hiérarchique de la Figure **2**, ce qui est cohérent avec leur écart en discrimination.

Une des façons de lire la performance globale de la procédure, est d'examiner les féquences d'allocations correctes des observations utilisées pour estimer les paramètres. Ceci figure au Tableau **9**, qui montre qu'elles sont toutes bien allouées. La vraie performance est cependant celle de l'inférence, à savoir l'allocation correcte de nouvelles observations, c'est-à-dire les prédictions.

Tableau 9. Allocations des deux groupes d'observations.

	$\mathbf{C}_{\text{préd.}}$	$\mathbf{F}_{\text{préd.}}$
$\mathbf{C}_{\text{oberv.}}$	5	0
$\mathbf{F}_{\text{observ}}$	0	7
Total	*5*	*7*

g Projections graphiques

Les observations sont à présent projetées sur des axes principaux, illustrant de la sorte les procédures dites "de visualisation" dans l'exposé. La première projection est faite sur des composantes principales, le grand classique des réductions de dimensions.

Le Tableau **10** en résume des résultats techniques, à savoir les valeurs propres – donnant l'extension des projections sur les gros axes principaux – ce qui correspond à leur variance. Leur valeur est d'autant plus élevée que la projection conserve l'information exprimée par les écarts par rapport au centroïde.

Tableau 10. Valeurs propres et variances des composantes principales

Valeurs propres	% total Variance	Cumul. des Valeurs propres	% Cumulé des val. propres
3,13	52,24	3,13	52,24
1,66	27,81	4,80	80,06
0 ,61	10,19	5,41	90,25

Les projections sur les axes principaux impliquent les coordonnées des variables, qui les situent – c'est le rôle de ces coordonnées – dans ce nouvel espace orthogonal. Celles-ci sont présentées au Tableau **11** pour les trois axes primordiaux.

Tableau 11. Valeurs propres et variances des composantes principales

	Axe 1	Axe 2	Axe 3
Test	,817	,057	-,307
Force	,961	-,145	-,007
Vitesse	-,806	-,401	-,049
Endur.	-,014	,890	-,423
Poids	,932	-,109	,302
Taille	-,141	,823	,494

Le modèle des composantes étant identifié et déterminé, il est possible de le résoudre en fonction des observations, que l'on peut alors projeter sur ces mêmes axes principaux. Plusieurs logiciels d'analyse de données, dont "Statistica, les appellent les "scores"; on les retrove au Tableau **12**.

Tableau 12. Scores des observations sur trois axes principaux.

Observ.	Axe 1	Axe 2	Axe 3
C1	1,83	-1,42	0,25
C2	0,76	-0,16	-0,39
F1	-0,57	-1,86	-0,26
C3	-0,22	1,19	0,02
F2	-1,27	-0,09	-0,09
F3	-0,27	1,06	0,01
F4	-1,25	-1,13	-0,48
C4	1,58	0,56	0,56
F5	0,55	0,33	-1,31
F6	-0,50	0,10	2,75
C5	-0,10	0,99	-0,90
F7	-0,53	0,41	-0,16

Ces projections peuvent être représentées graphiquement. Toutefois, on va plutôt en montrer, comme dernier soupir, un aspect qui va déclencher l'hilarité. Ces mêmes données sont soumises à une autre méthode de projection sur un référentiel de moindres dimensions, via une précure de "mutidimensial scaling", dite habituellement "MDS".

Au lieu de solutions algébriques comme celles des composantes principales et de l'analyse factorielle, le MDS recherche des axes privilégiés ("de synthèse") par des procédures de convergence.

Ici c'est le modèle élémentaire qui a été appliqué, soit "MDSCAL"; sa version métrique est fondée sur une matrice de distances, et c'est bien sûr la distance de Minkovski de degré 1, déjà éditée ci-dessus, qui a été reprise. Les résultats techniques (à l'itération finale) figurent au Tableau **13**.

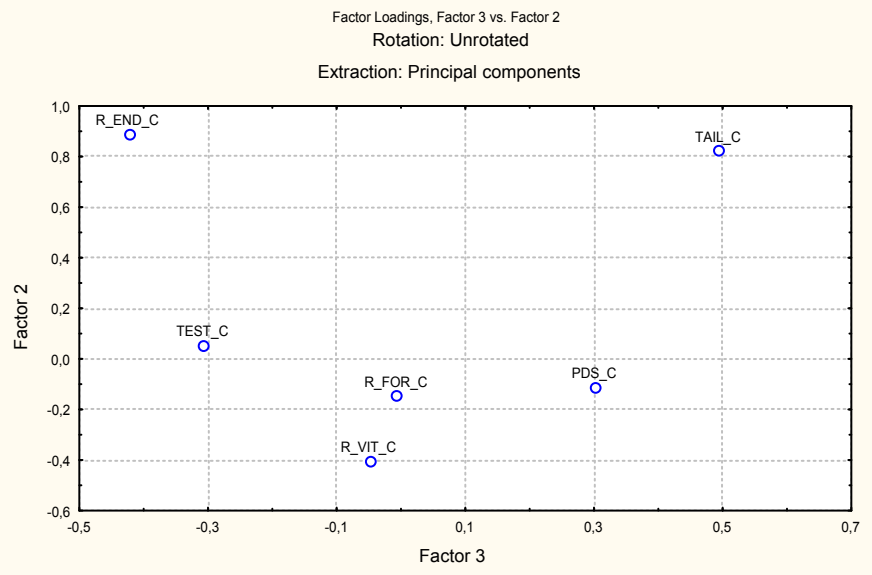
Tableau 13. Le modèle MDSCAL appliqué aux distances entre les variables

Stress brut = 0,000. Stress des ajustements= 0,000005
Le "stress" très faible indique que les écarts résiduels de l'ajustement sont minimes. Les coordonnées des variables sur les (2) nouveaux axes sont les suivantes.

Observ.	**Axe 1**	**Axe 2**
Test	-0,38	0,06
Force	-0,70	0,12
Vitesse	0,27	-1,05
Endur	0,37	-0,04
Poids	-1,09	0,37
Taille	1,53	0,542

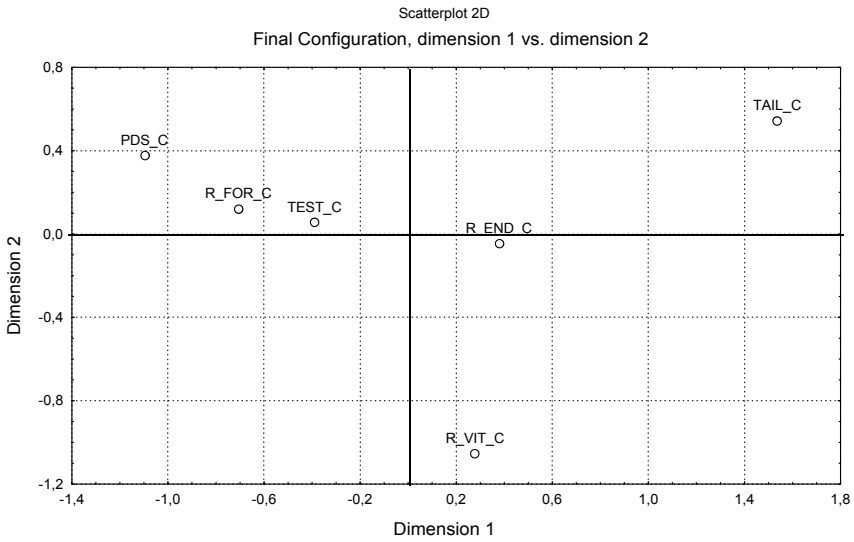
Comparons, in fine, les représentations par les composantes principales et par le MDS. À cette fin, la Figure **3** montre la projection sur les composantes principales *2 et 3*, et *non* sur les dominantes *1 et 2*.

Figure 3. Pojection sur les composantes 2 et 3



Pour le MDS, en revanche, la projection est faite sur les deux premiers axes mis en évidence.

Figure 4. Pojection sur les composantes 2 et 3



La ressemblance entre ces deux configurations est intéressante, montrant l'opposition d'une part entre la vitesse et les autres variables de performance, et d'autre part l'opposition de la taille et du poids relativement à ces variables. Les premiers axes ne montrent pas cette interprétation: cela confirme que souvent le premier axe principal reflète plutôt des relations de "grandeur" des données, lorsque celles-ci ne sont pas standardisées.

13 Document P.2. "Image de marque" de banques belges

a Les données

Le Tableau **14** présente un extrait typique des données initiales, annoncé dans le texte de l'exposé, où des "attributs" sont présentés à des répondants qui doivent donner le degré d'association présente *à leurs yeux* entre chacun de ces attributs et neuf institutions financières belges de la belle époque.

Les scores des réponses vont de 0 à 5, et ils ont été ici (pour être bref) agrégés sur l'ensemble des répondants. L'échantillon de l'enquête est fortement biaisé puisque ce sont tous des étudiant(e)s.

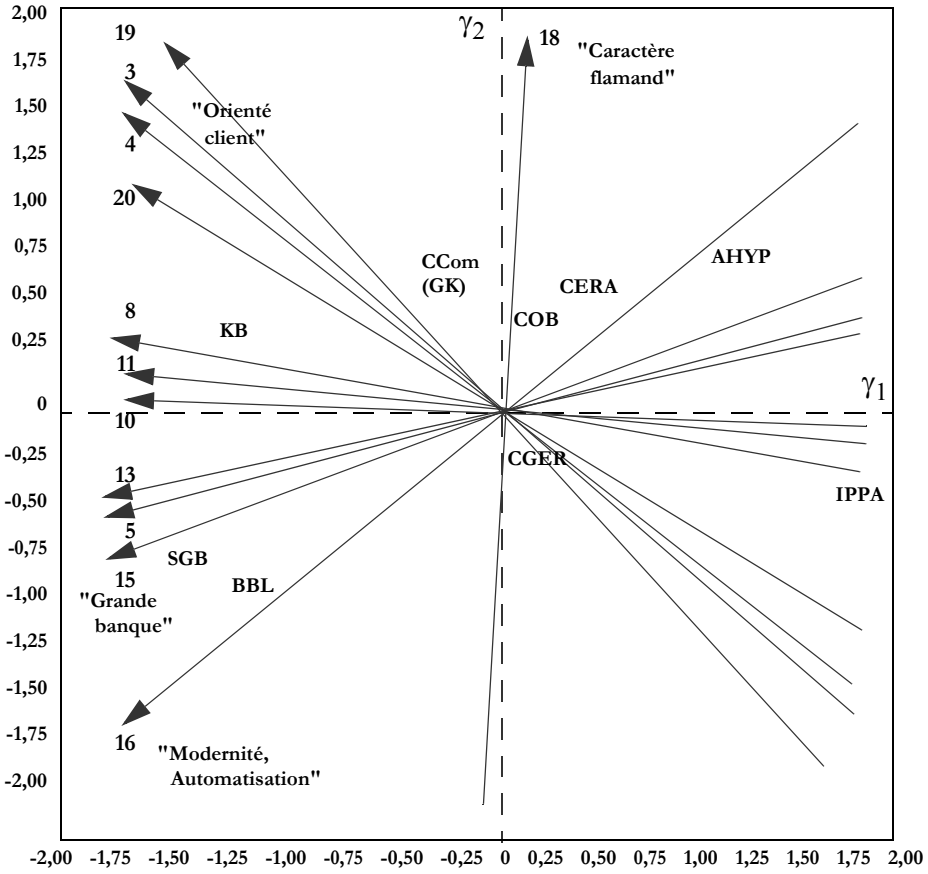
Tableau 14. Scores d'associations entre des attributs et des banques (extrait)

N°	Attributs	CGER	COB	BBL	CERA	CCom	SGB	KB	IPPA	AHYP
1	Publicité attractive	3,20	3,37	4,10	3,46	3,58	3,63	3,8		
2	Personnel amical									
3	Orienté jeunes									
4	Parking aisé	2,90	3,07	3,37	3,44	3,48	3,70	3,53	3,00	3,00
5	Stimule l'économie									
6	Discrétion									
7	Sponsoring									
8	Nombre de filiales	3,88	3,35	4	3,42	3,61	4,22	4,17	2,60	3,29
9	Temps d'attente									
10	Compétence personnel									
11	Orienté vers progrès									
12	Conditions financières									
13	Sécurité						4,01	4,03		
14	Qualité d'information									
15	Réputation mondiale	2,65	2,45	3,64	2,40	2,40	3,89	3,50	1,80	2,15
16	Automatisation	3,54	3,55	3,96	3,25	3,38	4,05	3,94	2,8	3,00
17	Horaire d'ouverture									
18	Caractère flamand	3,33	3,58	2,53	3,91	3,55	2,57	3,76	3,15	3,93
19	Aménagement bureaux							3,82	2,27	
20	Moderne									

Ce tableau a été soumis au "Multidimensional Scaling", déjà cité dans le document **1**; le résultat de base est de projeter les colonnes – ici les banques – sur les dimensions obtenues. Ceci fait, la procédure a été enrichie de régressions effectuées sur les coordonnées des variables sur les axes; les cosinus directeurs impliqués par les coefficients de régression (qui correspondent aux pentes de droites de relations) donnent les axes privilégiés des lignes – ici les attributs – par rapport au référentiel orthogonal.

Le résultat commun aux deux phases est porté à la Figure **5**.

Figure 5. L'image de marque des banques et les directions de l'espace



Cette belle image de marque finit dans les grandes directions de l'espace, montrées en plus gras sur la Figure **5**, et l'exposé laisse à présent au Lecteur le temps de prendre un petit pékêt bien mérité.

D'ailleurs,

> *Quand le seul outil dont vous disposez est un marteau, tout tend à ressembler à une enclume*
> (N.d.l.r. : dans les *jeunes* proverbes chnois, c'est à une faucille).